Course Reader for CS109

CS109
Department of Computer Science
Stanford University
Sept 2021
V 0.2

Get Started

Acknowledgements: This book was written based on notes from Chris Piech for Stanford's CS109 course, Probability for Computer scientists. The course was originally designed by Mehran Sahami and followed the Sheldon Ross book Probability Theory from which we take inspiration. The course has since been taught by Lisa Yan, Jerry Cain and David Varodayan and their ideas and feedback have improved this reader.

This course reader is open source. Want to make your mark? Keen to fix a typo? Download the github project and publish a pull request. We will credit all contributors.

Contributors: Tim Gianitsos, Logan Bhamidipaty

Introduction

Probability is the math of the future. Your ability to program can both illuminate the complexities of probability. But more, the intersection of coding and probability has created a beautiful field of its own.

Notation Reference

Core Probability

Notation	Meaning
	Capital letters can denote events
	Sometimes they denote sets
	Size of an event or set
	Complement of an event or set
	And of events (aka intersection)
	And of events (aka intersection)
	And of events (aka intersection)
	Or of events (aka union)
	Or of events (aka union)
	The probability of an event
	The conditional probability of an event given
	The probability of event and
	The conditional probability of an event given both and
	factorial
	Binomial coefficient
	Multinomial coefficient

Random Variables

Notation	Meaning
	Lower case letters denote regular variables
	Capital letters are used to denote random variables
	Capital is reserved for constants
	Expectation of
	Variance of
	Probability mass function (PMF) of , evaluated at

Notation	Meaning
	Probability mass function (PMF) of , evaluated at
	Probability density function (PDF) of , evaluated at
	Probability density function (PDF) of , evaluated at
	Joint probability density
	Conditional probability density
or	Cumulative distribution function (CDF) of
IID	Independent and Identically Distributed

Parametric Distributions

Notation	Meaning
	is a Bernoulli random variable
	is a Binomial random variable
	is a Poisson random variable
	is a Geometric random variable
	is a Negative Binomial random variable
	is a Uniform random variable
	is a Exponential random variable
	is a Beta random variable

Random Variable Reference

Discrete Random Variables

Bernoulli Random Variable

Notation: $X \sim \text{Bern}(p)$

Description: A boolean variable that is 1 with probability p

Parameters: p, the probability that $X = 1$.

Support: x is either 0 or 1

PMF equation: $$\text{P}(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

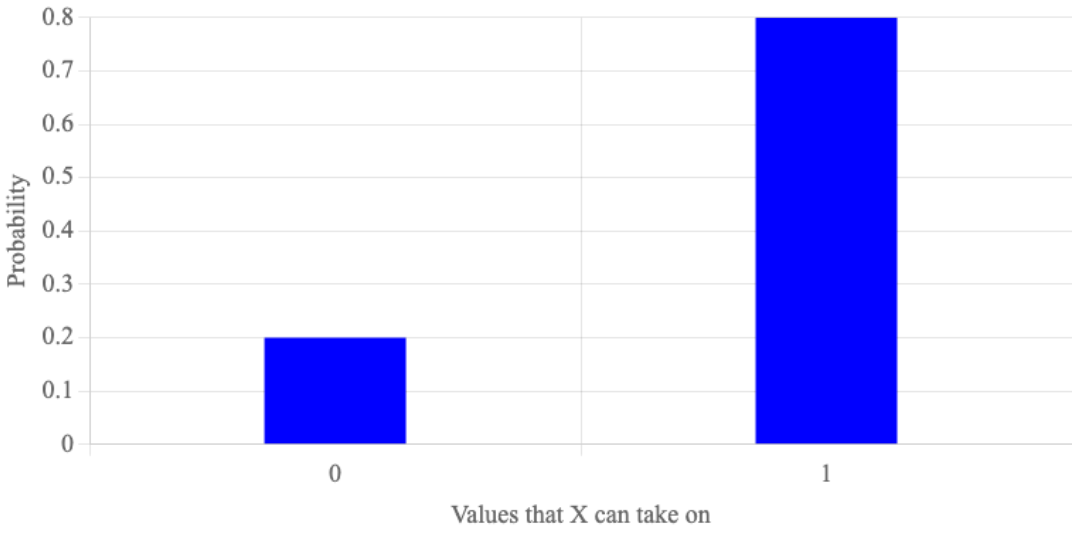
PMF (smooth): $\text{P}(X = x) = p^x (1-p)^{1-x}$

Expectation: $\text{E}[X] = p$

Variance: $\text{Var}(X) = p(1-p)$

PMF graph:

Parameter p: 0.80



Binomial Random Variable

Notation: $X \sim \text{Bin}(n, p)$

Description: Number of "successes" in n identical, independent experiments each with probability of success p.

Parameters: $n \in \{0, 1, \ldots\}$, the number of experiments.
$p \in [0, 1]$, the probability that a single experiment gives a "success".

Support: $x \in \{0, 1, \ldots, n\}$

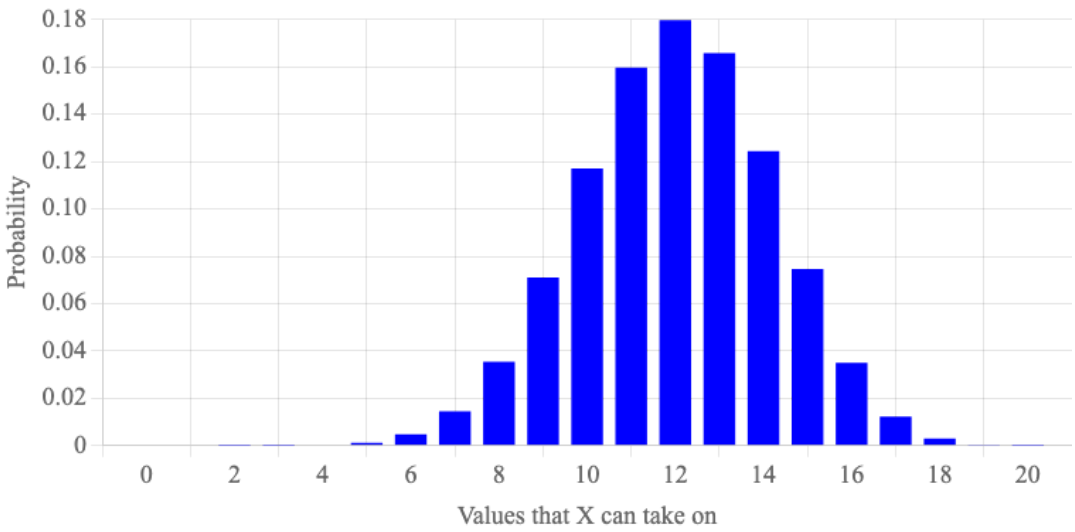
PMF equation: $$\text{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Expectation: $\text{E}[X] = n \cdot p$

Variance: $\text{Var}(X) = n \cdot p \cdot (1-p)$

PMF graph:

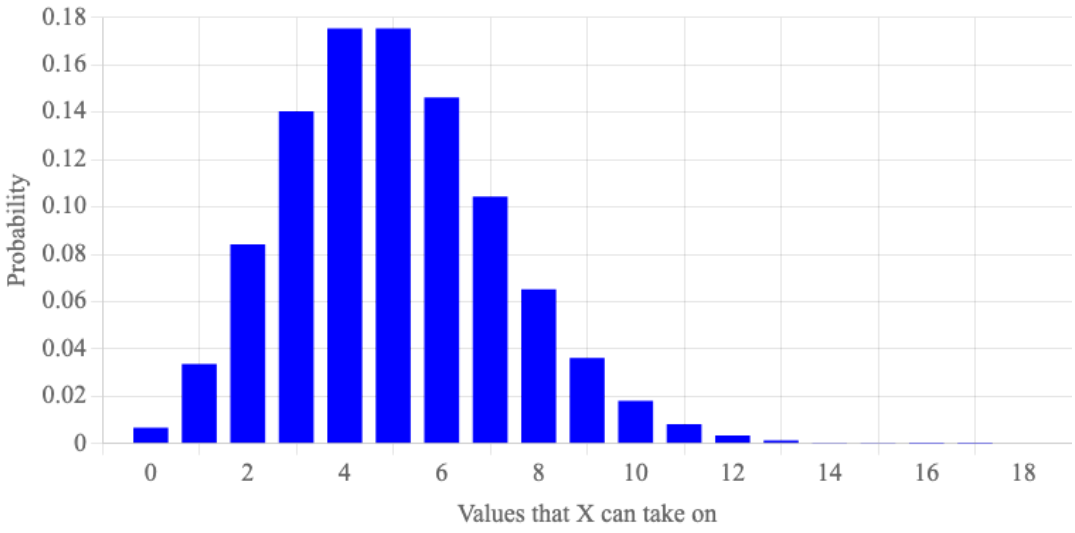
Parameter n: 20 Parameter p: 0.60



Poisson Random Variable

Notation:	$X \sim \text{Poi}(\lambda)$
Description:	Number of events in a fixed time frame if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.
Parameters:	$\lambda \in \{0, 1, \ldots\}$, the constant average rate.
Support:	$x \in \{0, 1, \ldots\}$
PMF equation:	$\text{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
Expectation:	$\text{E}[X] = \lambda$
Variance:	$\text{Var}(X) = \lambda$
PMF graph:	

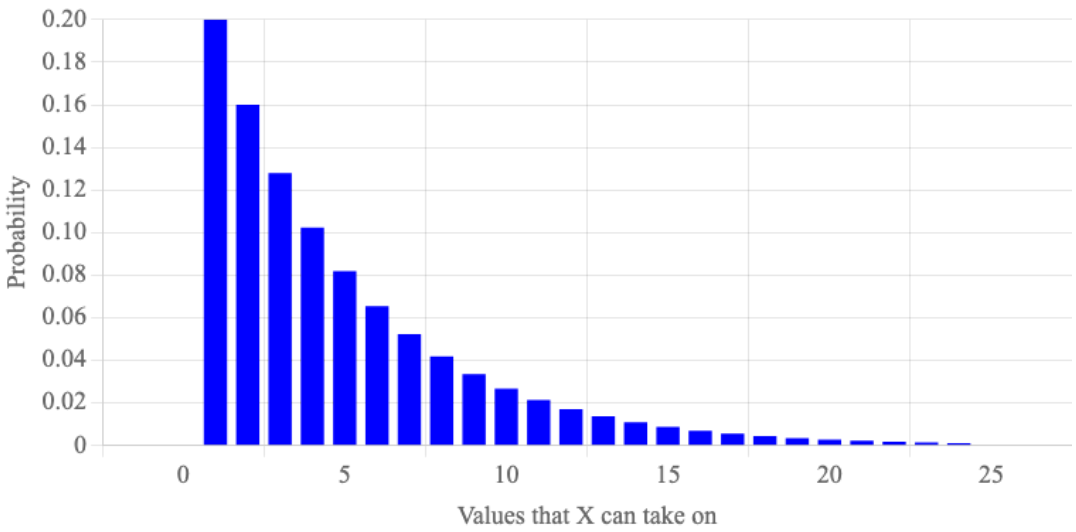
Parameter λ: 5



Geometric Random Variable

Notation:	$X \sim \text{Geo}(p)$
Description:	Number of experiments until a success. Assumes independent experiments each with probability of success p.
Parameters:	$p \in [0, 1]$, the probability that a single experiment gives a "success".
Support:	$x \in \{1, \ldots, \infty\}$
PMF equation:	$\text{P}(X = x) = (1-p)^{x-1}p$
Expectation:	$\text{E}[X] = \frac{1}{p}$
Variance:	$\text{Var}(X) = \frac{1-p}{p^2}$
PMF graph:	

Parameter p: 0.20

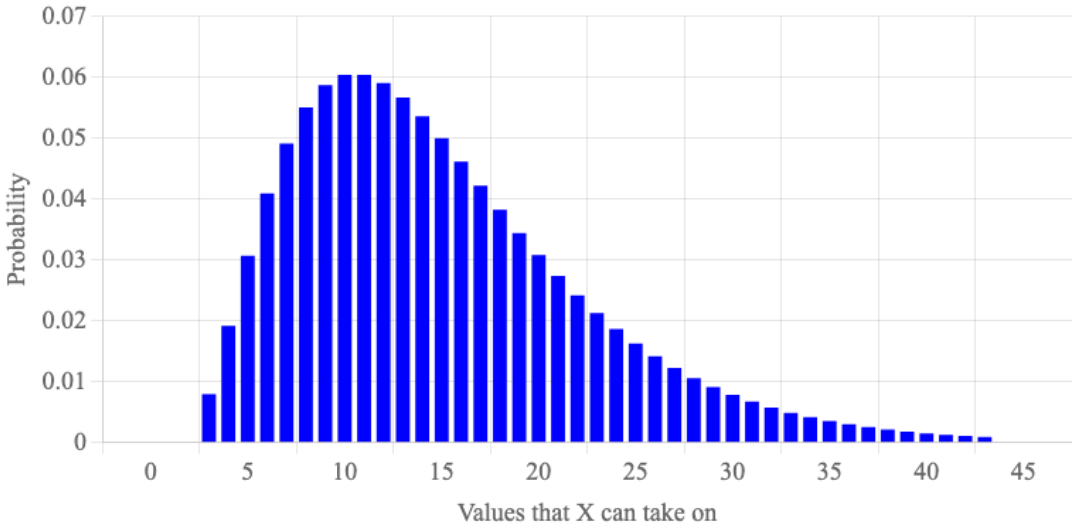


Negative Binomial Random Variable

Notation:	$X \sim \text{NegBin}(r, p)$
Description:	Number of experiments until r successes. Assumes each experiment is independent with probability of success p.
Parameters:	$r > 0$, the number of success we are waiting for. $p \in [0, 1]$, the probability that a single experiment gives a "success".
Support:	$x \in \{r, \dots, \infty\}$
PMF equation:	$\text{P}(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$
Expectation:	$\text{E}[X] = \frac{r}{p}$
Variance:	$\text{Var}(X) = \frac{r \cdot (1-p)}{p^2}$

PMF graph:

Parameter r: 3 Parameter p: 0.20



Continuous Random Variables

Uniform Random Variable

Notation:	$X \sim \text{Uni}(\alpha, \beta)$
Description:	A continuous random variable that takes on values, with equal likelihood, between α and β
Parameters:	$\alpha \in \mathbb{R}$, the minimum value of the variable. $\beta \in \mathbb{R}$, $\beta > \alpha$, the maximum value of the variable.
Support:	$x \in [\alpha, \beta]$
PDF equation:	$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{else} \end{cases}$

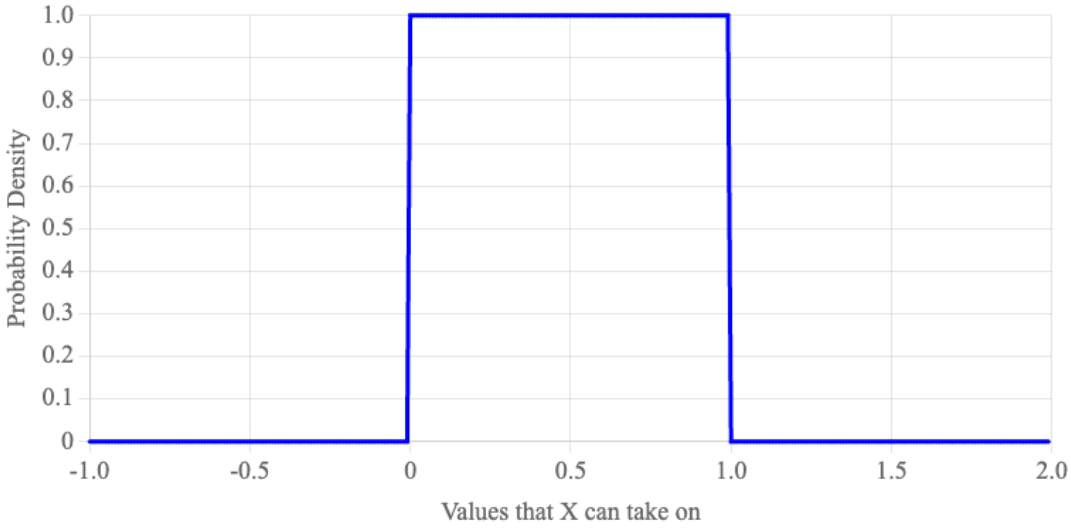
CDF equation: $F(x) = \begin{cases} \frac{x-\alpha}{\beta-\alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x < \alpha \\ 1 & \text{for } x > \beta \end{cases}$

Expectation: $\mathrm{E}[X] = \frac{1}{2}(\alpha + \beta)$

Variance: $\mathrm{Var}(X) = \frac{1}{12}(\beta - \alpha)^2$

PDF graph:

Parameter α: 0 Parameter β: 1



Exponential Random Variable

Notation: $X \sim \mathrm{Exp}(\lambda)$

Description: Time until next events if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.

Parameters: $\lambda \in \{0, 1, \ldots\}$, the constant average rate.

Support: $x \in \mathbb{R}^+$

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

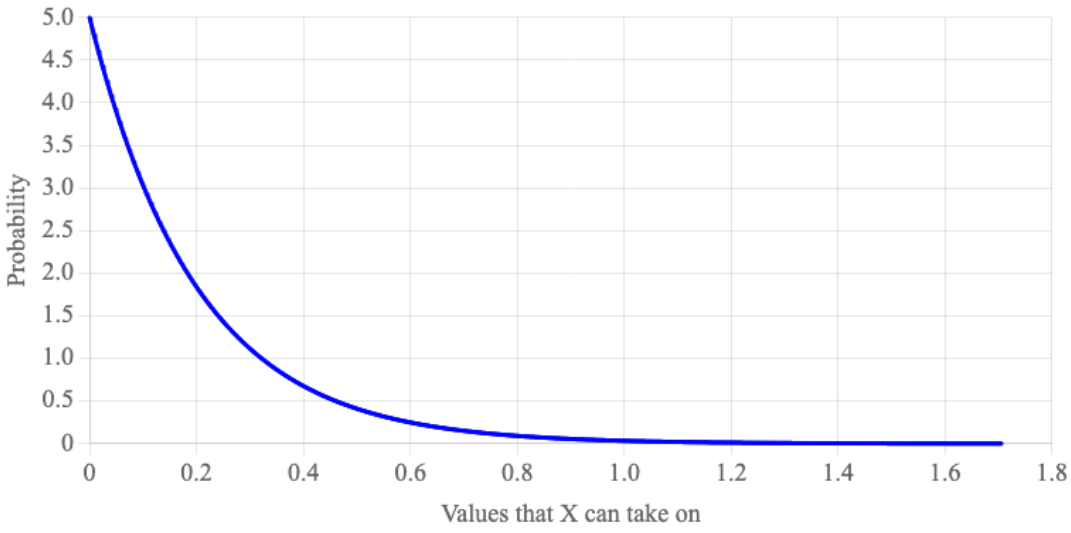
CDF equation: $F(x) = 1 - e^{-\lambda x}$

Expectation: $\mathrm{E}[X] = 1/\lambda$

Variance: $\mathrm{Var}(X) = 1/\lambda^2$

PDF graph:

Parameter λ: 5



Normal (aka Gaussian) Random Variable

Notation: $X \sim \mathrm{N}(\mu, \sigma^2)$

Description: A common, naturally occuring distribution.

Parameters: $\mu \in \mathbb{R}$, the mean.
$\sigma^2 \in \mathbb{R}$, the variance.

Support: $x \in \mathbb{R}$

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

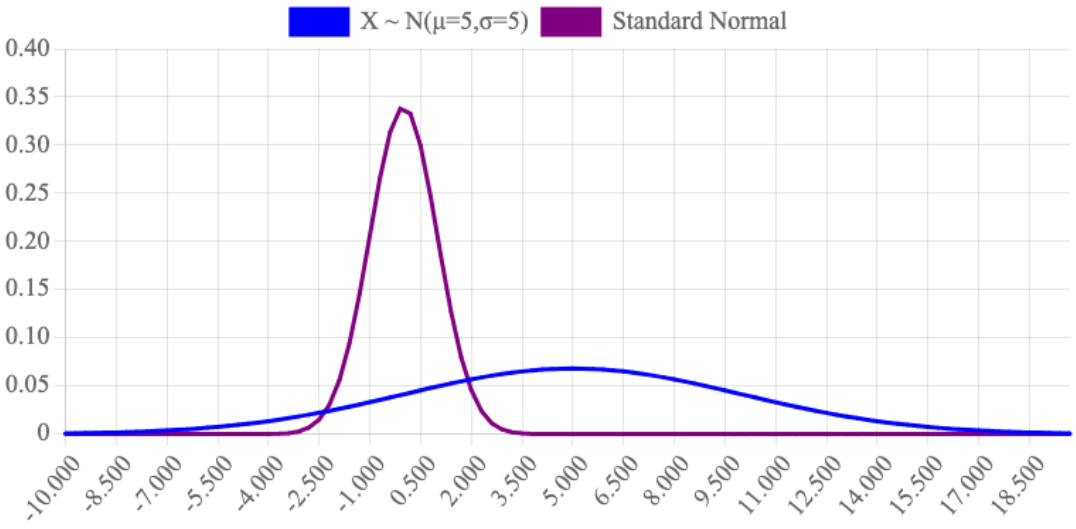
CDF equation: $F(x) = \phi(\frac{x-\mu}{\sigma})$ Where ϕ is the CDF of the standard normal

Expectation: $\mathrm{E}[X] = \mu$

Variance: $\mathrm{Var}(X) = \sigma^2$

PDF graph:

Parameter μ: 5 Parameter σ: 5



Beta Random Variable

Notation: $X \sim \mathrm{Beta}(a, b)$

Description: A belief distribution over the value of a probability p from a Binomial distribution after observing $a - 1$ successes and $b - 1$ fails.

Parameters: $a > 0$, the number successes + 1
$b > 0$, the number of fails + 1

Support: $x \in [0, 1]$

PDF equation: $f(x) = B \cdot x^{a-1} \cdot (1-x)^{b-1}$

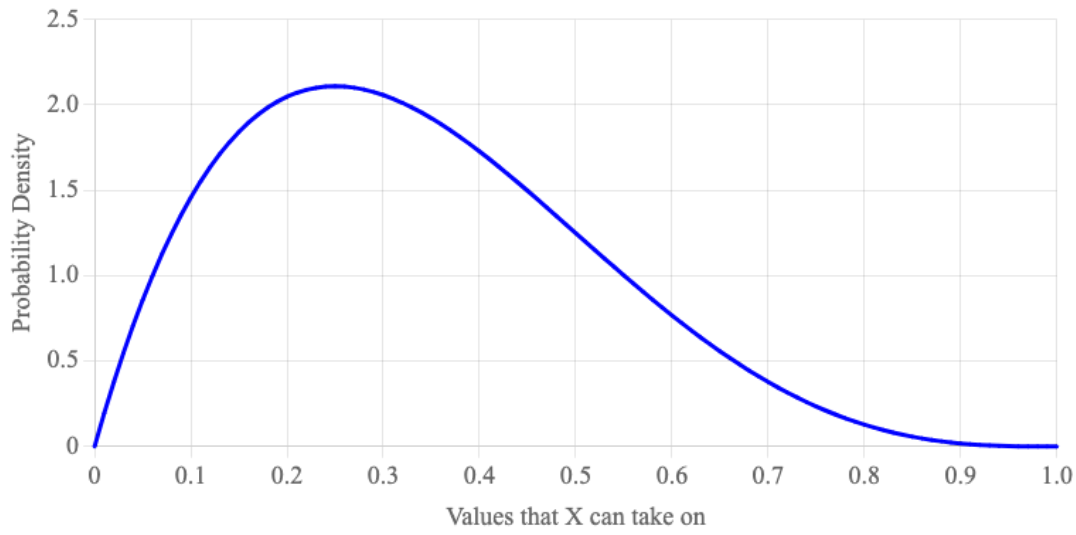
CDF equation: No closed form

Expectation: $\mathrm{E}[X] = \frac{a}{a+b}$

Variance: $\mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$

PDF graph:

Parameter a: 2 Parameter b: 4



Calculators

Factorial Calculator $n!$

n 10

```
factorial(n)
```

Combination Calculator $\binom{n}{k}$

n 10

k 6

```
combination(n,k)
```

Phi Calculator, $\Phi(x)$

x 0.7

```
phi(x)
```

Inverse Phi Calculator, $\Phi^{-1}(y)$

y 0.7

```
inverse_phi(y)
```

Norm CDF Calculator

x 0.0

mu 0

std 1

```
norm.cdf(x, mu, std)
```

Beta CDF Calculator

x 0.5

a 3

b 4

```
beta.cdf(x, a, b)
```

Counting

Although you may have thought you had a pretty good grasp on the notion of counting at the age of three, it turns out that you had to wait until now to learn how to really count. Aren't you glad you took this class now?! But seriously, counting is like the foundation of a house (where the house is all the great things we will do later in this book, such as machine learning). Houses are awesome. Foundations, on the other hand, are pretty much just concrete in a hole. But don't make a house without a foundation. It won't turn out well.

1. Counting with Steps

Definition: Step Rule of Counting (aka Product Rule of Counting)

If an experiment has two parts, where the first part can result in one of m outcomes and the second part can result in one of n outcomes regardless of the outcome of the first part, then the total number of outcomes for the experiment is $m \cdot n$.

Rewritten using set notation, the Step Rule of Counting states that if an experiment with two parts has an outcome from set A in the first part, where $|A| = m$, and an outcome from set B in the second part (where the number of outcomes in B is the same regardless of the outcome of the first part), where $|B| = n$, then the total number of outcomes of the experiment is $|A||B| = m \cdot n$.

Simple Example: Consider a hash table with 100 buckets. Two arbitrary strings are independently hashed and added to the table. How many possible ways are there for the strings to be stored in the table? Each string can be hashed to one of 100 buckets. Since the results of hashing the first string do not impact the hash of the second, there are 100 * 100 = 10,000 ways that the two strings may be stored in the hash table.

Peter Norvig, the author of the cannonical text book "Artificial Intelligence" made the following compelling point on why computer scientists need to know how to count. To start, lets set a baseline for a really big number: The number of atoms in the observable universe, often estimated to be around 10 to the 80th power (10^{80}). There certainly are a lot of atoms in the universe. As a leading expert said,

> "Space is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the chemist, but that's just peanuts to space." - Douglas Adams

This number is often used to demonstrate tasks that computers will never be able to solve. Problems can quickly grow to an absurd size, and we can understand why using the Step Rule of Counting.

There is an art project to display every possible picture. Surely that would take a long time, because there must be many possible pictures. But how many? We will assume the color model known as True Color, in which each pixel can be one of $2^{24} \approx 17$ million distinct colors.

How many distinct pictures can you generate from (a) a smart phone camera shown with 12 million pixels, (b) a grid with 300 pixels, and (c) a grid with just 12 pixels?

(a) 12 million pixels

(b) 300 pixels

(c) 12 pixels

Answer: We can use the steps rule of counting. An image can be created one pixel at a time, step by step. Each time we chose a pixel you can select its color out of 17 million choices. An array of n pixels produces (17 million)^n different pictures. (17 million)$^{12} \approx 10^{86}$, so the tiny 12-pixel grid produces a million times more pictures than the number of atoms in the universe! How about the 300 pixel array? It can produce 10^{2167} pictures. You may think the number of atoms in the universe is big, but that's just peanuts to the number of pictures in a 300-pixel array. And 12M pixels? $10^{86696638}$ pictures.

Example: Unique states of Go

For example a Go board has 19 × 19 points where a user can place a stone. Each of the points can be empty or occupied by black or white stone. By the Step Rule of Counting, we can compute the number of unique board configurations.

In go there are 19x19 points. Each point can have a black stone, white stone, or no stone at all.

Here we are going to construct the board one point at a time, step by step. Each time we add a point we have a unique choice where we can decide to make the point one of three options: {Black, White, No Stone}. Using this construction we can apply the Step Rule of Counting. If there was only one point, there would be three unique board configurations. If there were four points you would have $3 \cdot 3 \cdot 3 \cdot 3 = 81$ unique combinations. In Go there are $3^{(19\times19)} \approx 10^{172}$ possible board positions. The way we constructed our board didn't take into account which ones were illegal by the rules of Go. It turns out that "only" about 10^{170} of those positions are legal. That is about the square of the number of atoms in the universe. In otherwords: if there was another universe of atoms for every single atom, only then would there be as many atoms in the universe as there are unique configurations of a Go board.

As a computer scientist this sort of result can be very important. While computers are powerful, an algorithm which needed to store each configuration of the board would not be a reasonable approach. No computer can store more information than atoms in the universe squared!

The above argument might leave you feeling like some problems are incredibly hard as a result of the product rule of counting. Let's take a moment to talk about how the product rule of counting can help! Most logrithmic time algorithms leverage this principle.

Imagine you are building a machine learning system that needs to learn from data and you want to synthetically generate 10 million unique data points for it. How many steps would you need to encode to get to 10 million? Assuming that at each step you have a binary choice, the number of unique data points you produce will be 2^n by the Steps Rule of counting. If we chose n such that $\log_2 10,000,000 < n$. You would only need to encode $n = 24$ binary decisions.

Example: Rolling two dice. Two 6-sided dice , with faces numbered 1 through 6, are rolled. How many possible outcomes of the roll are there?

Solution: Note that we are not concerned with the total value of the two die ("die" is the singular form of "dice"), but rather the set of all explicit outcomes of the rolls. Since the first die can come up with 6 possible values and the second die similarly can have 6 possible values (regardless of what appeared on the first die), the total number of potential outcomes is 36 (= 6 × 6). These possible outcomes are explicitly listed below as a series of pairs, denoting the values rolled on the pair of dice:

(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)

2. Counting with **or**

If you want to consider the total number of unique outcomes, when outcomes can come from source A **or** source B, then the equation you use depends on whether or not there are outcomes which are both in A and B. If not, you can use the simpler "Mutually Exclusive Counting" rule. Otherwise you need to use the slightly more involved Inclusiong Exclusion rule.

Definition: Mutually Exclusive Counting

If the outcome of an experiment can either be drawn from set A or set B, and sets A and B, where none of the outcomes in set A is the same as the any of the outcomes in set B (called mutual exclusion), then there are $|A \text{ or } B| = |A| + |B|$ possible outcomes of the experiment.

Example: Sum of Routes. A route finding algorithm needs to find routes from Nairobi to Dar Es Salaam. It finds routes that either pass through Mt Kilimanjaro or Mombasa. There are 20 routes that pass through Mt Kilimanjaro, 15 routes that pass through Mombasa and 0 routes which pass through both Mt Kilimanjaro and Mombasa. How many routes are there total?

Solution: Routes can come from either Mt Kilimanjaro **or** Mombasa. The two sets of routes are mutually exclusive as there are zero routes which are in both groups. As such the total number of routes is addition: 20 + 15 = 35.

If you can show that two groups are mutually exclusive counting becomes simple addition. Of course not all sets are mutually exclusive. In the example above, imagine there had been a single route which went through both Mt Kilimanjaro and Mombasa. We would have double counted that routes because it would be included in both the sets. If sets are not mutually exclusive, counting the **or** is still addition, we simply need to take into account any double counting.

Definition: Inclusion Exclusion Counting

If the outcome of an experiment can either be drawn from set A or set B, and sets A and B may potentially overlap (i.e., it is not the case that A and B are mutually exclusive), then the number of outcomes of the experiment is $|A \text{ or } B| = |A| + |B| - |A \text{ and } B|$.

Note that the Inclusion-Exclusion Principle generalizes the Sum Rule of Counting for arbitrary sets A and B. In the case where $A \text{ and } B = \emptyset$, the Inclusion-Exclusion Principle gives the same result as the Sum Rule of Counting since $|A \text{ and } B| = 0$.

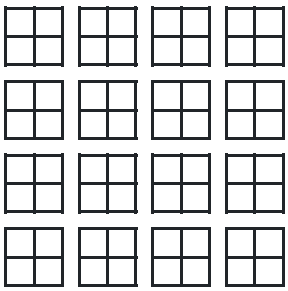
Example: An 8-bit string (one byte) is sent over a network. The valid set of strings recognized by the receiver must either start with "01" or end with "10". How many such strings are there?

Solution: The potential bit strings that match the receiver's criteria can either be the 64 strings that start with "01" (since that last 6 bits are left unspecified, allowing for $2^6 = 64$ possibilities) or the 64 strings that end with "10" (since the first 6 bits are unspecified). Of course, these two sets overlap, since strings that start with "01" and end with "10" are in both sets. There are 2^4 = 16 such strings (since the middle 4 bits can be arbitrary). Casting this description into corresponding set notation, we have: $|A|$ = 64, $|B|$ = 64, and $|A \text{ and } B|$ = 16, so by the Inclusion-Exclusion Principle, there are 64 + 64 − 16 = 112 strings that match the specified receiver's criteria.

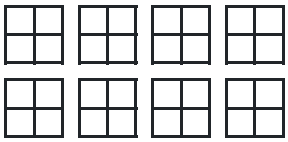
3. Overcounting and Correcting

One strategy for counting is sometimes to overcount a solution and the correct for any duplicates. This is especially common when it is easier to generate all outcomes under some relaxed assumptions, or someone introduces contraints. If you can argue that you have over-counted each element the same multiple number of times, you can simply correct by using division. If you can count exactly how many elements were over-counted you can correct using subtraction.

As a simple example to demonstrate the point, lets revisit the problem of generating all images, but this time lets just have 4 pixels (2x2) and each pixel can only be **blue** or white. How many unique images are there? Generating any image is a four step process where you chose each pixel one at a time. Since each pixel has two choices there are $2^4 = 16$ unique images (they are not exactly Picasso — but hey its 4 pixels):



Now lets say we add in new "constraint" that we only want to accept pictures which have an odd number of pixels turned blue. There are two ways of getting to the answer. You could start out with the original 16 and work out that you need to subtract off 8 images that have either 0, 2 or 4 blue pixels (which is easier to work out after the next chapter). Or you could have counted up using Mutually Exclusive Counting: there are 4 ways of making an image with 1 pixel and 4 ways of making an image with 3. Both approaches lead to the same answer, 8.



Next lets add a much harder constraint: mirror indistinction. If you can flip any image horizontally to create another, they are no longer considered unique. For example these two both show up in our set of 8 odd-blue pixel images, but they are now considered to be the same (the are indistinct after a horizontal flip):

How many images have an odd number of pixels taking into account mirror indistinction? The answer is that for each unique image with odd numbers of blue pixels, under this new constraint, you have counted it twice: itself and its horizonal flip. To convince yourself that each image has been counted *exactly* twice you can look at all of the example in the set of 8 imagines with odd number of blue pixels. Each image is next to one which is indistinct after a horizontal flip. Since each image was counted exactly twice in the set of 8, we can divide by two to get the updated count. If we list them out we can confirm that there are 8/2=4 images left after this last constraint:

Applying any math (counting included) to novel contexts can be as much an art as it is a science. In the next chapter we will build a useful toolset from the basic first principles of counting by steps, and counting by "or".

Combinatorics

Counting problems can be approached from the basic building blocks described in the first section: Counting. However some counting problems are so ubiquitous in the world of probability that it is worth knowing a few higher level counting abstractions. When solving problems, if you can find the analogy from these canonical examples you can build off of the corresponding combinatorics formulas:

1. Permutations of Distinct Objects
2. Permutations with Indistinct Objects
3. Combinations with Distinct Objects
4. Bucketing with Distinct Objects
5. Bucketing with Indistinct Objects
6. Bucketing into Fixed Sized Containers

While these are by no means the only common counting paradigms, it is a helpful set.

1. Permutations of Distinct Objects

> ***Definition***: Permutation Rule
>
> A permutation is an ordered arrangement of n distinct object. Those n objects can be permuted in $n \cdot (n–1) \cdot (n–2) \cdots 2 \cdot 1 = n!$ ways.

This changes slightly if you are permuting a subset of distinct objects, or if some of your objects are indistinct. We will handle those cases shortly! Note that unique is a synonym for distinct.

Example: How many unique orderings of characters are possible for the string "BAYES"? ***Solution***: Since the order of characters is important, we are considering all permutations of the 5 distinct characters B, A, Y, E, and S: $5! = 120$. Here is the full list:

BAYES, BAYSE, BAEYS, BAESY, BASYE, BASEY, BYAES, BYASE, BYEAS, BYESA, BYSAE, BYSEA, BEAYS, BEASY, BEYAS, BEYSA, BESAY, BESYA, BSAYE, BSAEY, BSYAE, BSYEA, BSEAY, BSEYA, ABYES, ABYSE, ABEYS, ABESY, ABSYE, ABSEY, AYBES, AYBSE, AYEBS, AYESB, AYSBE, AYSEB, AEBYS, AEBSY, AEYBS, AEYSB, AESBY, AESYB, ASBYE, ASBEY, ASYBE, ASYEB, ASEBY, ASEYB, YBAES, YBASE, YBEAS, YBESA, YBSAE, YBSEA, YABES, YABSE, YAEBS, YAESB, YASBE, YASEB, YEBAS, YEBSA, YEABS, YEASB, YESBA, YESAB, YSBAE, YSBEA, YSABE, YSAEB, YSEBA, YSEAB, EBAYS, EBASY, EBYAS, EBYSA, EBSAY, EBSYA, EABYS, EABSY, EAYBS, EAYSB, EASBY, EASYB, EYBAS, EYBSA, EYABS, EYASB, EYSBA, EYSAB, ESBAY, ESBYA, ESABY, ESAYB, ESYBA, ESYAB, SBAYE, SBAEY, SBYAE, SBYEA, SBEAY, SBEYA, SABYE, SABEY, SAYBE, SAYEB, SAEBY, SAEYB, SYBAE, SYBEA, SYABE, SYAEB, SYEBA, SYEAB, SEBAY, SEBYA, SEABY, SEAYB, SEYBA, SEYAB

Example: a smart-phone has a 4-digit passcode. Suppose there are 4 smudges over 4 digits on the screen. How many distinct passcodes are possible?
Solution: Since the order of digits in the code is important, we should use permutations. And since there are exactly four smudges we know that each number in the passcode is distinct. Thus, we can plug in the permutation formula: 4! = 24.

2. Permutations of Indistinct Objects

Definition: Permutations of In-Distinct Objects

Generally when there are n objects and:

n_1 are the same (indistinguishable) and
n_2 are the same and
...
n_r are the same, then the number of distinct permutations is:

$$\text{Number of unique orderings} = \frac{n!}{n_1!n_2!\cdots n_r!}$$

Example: How many distinct bit strings can be formed from three 0's and two 1's?
Solution: 5 total digits would give 5! permutations. But that is assuming the 0's and 1's are distinguishable (to make that explicit, let's give each one a subscript). Here are the $3! \cdot 2! = 12$ different ways that we could have arrived at the identical string "01100" if we thought of each 0 and 1 as unique.

$0_1\ 1_0\ 1_1\ 0_2\ 0_3$
$0_1\ 1_0\ 1_1\ 0_3\ 0_2$
$0_2\ 1_0\ 1_1\ 0_1\ 0_3$
$0_2\ 1_0\ 1_1\ 0_3\ 0_1$
$0_3\ 1_0\ 1_1\ 0_1\ 0_2$
$0_3\ 1_0\ 1_1\ 0_2\ 0_1$

$0_1\ 1_1\ 1_0\ 0_2\ 0_3$
$0_1\ 1_1\ 1_0\ 0_3\ 0_2$
$0_2\ 1_1\ 1_0\ 0_1\ 0_3$
$0_2\ 1_1\ 1_0\ 0_3\ 0_1$
$0_3\ 1_1\ 1_0\ 0_1\ 0_2$
$0_3\ 1_1\ 1_0\ 0_2\ 0_1$

Since identical digits are indistinguishable, all the listed permutations are the same. For any given permutation, there are 3! ways of rearranging the 0's and 2! ways of rearranging the 1's (resulting in indistinguishable strings). We have over-counted. Using the formula for permutations of indistinct objects, we can correct for the over-counting:

$$\text{Total} = \frac{5!}{3! \cdot 2!} = \frac{120}{6 \cdot 2} = 10$$

Example: How many *distinct* orderings of characters are possible for the string "MISSISSIPPI"?

In the case of the string "MISSISSIPPI", we should separate the characters into four distinct groups of indistinct characters: one "M", four "I"s, four "S"s, and two "P"s. The number of distinct orderings as:

$$\frac{11!}{1!4!4!2!} = 34,650$$

Example: Consider the 4-digit passcode smart-phone from before. How many distinct passcodes are possible if there are 3 smudges over 3 digits on the screen?

Solution: One of 3 digits is repeated, but we don't know which one. We can solve this by making three cases, one for each digit that could be repeated (each with the same number of permutations). Let A, B, C represent the 3 digits, with C repeated twice. We can initially pretend the two C's are distinct $[A, B, C_1, C_2]$. Then each case will have 4! permutations: However, then we need to eliminate the double-counting of the permutations of the identical digits (one A, one B, and two C's):

$$\frac{4!}{2! \cdot 1! \cdot 1!}$$

Adding up the three cases for the different repeated digits gives

$$3 \cdot \frac{4!}{2! \cdot 1! \cdot 1!} = 3 \cdot 12 = 36$$

Part B: What if there are 2 smudges over 2 digits on the screen?

Solution: There are two possibilities: 2 digits used twice each, or 1 digit used 3 times, and other digit used once.

$$\frac{4!}{2! \cdot 2!} + 2 \cdot \frac{4!}{3! \cdot 1!} = 6 + (2 \cdot 4) = 6 + 8 = 14$$

You can use the power of computers to enumerate all permutations. Here is sample python which uses the built in itertools library:

```
>>> import itertools

# get all 4! = 24 permutations of 1,2,3,4 as a list:
>>> list(itertools.permutations([1,2,3,4]))
[(1, 2, 3, 4), (1, 2, 4, 3), (1, 3, 2, 4), (1, 3, 4, 2), (1, 4, 2, 3), (1, 4, 3, 2),
(2, 1, 3, 4), (2, 1, 4, 3), (2, 3, 1, 4), (2, 3, 4, 1), (2, 4, 1, 3), (2, 4, 3, 1),
(3, 1, 2, 4), (3, 1, 4, 2), (3, 2, 1, 4), (3, 2, 4, 1), (3, 4, 1, 2), (3, 4, 2, 1),
(4, 1, 2, 3), (4, 1, 3, 2), (4, 2, 1, 3), (4, 2, 3, 1), (4, 3, 1, 2), (4, 3, 2, 1)]

# get all 3!/2! = 3 unique permutations of 1,1,2 as a set:
>>> set(itertools.permutations([1,1,2]))
{(1, 2, 1), (2, 1, 1), (1, 1, 2)}
```

3. Combinations of Distinct Objects

> ***Definition***: Combinations
>
> A combination is an unordered selection of r objects from a set of n objects. If all objects are distinct, and objects are not "replaced" once selected, then the number of ways of making the selection is:
>
> $$\text{Number of unique selections} = \frac{n!}{r!(n-r)!} = \binom{n}{r}$$

Here are all the $10 = \binom{5}{3}$ ways of choosing three items from a list of 5 unique numbers:

```
# Get all ways of chosing three numbers from [1,2,3,4,5]
>>> list(itertools.combinations([1,2,3,4,5], 3))
[(1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 3, 4), (1, 3, 5), (1, 4, 5), (2, 3, 4), (2, 3,
5), (2, 4, 5), (3, 4, 5)]
```

Notice how order doesn't matter. Since (1, 2, 3) is in the set of combinations, we don't also include (3, 2, 1) as this is considered to be the same selection. Note that this formula does not work if some of the objects are indistinct form one another.

How did we get the formula $\frac{n!}{r!(n-r!)}$? Consider this general way to select r unordered objects from a set of n objects, e.g., “7 choose 3”:

1. First consider permutations of all n objects. There are $n!$ ways to do that.
2. Then select the first r in the permutation. There is one way to do that.
3. Note that the order of r selected objects is irrelevant. There are $r!$ ways to permute them. The selection remains unchanged.
4. Note that the order of $(n-r)$ unselected objects is irrelevant. There are $(n-r)!$ ways to permute them. The selection remains unchanged.

$$\text{Total} = \frac{n!}{r! \cdot (n-r)!} = \binom{n}{r}$$

Example: In the Hunger Games, how many ways are there of choosing 2 villagers from district 12, which has a population of 8,000?

Solution: This is a straightforward combinations problem. $\binom{8000}{2}$ =31,996,000.

Part A: How many ways are there to select 3 books from a set of 6? **Solution:** If each of the books are distinct, then this is another straightforward combination problem. There are $\binom{6}{3} = \frac{6!}{3!3!} = 20$ ways.

Part B: How many ways are there to select 3 books if there are two books that should not both be chosen together? For example, if you are choosing 3 out of 6 probability books, don't choose both the 8th and 9th edition of the Ross textbook). \paragraph{Solution:} This problem is easier to solve if we split it up into cases. Consider the following three different cases:
Case 1: Select the 8th Ed. and 2 other non-9th Ed.: There are $\binom{4}{2}$ ways of doing so.
Case 2: Select the 9th Ed. and 2 other non-8th Ed.: There are $\binom{4}{2}$ ways of doing so.
Case 3: Select 3 from the books that are neither the 8th nor the 8th edition: There are $\binom{4}{3}$ ways of doing so.
Using our old friend the Sum Rule of Counting, we can add the cases:

$$\text{Total} = 2 \cdot \binom{4}{2} + \binom{4}{3} = 16$$

Alternatively, we could have calculated all the ways of selecting 3 books from 6, and then subtract the "forbidden" ones (i.e., the selections that break the constraint).
Forbidden Case: Select 8th edition and 9th edition and 1 other book. There are $\binom{4}{1}$ ways of doing so (which equals 4). Total = All possibilities - forbidden = 20 - 4 = 16 Two different ways to get the same right answer!

4. Bucketing with Distinct Objects

In this section we are going to be counting the many different ways that we can think of stuffing elements into containers. (It turns out that Jacob Bernoulli was into voting and ancient Rome. And in ancient Rome they used urns for ballot boxes. For this reason many books introduce this through counting ways to put balls in urns.) This "bucketing" or "group assignment" process is a useful metaphor for many counting problems.

The most common case that we will want to consider is when all of the items you are putting into buckets are distinct. In that case you can think of bucketing as a series of steps, and employ the step rule of counting. The first step? You put the first distinct item into a bucket (there are number-of-buckets ways to do this). Second step? You put the second distinct item into a bucket (again, there are number-of-buckets ways to do this).

Bucketing Distinct Items:

Suppose you want to place n distinguishable items into r containers. The number of ways of doing so is:

$$r^n$$

You have n steps (place each item) and for each item you have r choices

Problem: Say you want to put 10 distinguishable balls into 5 urns (No! Wait! Don't say that! Not urns!). Okay, fine. No urns. Say we are going to put 10 different strings into 5 buckets of a hash table. How many possible ways are there of doing this?

Solution: You can think of this as 10 independent experiments each with 5 outcomes. Using our rule for bucketing with distinct items, this comes out to 5^{10}.

5. Bucketing with Indistinct Objects

While the previous example allowed us to put n distinguishable objects into r distinct groups, the more interesting problem is to work with n indistinguishable objects.

Divider Method:

Suppose you want to place n indistinguishable items into r containers. The divider method works by imagining that you are going to solve this problem by sorting two types of objects, your n original elements and $(r-1)$ dividers. Thus, you are permuting $n+r-1$ objects, n of which are same (your elements) and $r-1$ of which are same (the dividers). Thus the total number of outcomes is:

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

Part A: Say you are a startup incubator and you have \$10 million to invest in 4 companies (in \$1 million increments). How many ways can you allocate this money?
Solution: This is just like putting 10 balls into 4 urns. Using the Divider Method we get:

$$\text{Total ways} = \binom{10+4-1}{10} = \binom{13}{10} = 286$$

This problem is analogous to solving the integer equation $x_1 + x_2 + x_3 + x_4 = 10$, where x_i represents the investment in company i such that $x_i \geq 0$ for all $i = 1, 2, 3, 4$.

Part B: What if you know you want to invest at least \$3 million in Company 1?
Solution: There is one way to give \$3 million to Company 1. The number of ways of investing the remaining money is the same as putting 7 balls into 4 urns.

$$\text{Total Ways} = \binom{7+4-1}{7} = \binom{10}{7} = 120$$

This problem is analogous to solving the integer equation $x_1 + x_2 + x_3 + x_4 = 10$, where $x_1 \geq 3$ and $x_2, x_3, x_4 \geq 0$. To translate this problem into the integer solution equation that we can solve via the divider method, we need to adjust the bounds on x_1 such that the problem becomes $x_1 + x_2 + x_3 + x_4 = 7$, where x_i is defined as in Part A.

Part C: What if you don't have to invest all \$10 M? (The economy is tight, say, and you might want to save your money.)
Solution: Imagine that you have an extra company: yourself. Now you are investing \$10 million in 5 companies. Thus, the answer is the same as putting 10 balls into 5 urns.

$$\text{Total} = \binom{10+5-1}{10} = \binom{14}{10} = 1001$$

This problem is analogous to solving the integer equation $x_1 + x_2 + x_3 + x_4 + x_5 = 10$, such that $x_i \geq 0$ for all $i = 1, 2, 3, 4, 5$.

6. Bucketing into Fixed Sized Containers

Bucketing into Fixed Sized Containers:

If n objects are distinct, then the number of ways of putting them into r groups of objects, such that group i has size n_i, and $\sum_{i=1}^{r} n_i = n$, is:

$$\frac{n!}{n_1! n_2! \cdots n_r!} = \binom{n}{n_1, n_2, \ldots, n_r}$$

where $\binom{n}{n_1, n_2, \ldots, n_r}$ is special notation called the multinomial coefficient.

You may have noticed that this is the exact same formula as "Permutations With Indistinct Objects". There is a deap parallel. One way to imagine assigning objects into their groups would be to imagine the groups themselves as objects. You have one object per "slot" in a group. So if there were two slots in group 1, three slots in group 2, and one slot in group 3 you could have six objects (1, 1, 2, 2, 2, 3). Each unique permutation can be used to make a unique assignment.

Problem:

Company Camazon has 13 distinct new servers that they would like to assign to 3 datacenters, where Datacenter A, B, and C have 6, 4, and 3 empty server racks, respectively. How many different divisions of the servers are possible?

Solution: This is a straightforward application of our multinomial coefficient representation. Setting $n_1 = 6, n_2 = 4, n_3 = 3$, $\binom{13}{6,4,3} = 60,060$.

Another way to do this problem would be from first principles of combinations as a multipart experiment. We first select the 6 servers to be assigned to Datacenter A, in $\binom{13}{6}$ ways. Now out of the 7 servers remaining, we select the 4 servers to be assigned to Datacenter B, in $\binom{7}{4}$ ways. Finally, we select the 3 servers out of the remaining 3 servers, in $\binom{3}{3}$ ways. By the Product Rule of Counting, the total number of ways to assign all servers would be $\binom{13}{6}\binom{7}{4}\binom{3}{3} = \frac{13!}{6!4!3!} = 60,060$.

Definition of Probability

What does it mean when someone makes a claim like "the probability that you find a pearl in an oyster is 1 in 5,000?" or "the probability that it will rain tomorrow is 52%?

1. Events and Experiments

When we speak about probabilities, there is always an implied context, which we formally call the "experiment". For example: flipping two coins is something that probability folks would call an experiment. In order to precisely speak about probability, we must first define two sets: the set of all possible outcomes of an experiment, and the subset that we consider to be our event (what is a set?).

Definition: Sample Space, S

A Sample Space is set of all possible outcomes of an experiment. For example:

- Coin flip: S = {Heads, Tails}
- Flipping two coins: S = {(H, H), (H, T), (T, H), (T, T)}
- Roll of 6-sided die: S = {1, 2, 3, 4, 5, 6}
- The number of emails you receive in a day: $S = \{x | x \in \mathbb{Z}, x \geq 0\}$ (non-neg. ints)
- YouTube hours in a day: $S = \{x | x \in \mathbb{R}, 0 \leq x \leq 24\}$

Definition: Event, E

An Event is some subset of S that we ascribe meaning to. In set notation ($E \subseteq S$).For example:

- Coin flip is heads: E = {Heads}
- Greater than 1 head on 2 coin flips = {(H, H), (H, T), (T, H)}
- Roll of die is 3 or less: E = {1, 2, 3}
- You receive less than 20 emails in a day: $E = \{x | x \in Z, 0 \leq x < 20\}$ (non-neg. ints)
- Wasted day ($\geq$ 5 YouTube hours): $E = \{x | x \in R, 5 \leq x \leq 24\}$

Events can be represented as capital letters such as E or F.

[todo] In the world of probability, events are binary: they either happen or they don't.

2. Definition of Probability

It wasn't until the 20th century that humans figured out a way to precisely define what the word probability means:

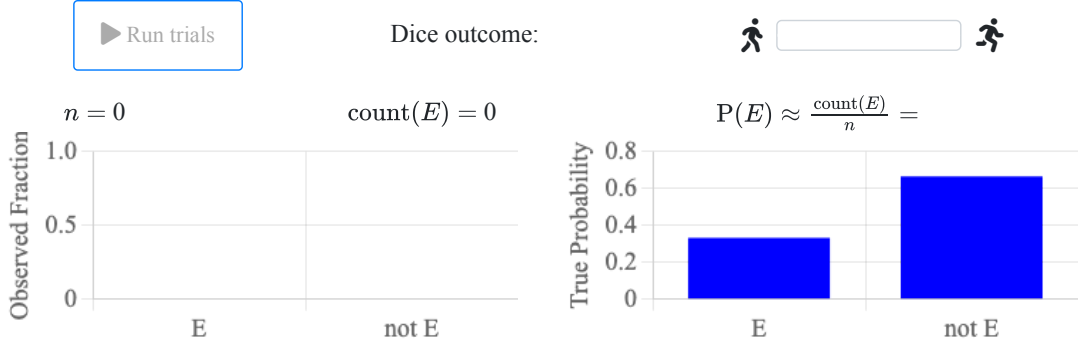
$$\mathrm{P}(\text{Event}) = \lim_{n \to \infty} \frac{\text{count}(\text{Event})}{n}$$

In English this reads: lets say you perform n trials of an "experiment" which could result in a particular "Event" occuring. The probability of the event occuring, $\mathrm{P}(\text{Event})$, is the ratio of trials that result in the event, written as $\text{count}(\text{Event})$, to the number of trials performed, n. In the limit, as your number of trials approaches infinity, the ratio will converve to the true probability. People also apply other semantics to the concept of a probability. One common meaning ascribed is that $\mathrm{P}(E)$ is a measure of the chance of event E occurring.

Example: Probability in the limit

Here we use the definition of probability to calculate the probability of event E, rolling a "5" or a "6" on a fair six-sided dice. Hit the "Run trials" button to start running trials of the experiment "roll dice". Notice how $\mathrm{P}(E)$, converges to $2/6$ or 0.33 repeating.

Event E: Rolling a 5 or 6 on a six-sided dice.



Measure of uncertainty: It is tempting to think of probability as representing some natural randomness in the world. That might be the case. But perhaps the world isn't random. I propose a deeper way of thinking about probability. There is so much that we as humans don't know, and probability is our robust language for expressing my belief that an event will happen given my limited knowledge. This interpretation acknowledges that your own uncertainty of an event. Perhaps if you knew the position of every water molecule, you could perfectly predict tomorrow's weather. But we don't have such knowledge and as such we use probability to talk about the chance of rain tomorrow given the information that we have access to.

Origins of probabilities: The different interpretations of probability are reflected in the many origins of probabilities that you will encounter in the wild (and not so wild) world. Some probabilities are calculated analytically using mathematical proofs. Some probabilities are calculated from data, experiments or simulations. Some probabilities are just made up to represent a belief. Most probabilities are generated from a combination of the above. For example, someone will make up a prior belief, that belief will be mathematically updated using data and evidence. Here is an example of calculating a probability from data:

Probabilities and simulations: Another way to compute probabilities is via simulation. For some complex problems where the probabilities are too hard to compute analytically you can run simulations using your computer. If your simulations generate believable trials from the sample space, then the probability of an event E is approximately equal to the fraction of simulations that produced an outcome from E. Again, by the definition of probability, as your number of simulations approaches infinity, the estimate becomes more accurate.

Probabilities and percentages: You might hear people refer to a probability as a percent. That the probability of rain tomorrow is 32%. The proper way to state this would be to say that 0.32 is the probability of rain. Percentages are simply probabilities multiplied by 100. "percent" is latin for "out of one hundred".

Problem: Use the definition of probability to approximate the answer to the question: "What is the probability a new-born elephant child is male?" Contrary to what you might think the gender outcomes of a newborn elephant are not equally likely between male and female. You have data from a report in Animal Reproductive Science which states that 3,070 elephants were born in Myanmar of which 2,180 were male [1]. Humans also don't have a 50/50 sex ratio at birth [2].

Answer: The Experiment is: A single elephant birth in Myanmar.
The sample space is the set of possible sexes assigned at birth, {Male, Female, Intersex}.
E is the event that a new-born elephan child is male, which in set notation is the subset {Male} of the sample space. The outcomes are not equally likely.

By the definition of probability, the ratio of trials that result in the event to the number of trials will tend to our desired probability:

$$\begin{aligned}\text{P}(\text{Born Male}) &= \text{P}(E) \\ &= \lim_{n\to\infty} \frac{\text{count}(E)}{n} \\ &\approx \frac{2,180}{3,070} \\ &\approx 0.710\end{aligned}$$

Since 3,000 is quite a bit less than infinity, this is an approximation. It turns out, however, to be a rather good one. A few important notes: there is no garuntee that our estimate applies to elephants outside Myanmar. Later in the class we will develop language for "how confident we can be in a number like 0.71 after 3,000 trials?" Using tools from later in class we can say that we have 98% confidence that the true probability is within 0.02 of 0.710.

3. Axioms of Probability

Here are some basic truths about probabilities that we accept as axioms:

Axiom 1: $0 \leq \mathrm{P}(E) \leq 1$	All probabilities are numbers between 0 and 1.
Axiom 2: $\mathrm{P}(S) = 1$	All outcomes must be from the Sample Space.
Axiom 3: If E and F are mutually exclusive, then $\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F)$	The probability of "or" for mutually exclusive events

These three axioms are formally called the Kolmogorov axioms and they are considered to be the foundation of probability theory. They are also useful identities!

You can convince yourself of the first axiom by thinking about the math definition of probability. As you perform trials of an experiment it is not possible to get more events than trials (thus probabilities are less than 1) and its not possible to get less than 0 occurrences of the event (thus probabilities are greater than 0). The second axiom makes sense too. If your event is the sample space, then each trial must produce the event. This is sort of like saying; the probability of you eating cake (event) if you eat cake (sample space that is the same as the event) is 1. The third axiom is more complex and in this textbook we dedicate an entire chapter to understanding it: Probability of or. It applies to events that have a special property called "mutual exclusion": the events do not share any outcomes.

These axioms have great historical significance. In the early 1900s it was not clear if probability was somehow different than other fields of math -- perhaps the set of techniques and systems of proofs from other fields of mathematics couldn't apply. Kolmogorov's great success was to show to the world that the tools of mathematics did infact apply to probability. From the foundation provided by this set of axioms mathematicians built the edifice of probability theory.

4. Provable Identities

We often refer to these as corollaries that are directly provable from the three axioms given above.

Identity 1: $\mathrm{P}(E^{\mathrm{C}}) = 1 - \mathrm{P}(E)$	The probability of event E not happening
Identity 2: If $E \subseteq F$, then $\mathrm{P}(E) \leq \mathrm{P}(F)$	Events which are subsets

This first identity is especially useful. For any event, you can calculate the probability of the event *not* occuring which we write in probability notation as E^{C}, if you know the probability of it occuring -- and vice versa. We can also use this identity to show you what it looks like to prove a theorem in probability.

Proof: $\mathrm{P}(E^{\mathrm{C}}) = 1 - \mathrm{P}(E)$

$\mathrm{P}(S) = \mathrm{P}(E \text{ or } E^{\mathrm{C}})$	E or E^{C} covers every outcome in the sample space
$\mathrm{P}(S) = \mathrm{P}(E) + \mathrm{P}(E^{\mathrm{C}})$	Events E and E^{C} are mututally exclusive
$1 = \mathrm{P}(E) + \mathrm{P}(E^{\mathrm{C}})$	Axiom 2 of probability
$\mathrm{P}(E^{\mathrm{C}}) = 1 - \mathrm{P}(E)$	By re-arranging

Equally Likely Outcomes

Some sample spaces have equally likely outcomes. We like those sample spaces, because there is a way to calculate probability questions about those sample spaces simply by counting. Here are a few examples where there are equally likely outcomes:

- Coin flip: S = {Head, Tails}
- Flipping two coins: S = {(H, H), (H, T), (T, H), (T, T)}
- Roll of 6-sided die: S = {1, 2, 3, 4, 5, 6}

Because every outcome is equally likely, and the probability of the sample space must be 1, we can prove that each outcome must have probability:

$$\mathrm{P}(\text{an outcome}) = \frac{1}{|S|}$$

Where |S| is the size of the sample space, or, put in other words, the total number of outcomes of the experiment. Of course this is only true in the special case where every outcome has the same likelihood.

Definition: Probability of Equally Likely Outcomes

If S is a sample space with equally likely outcomes, for an event E that is a subset of the outcomes in S:

$$\mathrm{P}(E) = \frac{\text{number of outcomes in } E}{\text{number of outcomes in } S} = \frac{|E|}{|S|}$$

There is some art form to setting up a problem to calculate a probability based on the equally likely outcome rule. (1) The first step is to explicitly define your sample space and to argue that all outcomes in your sample space are equally likely. (2) Next, you need to count the number of elements in the sample space and (3) finally you need to count the size of the event space. The event space must be all elements of the sample space that you defined in part (1). The first step leaves you with a lot of choice! For example you can decide to make indistinguishable objects distinct, as long as your calculation of the size of the event space makes the exact same assumptions.

Example: What is the probability that the sum of two die is equal to 7?

Buggy Solution: You could define your sample space to be all the possible sum values of two die (2 through 12). However this sample space fails the "equally likely" test. You are not equally likely to have a sum of 2 as you are to have a sum of 7.

Solution: Consider the sample space from the previous chapter where we thought of the die as distinct and enumerated all of the outcomes in the sample space. The first number is the roll on die 1 and the second number is the roll on die 2. Note that (1, 2) is distinct from (2, 1). Since each outcome is equally likely, and the sample space has exactly 36 outcomes, the likelihood of any one outcome is $\frac{1}{36}$. Here is a visualization of all outcomes:

(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)

The event (sum of dice is 7) is the subset of the sample space where the sum of the two dice is 7. Each outcome in the event is highlighted in **blue**. There are 6 such outcomes: (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1). Notice that (1, 6) is a different outcome than (6, 1). To make the outcomes equally likely we had to

make the die distinct.

$$\begin{aligned} \mathrm{P}(\text{Sum of two dice is } 7) &= \frac{|E|}{|S|} && \text{Since outcomes are equally likely} \\ &= \frac{6}{36} = \frac{1}{6} && \text{There are 6 outcomes in the event} \end{aligned}$$

Interestingly, this idea also applies to continuous sample spaces. Consider the sample space of all the outcomes of the computer function "random" which produces a real valued number between 0 and 1, where all real valued numbers are equally likely. Now consider the event E that the number generated is in the range [0.3 to 0.7]. Since the sample space is equally likely, $\mathrm{P}(E)$ is the ratio of the size of E to the size of S. In this case $\mathrm{P}(E) = \frac{0.4}{1} = 0.4$.

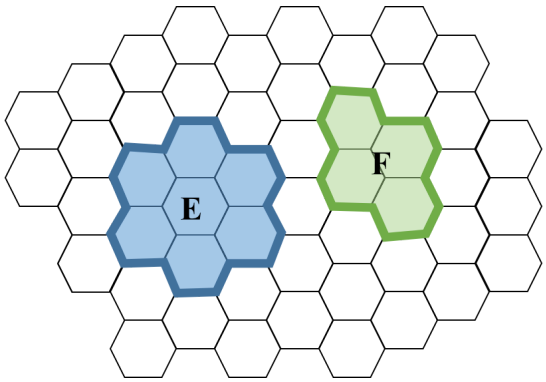
Probability of **or**

The equation for calculating the probability of either event E ***or*** event F happening, written $\mathrm{P}(E \text{ or } F)$ or equivalently as $\mathrm{P}(E \cup F)$, is deeply analogous to counting the size of two sets. As in counting, the equation that you can use depends on whether or not the events are "mutually exclusive". If events are mutually exclusive, it is very straightforward to calculate the probability of either event happening. Otherwise, you need the more complex "inclusion exclusion" formula.

1. Mutually exclusive events

Two events: E, F are considered to be mutually exclusive (in set notation $E \cap F = \emptyset$) if there are no outcomes that are in both events (recall that an event is a set of outcomes which is a subset of the sample space). In English, mutually exclusive means that two events can't both happen.

Mutual exclusion can be visualized. Consider the following visual sample space where each outcome is a hexagon. The set of all the fifty hexagons is the full sample space:



Example of two events: E, F, which are mutually exclusive.

Both events E and F are subsets of the same sample space. Visually, we can note that the two sets do not overlap. They are mutually exclusive: there is no outcome that is in both sets.

2. Prob of **or** for mutually exclusive events

Definition: Probability of **or** for mututally exclusive events
If two events: E, F are mutually exclusive then the probability of E ***or*** F occuring is:

$$\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F)$$

This property applies regardless of how you calculate the probability of E or F. Moreover, the idea extends to more than two events. Lets say you have n events $E_1, E_2, \ldots E_n$ where each event is mutually exclusive of one another (in other words, no outcome is in more than one event). Then:

$$\mathrm{P}(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_n) = \mathrm{P}(E_1) + \mathrm{P}(E_2) + \cdots + \mathrm{P}(E_n) = \sum_{i=1}^{n} \mathrm{P}(E_i)$$

You may have noticed that this is one of the axioms of probability. Though it might seem intuitive, it is one of three rules that we accept without proof.

Caution: Mutual exclusion only makes it easier to calculate the probability of E or F not other ways of combining events, such as E and F.

At this point we know how to compute the probability of the "or" of events if and only if they have the mutual exclusion property. What if they don't?

3. Prob of **or** for non-mutually exclusive events

Unfortunately, not all events are mutually exclusive. If you want to calculate $\mathrm{P}(E \text{ or } F)$ where the events E and F are ***not*** mutually exclusive you can ***not*** simply add the probabilities. As a simple sanity check, consider the event E: getting heads on a coin flip, where $\mathrm{P}(E) = 0.5$. Now imagine the sample space S, getting either a heads or a tails on a coin flip. These events are not mutually exclusive (the outcome heads is in both). If you incorrectly assumed they were mutually exclusive and tried to calculate $\mathrm{P}(E \text{ or } S)$ you would get this buggy derivation:

Buggy derivation: Incorrectly assuming mutual exclusion

Calculate the probability of E, getting an even number on a dice role (2, 4 or 6), or F, getting three or less (1, 2, 3) on the same dice role.

$$\begin{aligned} \mathrm{P}(E \text{ or } F) &= \mathrm{P}(E) + \mathrm{P}(F) && \text{Incorrectly assumes mutual exclusion} \\ &= 0.5 + 0.5 && \text{substitute the probabilities of } E \text{ and } S \\ &= 1.0 && \text{uh oh!} \end{aligned}$$

The probability can't be one since the outcome 5 is neither three or less nor even. The problem is that we double counted the probability of getting a 2, and the fix is to subtract out the probability of that doubly counted case.

What went wrong? If two events are not mutually exclusive, simply adding their probabilities double counts the probability of any outcome which is in both events. There is a formula for calculating ***or*** of two non-mutually exclusive events: it is called the "inclusion exclusion" principle.

Definition: Inclusion Exclusion principle
For any two events: E, F:

$$\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \text{ and } F)$$

This formula does have a version for more than two events, but it gets rather complex. For three events, E, F, and G the formula is:

$$\begin{aligned} \mathrm{P}(E \text{ or } F \text{ or } G) = \; & \mathrm{P}(E) + \mathrm{P}(F) + \mathrm{P}(G) \\ & - \mathrm{P}(E \text{ and } F) - \mathrm{P}(E \text{ and } G) - P(F \text{ and } G) \\ & + \mathrm{P}(E \text{ and } F \text{ and } G) \end{aligned}$$

For n events, $E_1, E_2, \ldots E_n$: build a running sum. Add all the probabilities of the events on their own. Then subtract all pairs of events. Then add all subsets of 3 events. Then subtract all subset of 4 events. Continue this process, up until n, adding the subsets if the size of subsets is odd, else subtracting them. The alternating addition and subtraction is where the name inclusion exclusion comes from. This is a complex process and you should first check if there is an easier way to calculate your probability.

Note that the inclusion exclusion principle also applies for mutually exclusive events. If two events are mutually exclusive $\mathrm{P}(E \text{ and } F) = 0$ since its not possible for both E and F to occur. As such the formula $\mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \text{ and } F)$ reduces to $\mathrm{P}(E) + \mathrm{P}(F)$.

The formulas for calculating the ***or*** of events that are not mutually exclusive often requires calculating the probability of the ***and*** of events. Learn more in the next section:

Conditional Probability

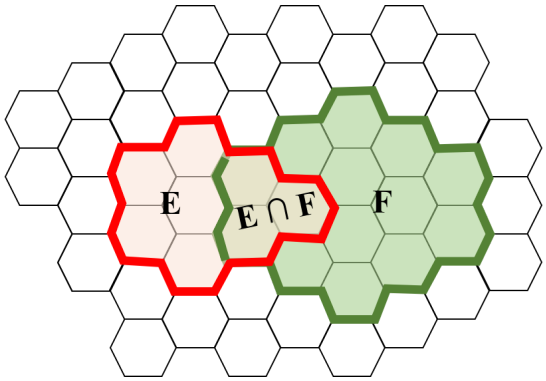
In English, a conditional probability states "what is the chance of an event E happening given that I have already observed some other event F". It is a critical idea in machine learning and probability because it allows us to update our probabilities in the face of new evidence.

When you condition on an event happening you are entering the universe where that event has taken place. Formally, once you condition on F the only outcomes that are now possible are the ones which are consistent with F. In other words your sample space will now be reduced to F. As an aside, in the universe where F has taken place, all rules of probability still hold!

Definition: Conditional Probability.
The probability of E given that (aka conditioned on) event F already happened:

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(E \text{ and } F)}{\mathrm{P}(F)}$$

Let's use a visualization to get an intuition for why the conditional probability formula is true. Again consider events E and F which have outcomes that are subsets of a sample space with 50 equally likely outcomes, each one drawn as a hexagon:



Conditioning on F means that we have entered the world where F has happened (and F, which has 14 equally likely outcomes, has become our new sample space). Given that event F has occurred, the conditional probability that event E occurs is the subset of the outcomes of E that are consistent with F. In this case we can visually see that those are the three outcomes in E and F. Thus we have the:

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(E \text{ and } F)}{\mathrm{P}(F)} = \frac{3/50}{14/50} = \frac{3}{14} \approx 0.21$$

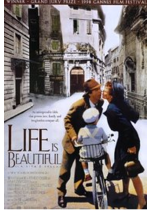
Even though the visual example (with equally likely outcome spaces) is useful for gaining intuition, conditional probability applies regardless of whether the sample space has equally likely outcomes!

1. Conditional Probability Example

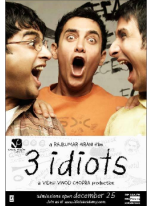
Let's use a real world example to better understand conditional probability: movie recommendation. Imagine a streaming service like Netflix wants to figure out the probability that a user will watch a movie E (for example, Life is Beautiful), based on knowing that they watched a different movie F (say Amélie). To start lets answer the simpler question, what is the probability that a user watches movie Life is Beautiful, E? We can solve this problem using the definition of probability and a dataset of movie watching [1]:

$$\begin{aligned}\mathrm{P}(E) &= \lim_{n\to\infty} \frac{\text{count}(E)}{n} \approx \frac{\#\text{ people who watched movie } E}{\#\text{ people on Netflix}} \\ &= \frac{1,234,231}{50,923,123} \approx 0.02\end{aligned}$$

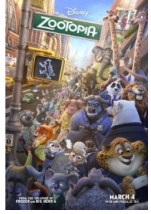
In fact we can do this for many movies E:



$\mathrm{P}(E) = 0.02$



$\mathrm{P}(E) = 0.01$



$\mathrm{P}(E) = 0.05$

$\mathrm{P}(E) = 0.09$

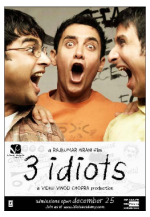
$\mathrm{P}(E) = 0.03$

Now for a more interesting question. What is the What is the probability that a user will watch the movie Life is Beautiful (E), given they watched Amelie (F)? We can use the definition of conditional probability.

$$\begin{aligned}
\mathrm{P}(E|F) &= \frac{\mathrm{P}(E \text{ and } F)}{\mathrm{P}(F)} && \text{Def of Cond Prob} \\
&\approx \frac{(\#\text{ who watched } E \text{ and } F)/(\#\text{ of people on Netflix})}{(\#\text{ who watched movie } F)/(\#\text{ people on Netflix})} && \text{Def of Prob} \\
&\approx \frac{\#\text{ of people who watched both } E \text{ and } F}{\#\text{ of people who watched movie } F} && \text{Simplifying}
\end{aligned}$$

If we let F be the event that someone watches the movie Amélie, we can now calculate $\mathrm{P}(E|F)$, the *conditional* probability that someone watches movie E:

$\mathrm{P}(E|F) = 0.09$



$\mathrm{P}(E|F) = 0.03$

$\mathrm{P}(E|F) = 0.05$

$\mathrm{P}(E|F) = 0.02$

$\mathrm{P}(E|F) = \mathbf{1.00}$

Why do some probabilities go up, some probabilities go down, and some probabilities are unchanged after we observe that the person has watched Amelie (F)? If you know someone watched Amelie, they are more likely to watch life is beautiful, and less likely to watch star wars. We have new information on the person!

2. The Conditional Paradigm

When you condition on an event you enter the universe where that event has taken place. In that new universe all the laws of probability still hold. Thus, as long as you condition consistently on the same event, every one of the tools we have learned still apply. Let's look at a few of our old friends when we condition consistently on an event (in this case G):

Name of Rule	**Original Rule**	**Rule Conditioned on G**
Axiom of probability 1	$0 \leq \mathrm{P}(E) \leq 1$	$0 \leq \mathrm{P}(E\|G) \leq 1$
Axiom of probability 2	$\mathrm{P}(S) = 1$	$\mathrm{P}(S\|G) = 1$
Axiom of probability 3	$\mathrm{P}(E \text{ or } F) = \mathrm{P}(E) + \mathrm{P}(F)$ for mutually exclusive events	$\mathrm{P}(E \text{ or } F\|G) = \mathrm{P}(E\|G) + \mathrm{P}(F\|G)$ for mutually exclusive events
Identity 1	$\mathrm{P}(E^{\mathrm{C}}) = 1 - \mathrm{P}(E)$	$\mathrm{P}(E^{\mathrm{C}}\|G) = 1 - \mathrm{P}(E\|G)$

3. Conditioning on Multiple Events

The conditional paradigm also applies to the definition of conditional probability! Again if we consistently condition on some event G occuring, the rule still holds:

$$\mathrm{P}(E|F, G) = \frac{\mathrm{P}(E \text{ and } F|G)}{\mathrm{P}(F|G)}$$

The term $\mathrm{P}(E|F,G)$ is new notation for conditioning on multiple events. You should read that term as "The probability of E occuring, given that both F and G have occured". This equation states that the definition for conditional probability of $E|F$ still applies in the universe where G has occured. Do you think that $\mathrm{P}(E|F,G)$ should be equal to $\mathrm{P}(E|F)$? The answer is: sometimes yes and sometimes no.

Independence

So far we have talked about mutual exclusion as an important "property" that two or more events can have. In this chapter we will introduce you to a second property: independence. Independence is perhaps one of the most important properties to consider! Like for mutual exclusion, if you can establish that this property applies (either by logic, or by declaring it as an assumption) it will make analytic probability calculations much easier!

Definition: Independence

Two events are said to be independent if knowing the outcome of one event does not change your belief about whether or not the other event will occur. For example, you might say that two separate dice rolls are independent of one another: the outcome of the first dice gives you no information about the outcome of the second -- and vice versa.

$$\mathrm{P}(E|F) = \mathrm{P}(E)$$

This definition is symmetric. If E is independent of F, then F is independent of E:

$$\mathrm{P}(F|E) = \mathrm{P}(F)$$

1. How to establish independence

How can you show that two or more events are independent? The default option is to show it mathematically. If you can show that $\mathrm{P}(E|F) = \mathrm{P}(E)$ then you have proven that the two events are indepedent. When working with probabilities that come from data, very few things will exactly match the mathematical definition of independence. That can happen for two reasons: first, events that are calculated from data or simulation are not perfectly precise and it can be impossible to know if a discreptancy between $\mathrm{P}(E)$ and $\mathrm{P}(E|F)$ is due to innacuracy in estimating probabilities, or dependence of events. Second, in our complex world many things actually influence each other, even if just a tiny amount. Despite that we often make the wrong, but useful, independence assumption. Since independence makes it so much easier for humans and machines to calculate composite probabilities, you may declare the events to be independent. It could mean your resulting calculation is slightly incorrect -- but this "modelling assumption" might make it feasible to come up with a result.

Independence is a property which is often "assumed" if you think it is reasonable that one event is unlikely to influence your belief that the other will occur (or if the influence is negligable). Let's worth through a few examples to better understand:

2. Conditional Independence

We saw earlier that the laws of probability still held if you consistently conditioned on an event. While the rules stay the same, the independence property might change. Events that were dependent can become independent when conditioning on an event. Events that were independent can become dependent.

Probability of **and**

The probability of the ***and*** of two events, say E and F, written $\mathrm{P}(E \text{ and } F)$, is the probability of both events happening. You might see equivalent notations $\mathrm{P}(EF)$, $\mathrm{P}(E \cap F)$ and $\mathrm{P}(E, F)$ to mean the probability of and. How you calculate the probability of event E and event F happening depends on whether or not the events are "independent". In the same way that mutual exclusion makes it easy to calculate the probability of the ***or*** of events, independence is a property that makes it easy to calculate the ***and*** of events.

1. Independent Events

If events are **independent** then calculating the probability of ***and*** becomes simple multiplication:

> ***Definition***: Probability of ***and*** for independent events.
> If two events: E, F are independent then the probability of E ***and*** F occuring is:
>
> $$\mathrm{P}(E \text{ and } F) = \mathrm{P}(E) \cdot \mathrm{P}(F)$$
>
> This property applies regardless of how the probabilities of E and F were calculated and whether or not the events are mutually exclusive.
>
> The independence principle extends to more than two events. For n events $E_1, E_2, \ldots E_n$ that are ***mutually*** independent of one another -- the independence equation also holds for all subsets of the events.
>
> $$\mathrm{P}(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_n) = \prod_{i=1}^{n} \mathrm{P}(E_i)$$

We can prove this equation by combining the definition of conditional probability and the definition of independence.

Proof: If E is independent of F then $\mathrm{P}(E \text{ and } F) = \mathrm{P}(E) \cdot \mathrm{P}(F)$

$$\begin{aligned}
\mathrm{P}(E|F) &= \frac{\mathrm{P}(E \text{ and } F)}{\mathrm{P}(F)} && \text{Definition of conditional probability} \\
\mathrm{P}(E) &= \frac{\mathrm{P}(E \text{ and } F)}{\mathrm{P}(F)} && \text{Definition of independence} \\
\mathrm{P}(E \text{ and } F) &= \mathrm{P}(E) \cdot \mathrm{P}(F) && \text{Rearranging terms}
\end{aligned}$$

See the chapter on independence to learn about when you can assume that two events are independent

2. Dependent Events

Events which are not independent are called ***dependent*** events. How can you calculate the **and** of dependent events? If your events are mutually exclusive you might be able to use a technique called DeMorgan's law, which we cover in a latter chapter. For the probability of and in dependent events there is a direct formula called the chain rule which can be directly derived from the definition of conditional probability:

> ***Definition***: The chain rule.
> The formula in the definition of conditional probability can be re-arranged to derive a general way of calculating the probability of the ***and*** of any two events:
>
> $$\mathrm{P}(E \text{ and } F) = \mathrm{P}(E|F) \cdot \mathrm{P}(F)$$

Of course there is nothing special about E that says it should go first. Equivalently:

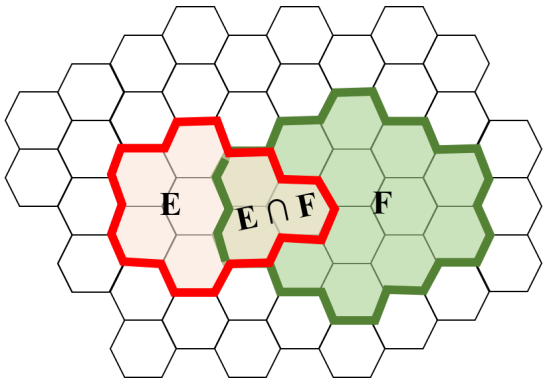
$$\mathrm{P}(E \text{ and } F) = \mathrm{P}(F \text{ and } E) = \mathrm{P}(F|E) \cdot \mathrm{P}(E)$$

We call this formula the "chain rule." Intuitively it states that the probability of observing events E ***and*** F is the probability of observing F, multiplied by the probability of observing E, given that you have observed F. It generalizes to more than two events:

$$\begin{aligned}\mathrm{P}(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_n) =& \mathrm{P}(E_1) \cdot \mathrm{P}(E_2|E_1) \cdot \mathrm{P}(E_3|E_1 \text{ and } E_2) \ldots \\ & \mathrm{P}(E_n|E_1 \ldots E_{n-1})\end{aligned}$$

Law of Total Probability

An astute person once observed that when looking at a picture, like the one we say for conditional probability:



that event E can be thought of as having two parts, the part that is in F, $(E \text{ and } F)$, and the part that isn't, $(E \text{ and } F^{\text{C}})$. This is true because F and F^{C} are (a) mutually exclusive sets of outcomes which (b) together cover the entire sample space. After further investigation this proved to be mathematically true, and there was much rejoicing:

$$\text{P}(E) = \text{P}(E \text{ and } F) + \text{P}(E \text{ and } F^{\text{C}})$$

This observation proved to be particularly useful when it was combined with the chain rule and gave rise to a tool so useful, it was given the big name, law of total probability.

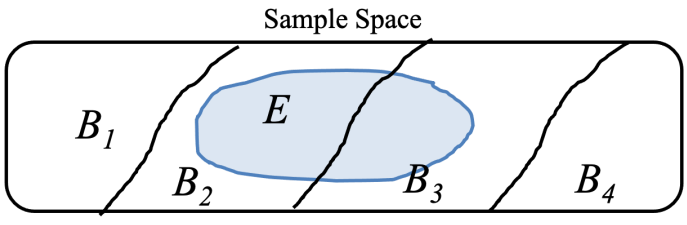
The Law of Total Probability
If we combine our above observation with the chain rule, we get a very useful formula:

$$\text{P}(E) = \text{P}(E|F)\,\text{P}(F) + \text{P}(E|F^{\text{C}})\,\text{P}(F^{\text{C}})$$

There is a more general version of the rule. If you can divide your sample space into any number of mutually exclusive events: $B_1, B_2, \ldots B_n$ such that every outcome in sample space fall into one of those events, then:

$$\begin{aligned} \text{P}(E) &= \sum_{i=1}^{n} \text{P}(E \text{ and } B_i) && \text{Extension of our observation} \\ &= \sum_{i=1}^{n} \text{P}(E|B_i)\,\text{P}(B_i) && \text{Using chain rule on each term} \end{aligned}$$

We can build intuition for the general version of the law of total probability in a similar way. If we can divide a sample space into a set of several mutually exclusive sets (where the or of all the sets covers the entire sample space) then any event can be solved for by thinking of the likelihood of the event and each of the mutually exclusive sets.



In the image above, you could compute $\text{P}(E)$ to be equal to $\text{P}\left[(E \text{ and } B_1) \text{ or } (E \text{ and } B_2) \ldots\right]$. Of course this is worth mentioning because there are many real world cases where the sample space can be discretized into several mutual exclusive events. As an example, if you were thinking about the probability of the location of an object on earth, you could discretize the area over which you are tracking into a grid.

Bayes' Theorem

Bayes' Theorem is one of the most ubiquitous results in probability for computer scientists. In a nutshell, Bayes' theorem provides a way to convert a conditional probability from one direction, say $\mathrm{P}(E|F)$, to the other direction, $\mathrm{P}(F|E)$.

Bayes' theorem is a mathematical identity which we can derive ourselves. Start with the definition of conditional probability and then expanding the and term using the chain rule:

$$\begin{aligned}\mathrm{P}(F|E) &= \frac{\mathrm{P}(F \text{ and } E)}{\mathrm{P}(E)} && \text{Def of conditional probability} \\ &= \frac{\mathrm{P}(E|F) \cdot \mathrm{P}(F)}{\mathrm{P}(E)} && \text{Substitute the chain rule for } \mathrm{P}(F \text{ and } E)\end{aligned}$$

This theorem makes no assumptions about E or F so it will apply for any two events. Bayes' theorem is exceptionally useful because it turns out to be the ubiquitous way to answer the question: "how can I update a belief about something, which is not directly observable, given evidence." This is for good reason. For many "noisy" measurements it is straightforward to estimate the probability of the noisy observation given the true state of the world. However, what you would really like to know is the conditional probability the other way around: what is the probability of the true state of the world given evidence. There are countless real world situations that fit this situation:

Example 1: Medical tests
What you want to know: Probability of a disease given a test result
What is easy to know: Probability of a test result given the true state of disease
Causality: We believe that diseases influences test results

Example 2: Student ability
What you want to know: Student knowledge of a subject given their answers
What is easy to know: Likelihood of answers given a student's knowledge of a subject
Causality: We believe that ability influences answers

Example 3: Cell phone location
What you want to know: Where is a cell phone, given noisy measure of distance to tower
What is easy to know: Error in noisy measure, given the true distance to tower
Causality: We believe that cell phone location influences distance measure

There is a pattern here: in each example we care about knowing some unobservable -- or hard to observe -- state of the world. This state of the world "causes" some easy-to-observe evidence. For example: having the flu (something we would like to know) ***causes*** a fever (something we can easily observe), not the other way around. We often call the unobservable state the "belief" and the observable state the "evidence". For that reason lets rename the events! Lets call the unobservable thing we want to know B for belief. Lets call the thing we have evidence of E for evidence. This makes is clear that Bayes' theorem allows us to calculate an updated belief given evidence: $\mathrm{P}(B|E)$

Definition: Bayes' Theorem
The most common form of Bayes' Theorem is **Bayes' Theorem Classic**:

$$\mathrm{P}(B|E) = \frac{\mathrm{P}(E|B) \cdot \mathrm{P}(B)}{\mathrm{P}(E)}$$

There are names for the different terms in the Bayes' Rule formula. The term $\mathrm{P}(B|E)$ is often called the "posterior": it is your updated belief of B after you take into account evidence E. The term $\mathrm{P}(B)$ is often called the "prior": it was your belief before seeing any evidence. The term $\mathrm{P}(E|B)$ is called the update and $\mathrm{P}(E)$ is often called the normalization constant.

There are several techniques for handling the case where the denominator is not know. One technique is to use the law of total probability to expand out the term, resulting in another formula, called **Bayes' Theorem with Law of Total Probability**:

$$\mathrm{P}(B|E) = \frac{\mathrm{P}(E|B) \cdot \mathrm{P}(B)}{\mathrm{P}(E|B) \cdot \mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}}) \cdot \mathrm{P}(B^{\mathrm{C}})}$$

Recall the law of total probability which is responsible for our new denominator:

$$\mathrm{P}(E) = \mathrm{P}(E|B) \cdot \mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}}) \cdot \mathrm{P}(B^{\mathrm{C}})$$

A common scenario for applying the Bayes' Rule formula is when you want to know the probability of something “unobservable” given an “observed” event. For example, you want to know the probability that a student understands a concept, given that you observed them solving a particular problem. It turns out it is much easier to first estimate the probability that a student can solve a problem given that they understand the concept and then to apply Bayes' Theorem. Intuitively, you can think about this as updating a belief given evidence.

1. Bayes' Theorem Applied

Sometimes the (correct) results from Bayes' Theorem can be counter intuitive. Here we work through a classic result: Bayes' applied to medical tests. We show a dynamic solution and present a visualization for understanding what is happening.

Example: Probability of a disease given a noisy test

In this problem we are going to calculate the probability that a patient has an illness given test-result for the illness. A positive test result means the test thinks the patient has the illness. You know the following information, which is typical for medical tests:

Natural % of population with illness: 13

Probability of a positive result given the patient has illness 0.92

Probability of a positive result given the patient does not have illness 0.10

The numbers in this example are from the Mamogram test for breast cancer. The seriousness of cancer underscores the potential for bayesian probability to be applied to important contexts. The natural occurence of breast cancer is 8%. The mamogram test returns a positive result 95% of the time for patients who have breast cancer. The test resturns a positive result 7% of the time for people who do not have breast cancer. In this demo you can enter different input numbers and it will reclaculate.

Answer

The probability that the patient has the illness given a positive test result is: 0.5789

Terms:
Let I be the event that the patient has the illness
Let E be the event that the test result is positive
$\mathrm{P}(I|E)$ = probability of the illness given a positive test. This is the number we want to calculate.
$\mathrm{P}(E|I)$ = probability of a positive result given illness = 0.92
$\mathrm{P}(E|I^{\mathrm{C}})$ = probability of a positive result given no illness = 0.10
$\mathrm{P}(I)$ = natural probability of the illness = 0.13

Bayes Theorem:
In this problem we know $\mathrm{P}(E|I)$ and $\mathrm{P}(E|I^{\mathrm{C}})$ but we want to know $\mathrm{P}(I|E)$. We can apply Bayes Theorem to turn our knowledge of one conditional into knowledge of the reverse.

$$\mathrm{P}(I|E) = \frac{\mathrm{P}(E|I)P(I)}{\mathrm{P}(E|I)\,\mathrm{P}(I) + \mathrm{P}(E|I^{\mathrm{C}})\,\mathrm{P}(I^{\mathrm{C}})} \qquad \text{Bayes' Theorem with Total Prob.}$$

Now all we need to do is plug values into this formula. The only value we don't explicitly have is $\mathrm{P}(I^{\mathrm{C}})$. But we can simply calculate it since $\mathrm{P}(I^{\mathrm{C}}) = 1 - \mathrm{P}(I)$. Thus:

$$P(I|E) = \frac{(0.92)(0.13)}{(0.92)(0.13) + (0.10)(1 - 0.13)} = 0.5789$$

2. Natural Frequency Intuition

One way to build intuition for Bayes Theorem is to think about "natural frequences". Let's take another approach at answer the probability question in the above example on belief of illness given a test. In this take, we are going to imagine we have a population of 1000 people. Let's think about how many of those have the illness and test positive and how many don't have the illness and test positive. This visualization is based off the numbers in the fields above. Feel free to change them!

There are many possibilities for how many people have the illness, but one very plaussible number is 1000, the number of people in our population, multiplied by the probability of the disease.

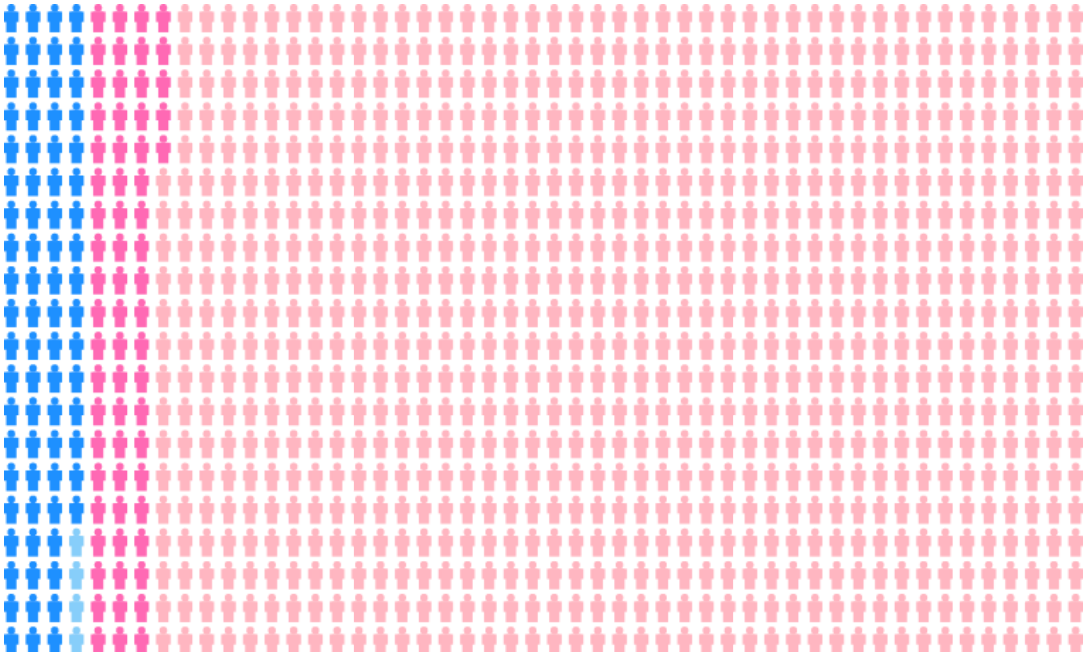
$1000 \times \mathrm{P}(\text{Illness})$ people have the illness
$1000 \times (1 - \mathrm{P}(\text{Illness}))$ people do not have the illness.

We are going to color people who **have the illness** in blue and those **without the illness** in pink (those colors do not imply gender!).

A certain number of people **with the illness will test positive** (which we will draw in Dark Blue) and a certain number of people **without the illness will test positive** (which we will draw in Dark Pink):

$1000 \times \mathrm{P}(\text{Illness}) \times \mathrm{P}(\text{Positive}|\text{Illness})$ people have the illness and test positive
$1000 \times \mathrm{P}(\text{Illness}^{\mathrm{C}}) \times \mathrm{P}(\text{Positive}|\text{Illness}^{\mathrm{C}})$ people do not have the illness and test positive.

Here is the whole population of 1000 people:



The number of people who **test positive and have the illness** is 76.
The number of people who **test positive and don't have the illness** is 65.
The total number of people who test positive is 141.

Out of the subset of people who test positive, the fraction that have the illness is 76/141 = 0.5390 which is a close approximation of the answer. If instead of using 1000 imaginary people, we had used more, the approximation would have been even closer to the actual answer (which we calculated using Bayes Theorem).

3. Bayes with the General Law of Total Probability

A classic challenge when applying Bayes' theorm is to calculate the probability of the normalization constant $\mathrm{P}(E)$ in the denominator of Bayes' Theorem. One common strategy for calculating this probability is to use the law of total probability. Our expanded version of Bayes' Theorem uses the simple version of the total law of probability: $\mathrm{P}(E) = \mathrm{P}(E|F)\,\mathrm{P}(F) + \mathrm{P}(E|F^c)\,\mathrm{P}(F^c)$. Sometimes you will want the more expanded version of the law of total probability: $\mathrm{P}(E) = \sum_i \mathrm{P}(E|B_i)\,\mathrm{P}(B_i)$. Recall that this only works if the events B_i are mutually exclusive and cover the sample space.

For example say we are trying to track a phone which could be in any one of n discrete locations and we have prior beliefs $\mathrm{P}(B_1) \ldots \mathrm{P}(B_n)$ as to whether the phone is in location B_i. Now we gain some evidence (such as a particular signal strength from a particular cell tower) that we call E and we need to update all of our probabilities to be $\mathrm{P}(B_i|E)$. We should use Bayes' Theorem!

The probability of the observation, assuming that the the phone is in location B_i, $\mathrm{P}(E|B_i)$, is something that can be given to you by an expert. In this case the probability of getting a particular signal strength given a location B_i will be determined by the distance between the cell tower and location B_i .

Since we are assuming that the phone must be in *exactly* one of the locations, we can find the probability of any of the event B_i given E by first applying Bayes' Theorem and then applying the general version of the law of total probability:

$$\begin{aligned}
\mathrm{P}(B_i|E) &= \frac{\mathrm{P}(E|B_i) \cdot \mathrm{P}(B_i)}{\mathrm{P}(E)} && \text{Bayes Theorem. What to do about } \mathrm{P}(E)\text{?} \\
&= \frac{\mathrm{P}(E|B_i) \cdot \mathrm{P}(B_i)}{\sum_{i=1}^{n} \mathrm{P}(E|B_i) \cdot \mathrm{P}(B_i)} && \text{Use General Law of Total Probability for } \mathrm{P}(E)
\end{aligned}$$

4. Unknown Normalization Constant, $\mathrm{P}(E)$

There are times when we would like to use Bayes' Theorem to update a belief, but we don't know the probability of E, $\mathrm{P}(E)$. All hope is not lost. This term is called the "normalization constant" because it is the same regardless of whether or not the event B happens. The most traditional solution is to use the law of total probability: $\mathrm{P}(E) = \mathrm{P}(E|B)\,\mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})$. Here are some other useful "tricks" for dealing with $\mathrm{P}(E)$.

We can make the normalization cancel out by calculating the ratio of $\frac{\mathrm{P}(B|E)}{\mathrm{P}(B^{\mathrm{C}}|E)}$. This fraction tells you how many times more likely it is that B will happen given E than not B:

$$\begin{aligned}
\frac{\mathrm{P}(B|E)}{\mathrm{P}(B^{\mathrm{C}}|E)} &= \frac{\frac{\mathrm{P}(E|B)\,\mathrm{P}(B)}{\mathrm{P}(E)}}{\frac{\mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})}{\mathrm{P}(E)}} && \text{Apply Bayes' Theorm to both terms} \\
&= \frac{\mathrm{P}(E|B)\,\mathrm{P}(B)}{\mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})} && \text{The term } \mathrm{P}(E) \text{ cancels}
\end{aligned}$$

We can always use the fact that either B will happen or it won't when consistently conditioned on E: $\mathrm{P}(B|E) + \mathrm{P}(B^{\mathrm{C}}|E) = 1$ to compute $\mathrm{P}(E)$. Note that this is the simply the first identity of probability, consistently conditioning:

$$\begin{aligned}
1 &= \mathrm{P}(B|E) + \mathrm{P}(B^{\mathrm{C}}|E) && \text{Either } B \text{ occurs or it doesn't} \\
1 &= \frac{\mathrm{P}(E|B)\,\mathrm{P}(B)}{\mathrm{P}(E)} + \frac{\mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})}{\mathrm{P}(E)} && \text{Apply Bayes' Theorem to both terms} \\
1 &= \frac{1}{\mathrm{P}(E)} \cdot \left[\mathrm{P}(E|B)\,\mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})\right] && \text{Factor out } 1/\mathrm{P}(E) \\
\mathrm{P}(E) &= \mathrm{P}(E|B)\,\mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}}) && \text{Rearrange terms}
\end{aligned}$$

If you look closely at the last line, you will notice that we have simply found a new way to derive the total law of probability for E. The law of total probability is truly a great way of dealing with $\mathrm{P}(E)$.

Log Probabilities

A log probability $\log \mathrm{P}(E)$ is simply the log function applied to a probability. For example if $\mathrm{P}(E) = 0.00001$ then $\log \mathrm{P}(E) = \log(0.00001) \approx -11.51$. Note that in this book, the default base is the natural base e. There are many reasons why log probabilities are an essential tool for digital probability: (a) computers can be rather limited when representing very small numbers and (b) logs have the wonderful ability to turn multiplication into addition, and computers are much faster at addition.

You may have noticed that the log in the above example produced a negative number. Recall that $\log b = c$, with the implied natural base e is the same as the statement $e^c = b$. It says that c is the exponent of e that produces b. If b is a number between 0 and 1, what power should you raise e to in order to produce b? If you raise e^0 it produces 1. To produce a number less than 1, you must raise e to a power less than 0. That is a long way of saying: if you take the log of a probability, the result will be a negative number.

$$\begin{aligned} 0 \leq \mathrm{P}(E) \leq 1 \qquad & \text{Axiom 1 of probability} \\ -\infty \leq \log \mathrm{P}(E) \leq 0 \qquad & \text{Rule for log probabilities} \end{aligned}$$

1. Products become Addition

The product of probabilities $\mathrm{P}(E)$ and $\mathrm{P}(F)$ becomes addition in logarithmic space:

$$\log(\mathrm{P}(E) \cdot \mathrm{P}(F)) = \log \mathrm{P}(E) + \log \mathrm{P}(F)$$

This is especially convenient because computers are much more efficient when adding than when multiplying. It can also make derivations easier to write. This is especially true when you need to multiply many probabilities together:

$$\log \prod_i \mathrm{P}(E_i) = \sum_i \log \mathrm{P}(E_i)$$

2. Representing Very Small Probabilities

Computers have the power to process many events and consider the probability of very unlikely situations. While computers are capable of doing all the computation, the floating point representation means that computers can not represent decimals to perfect precision. In fact, python is unable to represent any probability smaller than 2.225e-308. On the other hand the log of that same number is -307.652 is very easy for a computer to store.

Why would you care? Often in the digital world, computers are asked to reason about the probability of data, or a whole dataset. For example, perhaps your data is words and you want to reason about the probability that a given author would write these specific words. While this probability is very small (we are talking about an exact document) it might be larger than the probability that a different author would write a specific document with specific words. For these sort of small probabilities, if you use computers, you would need to use log probabilities.

Many Coin Flips

In this section we are going to consider the number of heads on n coin flips. This thought experiment is going to be a basis for much probability theory! It goes far beyond coin flips.

Say a coin comes up heads with probability p. Most coins are fair and as such come up heads with probability $p = 0.5$. There are many events for which coin flips are a great analogy that have different values of p so lets leave p as a variable. You can try simulating coins here. Note that **H** is short for Heads and **T** is short for Tails. We think of each coin as distinct:

Coin Flip Simulator

Number of flips n: 10 Probability of heads p: 0.60 New simulation

Simulator results:

T, H, T, H, T, H, H, H, H, T

Total number of heads: 6

Let's explore a few probability questions in this domain.

1. Warmups

What is the probability that all n flips are heads?

Lets say $n = 10$ this question is asking what is the probability of getting:

H, H, H, H, H, H, H, H, H, H

Each coin flip is independent so we can use the rule for probability of **and** with independent events. As such, the probability of n heads is p multiplied by itself n times: p^n. If $n = 10$ and $p = 0.6$ then the probability of n heads is around 0.006.

What is the probability that all n flips are tails?

Lets say $n = 10$ this question is asking what is the probability of getting:

T, T, T, T, T, T, T, T, T, T

Each coin flip is independent. The probability of tails on any coin flip is $1 - p$. Again, since the coin flips are independent, the probability of tails n times on n flips is $(1 - p)$ multiplied by itself n times: $(1 - p)^n$. If $n = 10$ and $p = 0.6$ then the probability of n tails is around 0.0001.

First k heads then $n - k$ tails

Lets say $n = 10$ and $k = 4$, this question is asking what is the probability of getting:

H, H, H, H, T, T, T, T, T, T

The coins are still independent! The first k heads occur with probability p^k the run of $n - k$ tails occurs with probability $(1 - p)^{n-k}$. The probability of k heads then $n - k$ tails is the product of those two terms: $p^k \cdot (1 - p)^{n-k}$

2. Exactly k heads

Next lets try to figure out the probability of exactly k heads in the n flips. Importantly we don't care where in the n flips that we get the heads, as long as there are k of them. Note that this question is different than the question of first k heads and then $n-k$ tails which requires that the k heads come first! That particular result does generate exactly k coin flips, but there are others.

There are many others! In fact any permutation of k heads and $n-k$ tails will satisfy this event. Lets ask the computer to list them all for exactly $k=4$ heads within $n=10$ coin flips. The output region is scrollable:

```
(H, H, H, H, T, T, T, T, T, T)
(H, H, H, T, H, T, T, T, T, T)
(H, H, H, T, T, H, T, T, T, T)
(H, H, H, T, T, T, H, T, T, T)
(H, H, H, T, T, T, T, H, T, T)
(H, H, H, T, T, T, T, T, H, T)
(H, H, H, T, T, T, T, T, T, H)
(H, H, T, H, H, T, T, T, T, T)
(H, H, T, H, T, H, T, T, T, T)
(H, H, T, H, T, T, H, T, T, T)
(H, H, T, H, T, T, T, H, T, T)
(H, H, T, H, T, T, T, T, H, T)
(H, H, T, H, T, T, T, T, T, H)
(H, H, T, T, H, H, T, T, T, T)
(H, H, T, T, H, T, H, T, T, T)
(H, H, T, T, H, T, T, H, T, T)
(H, H, T, T, H, T, T, T, H, T)
(H, H, T, T, H, T, T, T, T, H)
```

Exactly how many outcomes are there with $k=4$ heads in $n=10$ flips? 210. The answer can be calculated using permutations of indistinct objects:

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

The probability of exactly $k=4$ heads is the probability of the **or** of each of these outcomes. Because we consider each coin to be unique, each of these outcomes is "mutually exclusive" and as such if E_i is the outcome from the ith row,

$$\mathrm{P}(\text{exactly } k \text{ heads}) = \sum_{i=1}^{N} \mathrm{P}(E_i)$$

The next question is, what is the probability of each of these outcomes?

Here is a arbitrarily chosen outcome which satisfies the event of exactly $k=4$ heads in $n=10$ coin flips. In fact it is the one on row 128 in the list above:

```
T, H, T, T, H, T, T, H, H, T
```

What is the probability of getting the exact sequence of heads and tails in the example above? Each coin flip is still independent, so we multiply p for each heads and $1-p$ for each tails. Let E_{128} be the event of this exact outcome:

$$\mathrm{P}(E_{128}) = (1-p) \cdot p \cdot (1-p) \cdot (1-p) \cdot p \cdot (1-p) \cdot (1-p) \cdot p \cdot p \cdot (1-p)$$

If you rearrange these multiplication terms you get:

$$\begin{aligned} \mathrm{P}(E_{128}) &= p \cdot p \cdot p \cdot p \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) \\ &= p^4 \cdot (1-p)^6 \end{aligned}$$

There is nothing too special about row 128. If you chose any row, you would get k independent heads and $n-k$ independent tails. For any row i, $\mathrm{P}(E_i) = p^n \cdot (1-p)^{k-n}$. Now we are ready to calculate the probability of exactly k heads:

$$
\begin{aligned}
\mathrm{P}(\text{exactly } k \text{ heads}) &= \sum_{i=1}^{N} \mathrm{P}(E_i) && \text{Mutual Exclusion} \\
&= \sum_{i=1}^{N} p^k \cdot (1-p)^{n-k} && \text{Sub in } \mathrm{P}(E_i) \\
&= N \cdot p^k \cdot (1-p)^{n-k} && \text{Sum } N \text{ times} \\
&= \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} && \text{Perm of indistinct objects}
\end{aligned}
$$

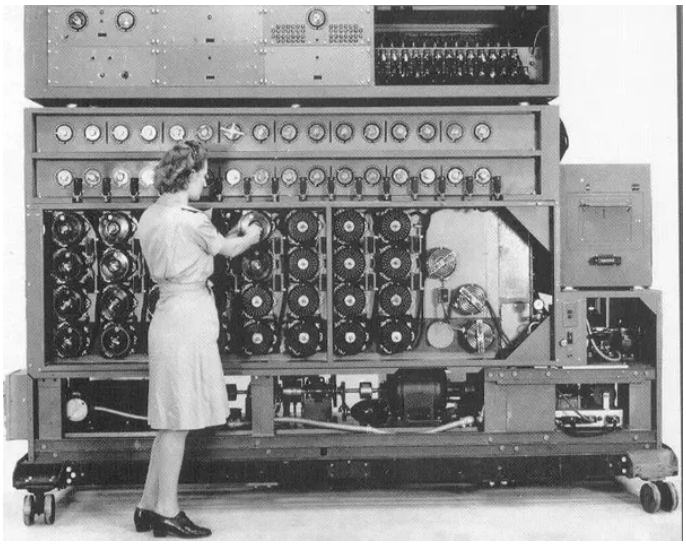
3. More than k heads

You can use the formula for exactly k heads to compute other probabilities. For example the probability of more than k heads is:

$$
\begin{aligned}
\mathrm{P}(\text{more than } k \text{ heads}) &= \sum_{i=k+1}^{n} \mathrm{P}(\text{exactly } i \text{ heads}) && \text{Mutual Exclusion} \\
&= \sum_{i=k+1}^{n} \binom{n}{i} \cdot p^i \cdot (1-p)^{n-i} && \text{Substitution}
\end{aligned}
$$

Enigma Machine

One of the very first computers was built to break the Nazi “enigma” codes in WW2. It was a hard problem because the “enigma” machine, used to make secret codes, had so many unique configurations. Every day the Nazi's would chose a new configuration and if they Allies could figure out the daily configuration, they could read all enemy messages. One solution was to try all configurations until one produced legible German. This begs the question: How many configurations are there?

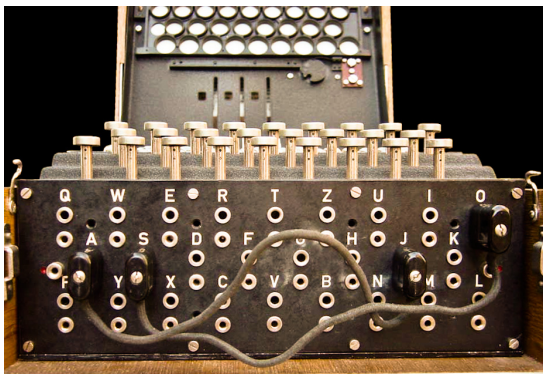


The WW2 machine built to search different enigma configurations.

The enigma machine has three rotors. Each rotor can be set to one of 26 different positions. How many unique configurations are there of the three rotors?

Using the steps rule of counting: $26 \cdot 26 \cdot 26 = 26^3 = 17,576$.

Whats more! The machine has a plug board which could swap the electrical signal for letters. On the plug board, wires can connect any pair of letters to produce a new configuration. A wire can’t connect a letter to itself. Wires are indistinct. A wire from ‘K’ to ’L’ is not considered distinct from a wire from ‘L’ to ’K’. We are going to work up to considering any number of wires.



The engima plugboard. For electical reasons each letter has two jacks and each plug has two prongs. Semantically this is equivalent to one plug location per letter.

One wire: How many ways are there to place exactly one wire that connects two letters?

Chosing 2 letters from 26 is a combination. Using the combination formula: $\binom{26}{2} = 325$.

Two wires: How many ways are there to place exactly two wires? Recall that wires are not considered distinct. Each letter can have at most one wire connected to it, thus you couldn’t have a wire connect ‘K’ to ‘L’ and another one connect ‘L’ to ‘X’

There are $\binom{26}{2}$ ways to place the first wire and $\binom{24}{2}$ ways to place the second wire. However, since the wires are indistinct, we have double counted every possibility. Because every possibility is counted twice we should divide by 2:

$$\text{Total} = \frac{\binom{26}{2} \cdot \binom{24}{2}}{2} = 44,850$$

Three wires: How many ways are there to place **exactly** three wires?

There are $\binom{26}{2}$ ways to place the first wire and $\binom{24}{2}$ ways to place the second wire. There are now $\binom{22}{2}$ ways to place the third. However, since the wires are indistinct, and our step counting implicitly treats them as distinct, we have overcounted each possibility. How many times is each pairing of three letters overcounted? Its the number of permutations of three distinct objects: 3!

$$\text{Total} = \frac{\binom{26}{2} \cdot \binom{24}{2} \cdot \binom{22}{2}}{3!} = 3,453,450$$

There is another way to arrive at the same answer. First we are going to chose the letters to be paired, then we are going to pair them off. There are $\binom{26}{6}$ ways to select the letters that are being wired up. We then need to pair off those letters. One way to think about pairing the letters off is to first permute them (6! ways) and then pair up the first two letters, the next two, the next two, and so on. For example, if our letters were {A,B,C,D,E,F} and our permutation was BADCEF, then this would correspond to wiring B to A and D to C and E to F. We are overcounting by a lot. First, we are overcounting by a factor of 3! since the ordering of the pairs doesn't matter. Second, we are overcounting by a factor of 2^3 since the ordering of the letters within each pair doesn't matter.

$$\text{Total} = \binom{26}{6} \frac{6!}{3! \cdot 2^3} = 3,453,450$$

Arbitrary wires: How many ways are there to place k wires, thus connecting $2 \cdot k$ letters? During WW2 the Germans always used a fixed number of wires. But one fear was that if they discovered the Enigma machine was cracked, they could simply use an arbitrary number of wires.

The set of ways to use exactly i wires is mutually exclusive from the set of ways to use exactly j wires if $i \neq j$ (since no way can use both exactly i and j wires). As such $\text{Total} = \sum_{k=0}^{13} \text{Total}_k$ Where Total_k is the number of ways to use exactly k wires. Continuing our logic for ways to used exact number of wires:

$$\text{Total}_k = \frac{\prod_{i=1}^{k} \binom{28-2i}{2}}{k!}$$

Bringing it all together:

$$\begin{aligned}
\text{Total} &= \sum_{k=0}^{13} \text{Total}_k \\
&= \sum_{k=0}^{13} \frac{\prod_{i=1}^{k} \binom{28-2i}{2}}{k!} \\
&= 532,985,208,200,576
\end{aligned}$$

The actual Enigma used in ww2 had exactly 10 wires connecting 20 letters allowing for 150,738,274,937,250 unique configuration. The enigma machine also chose the three rotors from a set of five adding another factor of $\binom{5}{3} = 60$.

When you combine the number of ways of setting the rotars, with the number of ways you could set the plug board you get the total number of configurations of an enigma machine. Thinking of this as two steps we can multiply the two numbers we earlier calculated: 17,576 · 150,738,274,937,250 · 60 $\approx 159 \cdot 10^{18}$ unique settings. So, Alan Turing and his team at Blechly Park to build a machine which could help test many configurations -- a predecesor to the first computers.

Serendipity

The word serendipity comes from the Persian fairy tale of the Three Princes of Serendip

Problem

What is the probability of a seredipitous encounter with a friend? Imagine you are live in an area with a large general population (eg Stanford with 17,000 students). A small subset of the population are friends. What are the chances that you run into at least one friend if you see a handful of people from the population? Assume that seeing each person from the population is equally likely.

Total Population

17000

Friends

150

People that you see

100

Calculate

Answer

The probability that you see at least one friend is:

Bacteria Evolution

A wonderful property of modern life is that we have anti-biotics to kill bacterial infections. However, we only have a fixed number of anti-biotic medicines, and bacteria are evolving to become resistent to our anti-biotics. In this example we are going to use probability to understand evolution of anti-biotic resistence in bacteria.

Imagine you have a population of 1 million infectious bacteria in your gut, 10% of which have a mutation that makes them slightly more resistant to anti-biotics. You take a course of anti-biotics. The probability that bacteria with the mutation survives is 20%. The probability that bacteria without the mutation survives is 1%.

What is the probability that a randomly chosen bacterium survives the anti-biotics?

Let E be the event that our bacterium survives. Let M be the event that a bacteria has the mutataion. By the By Law of Total Probability (LOTP):

$$\begin{aligned} \mathrm{P}(E) &= \mathrm{P}(E \text{ and } M) + \mathrm{P}(E \text{ and } M^{\mathrm{C}}) && \text{LOTP} \\ &= \mathrm{P}(E|M)\,\mathrm{P}(M) + \mathrm{P}(E|M^{\mathrm{C}})\,\mathrm{P}(M^{\mathrm{C}}) && \text{Chain Rule} \\ &= 0.20 \cdot 0.10 + 0.01 \cdot 0.90 && \text{Substituting} \\ &= 0.029 \end{aligned}$$

What is the probability that a surviving bacterium has the mutation?

Using the same events in the last section, this question is asking for $\mathrm{P}(M|E)$. We aren't givin the conditional probability in that direction, instead we know $P(E|M)$. Such situations call for Bayes' Theorem:

$$\begin{aligned} \mathrm{P}(M|E) &= \frac{\mathrm{P}(E|M)\,\mathrm{P}(M)}{\mathrm{P}(E)} && \text{Bayes} \\ &= \frac{0.20 \cdot 0.10}{\mathrm{P}(E)} && \text{Given} \\ &= \frac{0.20 \cdot 0.10}{0.029} && \text{Calculated} \\ &\approx 0.69 \end{aligned}$$

After the course of anti-biotics, 69% of bacteria have the mutation, up from 10% before. If this population is allowed to reproduce you will have a much more resistent set of bacteria!

Random Variables

A Random Variables (RV) is a variable that probabilistically takes on a value and they are one of the most important constructs in all of probability theory. You can think of an RV as being like a variable in a programming language, and in fact random variables are just as important to probability theory as variables are to programming. Random Variables take on values, have types and have domains over which they are applicable.

Random variables work with all of the foundational theory we have build up to this point. We can define events that occur if the random variable takes one values that satisfy a numerical test (eg does the variable equal 5, is the variable less than 8).

Lets look at a first example of a random variable. Say we flip three fair coins. We can define a random variable Y to be the total number of "heads" on the three coins. We can ask about the probability of Y taking on different values using the following notation:

Let Y be the number of heads on three coin flips

$\mathrm{P}(Y = 0) = 1/8$	(T, T, T)
$\mathrm{P}(Y = 1) = 3/8$	(H, T, T), (T, H, T), (T, T, H)
$\mathrm{P}(Y = 2) = 3/8$	(H, H, T), (H, T, H), (T, H, H)
$\mathrm{P}(Y = 3) = 1/8$	(H, H, H)
$\mathrm{P}(Y \geq 4) = 0$	

Even though we use the same notation for random variables and for events (both use capitol letters) they are distinct concepts. An event is a scenario, a random variable is an object. The scenario where a random variable takes on a particular value (or range of values) is an event. When possible, I will try and use letters E,F,G for events and X,Y,Z for random variables.

Using random variables is a convenient notation technique that assists in decomposing problems. There are many different types of random variables (indicator, binary, choice, Bernoulli, etc). The two main families of random variable types are discrete and continuous. Discrete random variables can only take on integer values. Continuous random variables can take on decimal values. We are going to develop our intuitions using discrete random variable and then introduce continuous.

1. Properties of random variables

There are many properties of a random variable of any random variable some of which we will dive into extensively. Here is a brief summary. Each random variable has:

Property	**Notation Example**	**Description**
Meaning		A semantic description of the random variable
Symbol	X	A letter used to denote the random variable
Support or Range	$\{0, 1, \ldots, 3\}$	the values the random variable can take on
Distribution Function (PMF or PDF)	$\mathrm{P}(X = x)$	A function which maps values the RV can take on to likelihood.
Expectation	$\mathrm{E}[X]$	A weighted average

Property	Notation Example	Description
Variance	$\text{Var}(X)$	A measure of spread
Standard Deviation	$\text{Std}(X)$	The square root of variance
Mode		The most likely value of the random variable

You should set a goal of deeply understanding what each of these properties mean. There are many more properties than the ones in the table above: properties like entropy, median, skew, kertosis.

2. Random variables vs Events

Random variables and events are two different concepts. An event is an outcome, or a set of outcomes, to an experiment. A random variable is a more like an experiment -- it will take on an outcome eventually. Probabilities are over events, so if you want to talk about probability in the context of a random variable, you must construct an event. You can make events by using any of the Relational Operators: $<$, $\leq$, $>$, $\geq$, $=$, or $\neq$ (not equal to). This is analogous to coding where you can use relational operators to create boolean expressions from numbers.

Lets continue our example of the random variable Y which represents the number of heads on three coin flips. Here are some events using the variable Y:

Event	Meaning	Probability Statement
$Y = 1$	Y takes on the value 1 (there was one heads)	$\text{P}(Y = 1)$
$Y < 2$	Y takes on 0 or 1 (note this Y can't be negative)	$\text{P}(Y < 2)$
$X > Y$	X takes on a value greater than the value Y takes on.	$\text{P}(X > Y)$
$Y = y$	Y takes on a value represented by non-random variable y	$\text{P}(Y = y)$

You will see many examples like this last one, $\text{P}(Y = y)$, in this text book as well as in scientific and math research papers. It allows us to talk about the likelihood of Y taking on a value, in general. For example, later in this book we will derive that for three coin flips where Y is the number of heads, the probability of getting exactly y heads is:

$$\text{P}(Y = y) = \frac{0.75}{y!(3-y)!} \qquad \text{If } 0 \leq y \leq 3$$

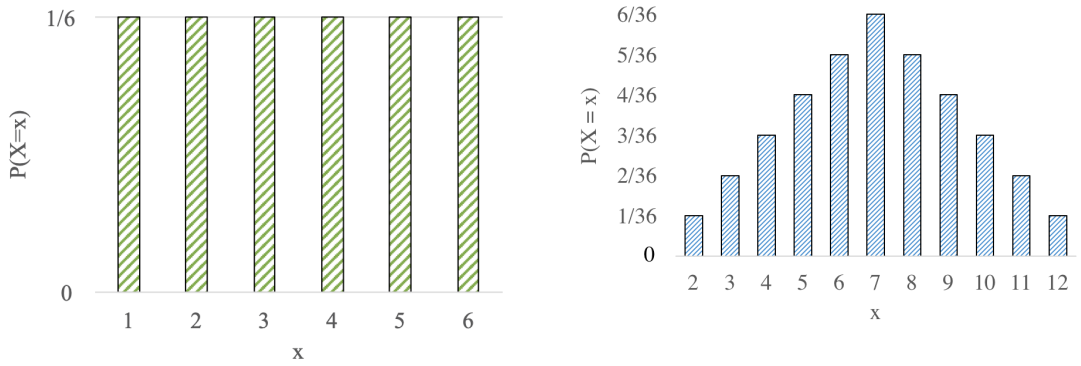
This statement above is a function which takes in a parameter y as input and returns the numeric probability $\text{P}(Y = y)$ as output. This particular expression allows us to talk about the probability that the number of heads is 0, 1, 2 or 3 all in one expression. You can plug in any one of those values for y to get the corresponding probability. It is customary to use lower-case symbols for non-random values. The use of an equals sign in the "event" can be confusing. For example what does this expression say $\text{P}(Y = 1) = 0.375$? It says that the probability that "Y takes on the value 4" is 0.375. For discrete random variables this function is called the "probability mass function" and it is the topic of our next chapter.

Probability Mass Functions

For a random variable, the most important thing to know is: how likely is each outcome? For a discrete random variable, this information is called the "**probability mass function**". The probability mass function (PMF) provides the "mass" (i.e. amount) of "probability" for each possible assignment of the random variable.

Formally, the probability mass function is a mapping between the values that the random variable could take on and the probability of the random variable taking on said value. In mathematics, we call these associations functions. There are many different ways of representing functions: you can write an equation, you can make a graph, you can even store many samples in a list. Let's start by looking at PMFs as graphs where the x-axis is the values that the random variable could take on and the y-axis is the probability of the random variable taking on said value.

In the following example, on the left we show a PMF, as a graph, for the random variable: X = the value of a six-sided die roll. On the right we show a contrasting example of a PMF for the random variable X = value of the sum of two dice rolls:



Left: the PMF of a single six-sided die roll. Right: the PMF of the sum of two dice rolls.

The sum of two dice example in the equally likely probability section. Again, the information that is provided in these graphs is the likelihood of a random variable taking on different values. In the graph on the right, the value "6" on the x-axis is associated with the probability $\frac{5}{36}$ on the y-axis. This x-axis refers to the event "the sum of two dice is 6" or $Y = 6$. The y-axis tells us that the probability of that event is $\frac{5}{36}$. In full: $\mathrm{P}(Y = 6) = \frac{5}{36}$. The value "2" is associated with "$\frac{1}{36}$" which tells us that, $\mathrm{P}(Y = 2) = \frac{1}{36}$, the probability that two dice sum to 2 is $\frac{1}{36}$. There is no value associated with "1" because the sum of two dice can not be 1. If you find this notation confusing, revisit the random variables section.

Here is the exact same information in equation form:

$$\mathrm{P}(X = x) = \frac{1}{6} \qquad \text{if } 1 \leq x \leq 6$$

$$\mathrm{P}(Y = y) = \begin{cases} \frac{(y-1)}{36} & \text{if } 1 \leq y \leq 7 \\ \frac{(13-y)}{36} & \text{if } 8 \leq y \leq 12 \end{cases}$$

As a final example, here is the PMF for Y, the sum of two dice, in Python code:

```
def pmf_sum_two_dice(y):
    # Returns the probability that the sum of two dice is y
    if y < 2 or y > 12:
        return 0
    if y <= 7:
        return (y-1) / 36
    else:
        return (13-y) / 36
```

1. Notation

You may feel that $\mathrm{P}(Y = y)$ is redundant notation. In probability research papers and higher-level work, mathematicians often use the shorthand $\mathrm{P}(y)$ to mean $\mathrm{P}(Y = y)$. This shorthand assumes that the lowercase value (e.g. y) has a capital letter counterpart (e.g. Y) that represents a random variable even though it's not written explicitly. In this book, we will often use the full form of the event $\mathrm{P}(Y = y)$, but we will occasionally use the shorthand $\mathrm{P}(y)$.

2. Probabilities Must Sum to 1

For a variable (call it X) to be a proper random variable it must be the case that if you summed up the values of $\mathrm{P}(X = k)$ for all possible values k that X can take on, the result must be 1:

$$\sum_k \mathrm{P}(X = k) = 1$$

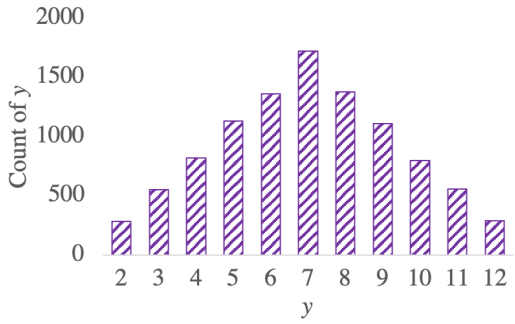
For further understanding, let's derive why this is the case. A random variable taking on a value is an event (for example $X = 2$). Each of those events is mutually exclusive because a random variable will take on exactly one value. Those mutually exclusive cases define an entire sample space. Why? Because X must take on *some* value.

3. Data to Histograms to Probability Mass Functions

One surprising way to store a likelihood function (recall that a PMF is the name of the likelihood function for *discrete* random variables) is simply a list of data. We simulated summing two die 10,000 times to make this example dataset:

```
[8, 4, 9, 7, 7, 7, 7, 5, 6, 8, 11, 5, 7, 7, 7, 6, 7, 8, 8, 9, 9, 4, 6, 7, 10,
12, 6, 7, 8, 9, 3, 7, 4, 9, 2, 8, 5, 8, 9, 6, 8, 7, 10, 7, 6, 7, 7, 5, 4, 6, 9,
5, 7, 4, 2, 11, 10, 11, 8, 4, 11, 9, 7, 10, 12, 4, 8, 5, 11, 5, 3, 9, 7, 5, 5,
5, 3, 8, 6, 11, 11, 2, 7, 7, 6, 5, 4, 6, 3, 8, 5, 8, 7, 6, 9, 4, 3, 7, 6, 6, 6,
5, 6, 10, 5, 9, 9, 8, 8, 7, 4, 8, 4, 9, 8, 5, 10, 10, 9, 7, 9, 7, 7, 10, 4, 7,
8, 4, 7, 8, 9, 11, 7, 9, 10, 10, 2, 7, 9, 4, 8, 8, 12, 9, 5, 11, 10, 7, 6, 4, 8,
9, 9, 6, 5, 6, 5, 6, 11, 7, 3, 10, 7, 3, 7, 7, 10, 3, 6, 8, 6, 8, 5, 10, 2, 7,
```

Note that this data, on its own, represents an approximation for the probability mass function. If you wanted to approximate $\mathrm{P}(Y = 5)$ you could simply count the number of times that "5" occurs in your data. This is an approximation based on the definition of probability. Here is the full histogram of the data, a count of times each value occurs:



A normalized histogram (where each value is divided by the length of your data list) is an approximation of the PMF. For a dataset of discrete numbers, a histogram shows the count of each value (in this case y). By the definition of probability, if you divide this count by the number of experiments run, you arrive at an approximation of the probability of the event $\mathrm{P}(Y = y)$. In our example, we have 10,000 elements in our dataset. The count of times that 3 occurs is 552. Note that:

$$\frac{\text{count}(Y = 5)}{n} = \frac{552}{10000} = 0.0552$$
$$\mathrm{P}(Y = 5) = \frac{4}{36} = 0.0555$$

In this case, since we ran 10,000 trials, the histogram is a very good approximation of the PMF. We use the sum of dice as an example because it is easy to understand. Datasets in the real world often represent more exciting events.

Expectation

A random variable is fully prepresented by its probability mass function (PMF), which represents each of the values the random variable can take on, and the corresponding probabilities. A PMF can be a lot of information. Sometimes it is useful to summarize the random variable! The most common, and arguably the most useful, summary of a random variable is its "**Expectation**".

Definition: Expectation

The expectation of a random variable X, writte $\mathrm{E}[X]$ is the average of all the values the random variable can take on, each weighted by the probability that the random variable will take on that value.

$$\mathrm{E}[X] = \sum_x x \cdot \mathrm{P}(X = x)$$

Expectation goes by many other names: Mean, Weighted Average, Center of Mass, 1st Moment. All of which are calculated using the same formula.

Recall that $\mathrm{P}(X = x)$, also written as $\mathrm{P}(x)$, is the probability mass function of the random variable X. Here is code that calculates the expectation of the sum of two dice, based off the probability mass function:

```
def expectation_sum_two_dice():
    exp_sum_two_dice = 0
    # sum of dice can take on the values 2 through 12
    for x in range(2, 13):
        pr_x = pmf_sum_two_dice(x) # pmf gives Pr(x)
        exp_sum_two_dice += x * pr_x
    return exp_sum_two_dice
```

If we worked it out manually we would get that if X is the sum of two dice, $\mathrm{E}[X] = 7$:

$$\mathrm{E}[X] = \sum_x x \cdot \mathrm{P}(X = x) = 2 \cdot \frac{1}{36} + 2 \cdot \frac{2}{36} + \cdots + 12\frac{1}{36} = 7$$

7 is the "average" number you expect to get if you took the sum of two dice near infinite times. In this case it also happens to be the same as the mode, the most likely value of the sum of two dice, but this is not always the case!

1. Properties of expectation

Property: Linearity of Expectation

$$E[aX + b] = a\,\mathrm{E}[X] + b$$

Property: Expectation of the Sum of Random Variables

$$E[X + Y] = E[X] + E[Y]$$

Property: Law of Unconcious Statistician

$$E[g(X)] = \sum_x g(x)\,\mathrm{P}(X = x)$$

One can also calculate the expected value of a function g(X) of a random variable X when one knows the probability distribution of X but one does not explicitly know the distribution of g(X). This theorem has the humorous name of "the Law of the Unconscious Statistician" (LOTUS), because it is so useful that

you should be able to employ it unconciously.

Property: Expectation of a Constant

$$E[a] = a$$

Sometimes in proofs, you will end up with the expectation of a constant (rather than a random variable). For example what does the $\mathrm{E}[5]$ mean? Since 5 is not a random variable, it does not change, and will always be 5, $\mathrm{E}[5] = 5$.

Variance

Definition: Variance of a Random Variable

The variance is a measure of the "spread" of a random variable around the mean. Variance for a random variable, X, with expected value $\mathrm{E}[X] = \mu$ is:

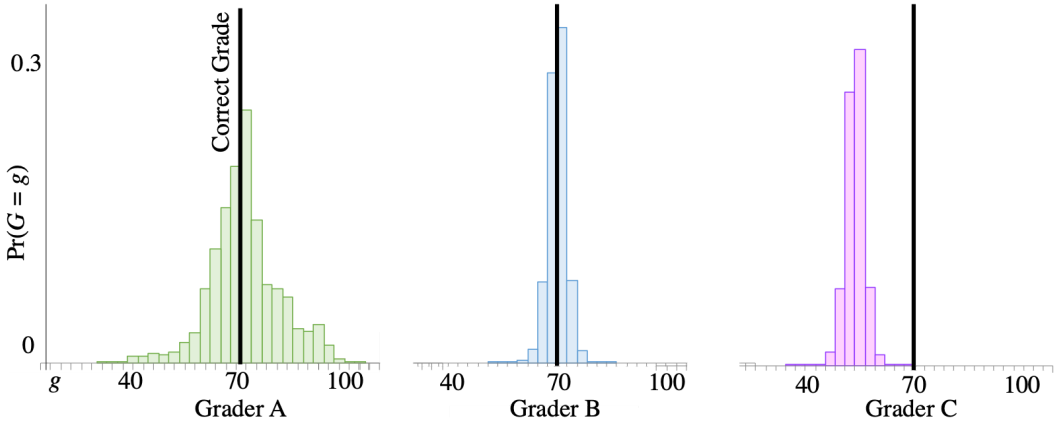
$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2]$$

Semantically, this is the average distance of a sample from the distribution to the mean. When computing the variance often we use a different (equivalent) form of the variance equation:

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

In the last section we showed that Expectation was a useful summary of a random variable (it calculates the "weighted average" of the random variable). One of the next most important properties of random variables to understand is variance: the measure of spread.

To start, lets consider probability mass functions for three sets of graders. When each of them grades an assigment, meant to receive a 70/100, they each have a probability distribution of grades that they could give.



Distributions of three types of peer graders. Data is from a massive online course.

The distribution for graders in group C have a different *expectation*. The average grade that they give when grading an assignment worth 70 is a 55/100. That is clearly not great! But what is the difference between graders A and B? Both of them have the same expected value (which is equal to the correct grade). The graders in group A have a higher "spread". When grading an assignment worth 70, they have a reasonable chance of giving it a 100, or of giving it a 40. Graders in group B have much less spread. Most of the probability mass is close to 70. You want graders like those in group B: in expectation they give the correct grade, and they have low spread. As an aside: scores in group B came from a probabilistic algorithm over peer grades.

Theorists wanted a number to describe spread. They invented variance to be the average of the distance between values that the random variable could take on and the mean of the random variable. There are many reasonable choices for the distance function, probability theorists chose squared deviation from the mean:

$$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2]$$

Proof: $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$

It is much easier to compute variance using $\mathrm{E}[X^2] - \mathrm{E}[X]^2$. You certainly don't need to know why its an equivalent expression, but in case you were wondering, here is the proof.

$$
\begin{aligned}
\operatorname{Var}(X) &= \mathrm{E}[(X-\mu)^2] && \text{Note: } \mu = \mathrm{E}[X] \\
&= \sum_x (x-\mu)^2 \, \mathrm{P}(x) && \text{Definition of Expectation} \\
&= \sum_x (x^2 - 2\mu x + \mu^2) \, \mathrm{P}(x) && \text{Expanding the square} \\
&= \sum_x x^2 \, \mathrm{P}(x) - 2\mu \sum_x x \, \mathrm{P}(x) + \mu^2 \sum_x \mathrm{P}(x) && \text{Propagating the sum} \\
&= \sum_x x^2 \, \mathrm{P}(x) - 2\mu \, \mathrm{E}[X] + \mu^2 \sum_x \mathrm{P}(x) && \text{Substitute def of expectation} \\
&= \mathrm{E}[X^2] - 2\mu \, \mathrm{E}[X] + \mu^2 \sum_x \mathrm{P}(x) && \text{LOTUS } g(x) = x^2 \\
&= \mathrm{E}[X^2] - 2\mu \, \mathrm{E}[X] + \mu^2 && \text{Since } \sum_x \mathrm{P}(x) = 1 \\
&= \mathrm{E}[X^2] - 2 \, \mathrm{E}[X]^2 + \mathrm{E}[X]^2 && \text{Since } \mu = \mathrm{E}[X] \\
&= \mathrm{E}[X^2] - \mathrm{E}[X]^2 && \text{Cancelation}
\end{aligned}
$$

1. Standard Deviation

Variance is especially useful for comparing the "spread" of two distributions and it has the useful property that it is easy to calculate. In general a larger variance means that there is more deviation around the mean — more spread. However, if you look at the leading example, the units of variance are the square of points. This makes it hard to interpret the numerical value. What does it mean that the spread is 52 points^2 ? A more interpretable measure of spread is the square root of Variance, which we call the Standard Deviation $\operatorname{Std}(X) = \sqrt{\operatorname{Var}(X)}$. The standard deviation of our grader is 7.2 points. In this example folks find it easier to think of spread in points rather than points ^{2}. As an aside, the standard deviation is the average distance of a sample (from the distribution) to the mean, using the euclidean distance function

Bernoulli Distribution

1. Parametric Random Variables

There are many classic and commonly-seen random variable abstractions that show up in the world of probability. At this point in the class, you will learn about several of the most significant parametric discrete distributions. When solving problems, if you can recognize that a random variable fits one of these formats, then you can use its pre-derived probability mass function (PMF), expectation, variance, and other properties. Random variables of this sort are called **parametric** random variables. If you can argue that a random variable falls under one of the studied parametric types, you simply need to provide parameters. A good analogy is a `class` in programming. Creating a parametric random variable is very similar to calling a constructor with input parameters.

2. Bernoulli Random Variables

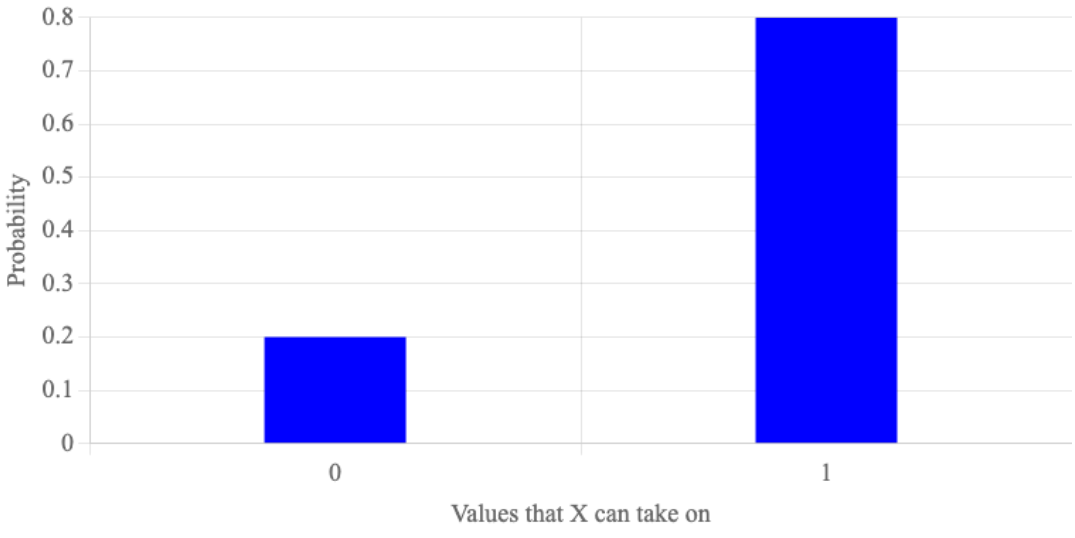
A Bernoulli random variable (also called a *boolean* or *indicator* random variable) is the simplest kind of parametric random variable. It can take on two values, 1 and 0. It takes on a 1 if an experiment with probability p resulted in success and a 0 otherwise. Some example uses include a coin flip, a random binary digit, whether a disk drive crashed, and whether someone likes a Netflix movie. Here p is the parameter, but different instances of Bernoulli random variables might have different values of p.

Here is a full description of the key properties of a Bernoulli random variable. If X is declared to be a Bernoulli random variable with parameter p, denoted $X \sim \text{Bern}(p)$:

Bernoulli Random Variable

Notation:	$X \sim \text{Bern}(p)$
Description:	A boolean variable that is 1 with probability p
Parameters:	p, the probability that $X = 1$.
Support:	x is either 0 or 1
PMF equation:	$\text{P}(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$
PMF (smooth):	$\text{P}(X = x) = p^x (1-p)^{1-x}$
Expectation:	$\text{E}[X] = p$
Variance:	$\text{Var}(X) = p(1-p)$
PMF graph:	

Parameter p: 0.80



Because Bernoulli distributed random variables are parametric, as soon as you declare a random variable to be of type Bernoulli you automatically can know all of these pre-derived properties! Some of these properties are straightforward to prove for a Bernoulli. For example, you could have solved for expectation:

Proof: Expectation of a Bernoulli. If X is a Bernoulli with parameter p, $X \sim \text{Bern}(p)$:

$$\begin{aligned} \text{E}[X] &= \sum_x x \cdot \text{P}(X = x) && \text{Definition of expectation} \\ &= 1 \cdot p + 0 \cdot (1 - p) && X \text{ can take on values 0 and 1} \\ &= p && \text{Remove the 0 term} \end{aligned}$$

Proof: Variance of a Bernoulli. If X is a Bernoulli with parameter p, $X \sim \text{Bern}(p)$:

To compute variance, first compute $E[X^2]$:

$$\begin{aligned} E[X^2] &= \sum_x x^2 \cdot \text{P}(X = x) && \text{LOTUS} \\ &= 0^2 \cdot p + 1^2 \cdot p \\ &= p \end{aligned}$$

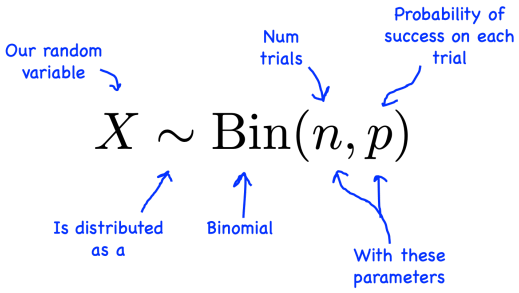
$$\begin{aligned} \text{Var}(X) &= E[X^2] - E[X]^2 && \text{Def of variance} \\ &= p - p^2 && \text{Substitute } E[X^2] = p, E[X] = p \\ &= p(1 - p) && \text{Factor out } p \end{aligned}$$

3. Indicator

Bernoulli random variables and indicator variables are two aspects of the same concept. A random variable I is an indicator variable for an event A if $I = 1$ when A occurs and $I = 0$ if A does not occur. $\text{P}(I = 1) = \text{P}(A)$ and $\text{E}[I] = \text{P}(A)$. Indicator random variables are Bernoulli random variables, with $p = \text{P}(A)$.

Binomial Distribution

In this section, we will discuss the binomial distribution. To start, imagine the following example. Consider n independent trials of an experiment where each trial is a "success" probability p. Let X be the number of successes in n trials. This situation is truly common in the natural world, and as such, there has been a lot of research into such phenomena. Random variables like X are called binomial random variables. If you can identify that a process fits this description, you can inherit many already proved properties such as the PMF formula, expectation, and variance!



Here are a few examples of binomial random variables:

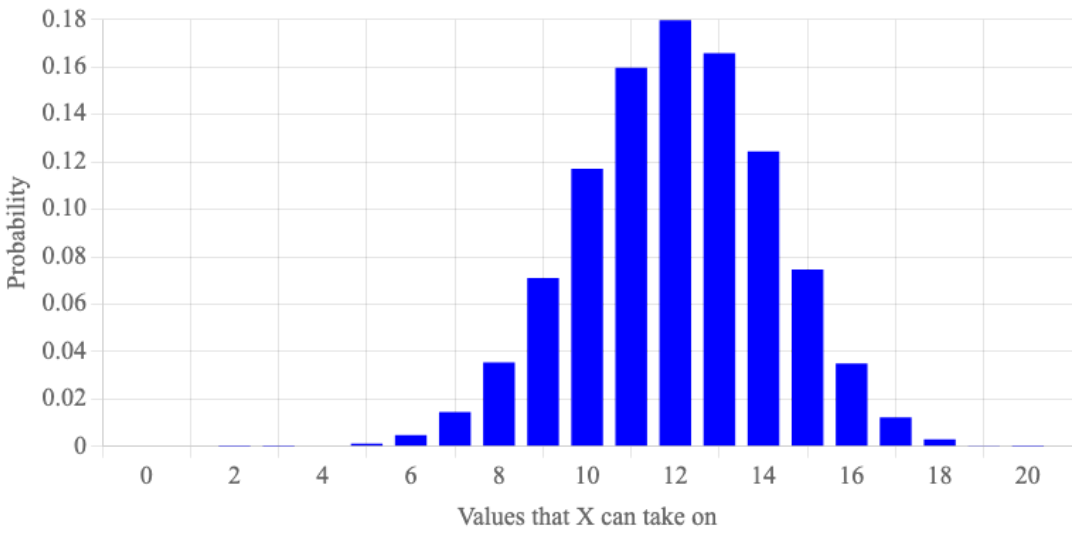
- # of heads in n coin flips
- # of 1's in randomly generated length n bit string
- # of disk drives crashed in 1000 computer cluster, assuming disks crash independently

Binomial Random Variable

Notation:	$X \sim \text{Bin}(n, p)$
Description:	Number of "successes" in n identical, independent experiments each with probability of success p.
Parameters:	$n \in \{0, 1, \ldots\}$, the number of experiments. $p \in [0, 1]$, the probability that a single experiment gives a "success".
Support:	$x \in \{0, 1, \ldots, n\}$
PMF equation:	$\text{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$
Expectation:	$\text{E}[X] = n \cdot p$
Variance:	$\text{Var}(X) = n \cdot p \cdot (1-p)$

PMF graph:

Parameter n: 20 Parameter p: 0.60



One way to think of the binomial is as the sum of n Bernoulli variables. Say that $Y_i \sim \text{Bern}(p)$ is an indicator Bernoulli random variable which is 1 if experiment i is a success. Then if X is the total number of successes in n experiments, $X \sim \text{Bin}(n, p)$:

$$X = \sum_{i=1}^{n} Y_i$$

Recall that the outcome of Y_i will be 1 or 0, so one way to think of X is as the sum of those 1s and 0s.

1. Binomial PMF

The most important property to know about a binomial is its PMF function:

Probability Mass Function for a Binomial

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

Probability that our variable takes on the value k

Recall, we derived this formula in Part 1. There is a complete example on the probability of k heads in n coin flips, where each flip is heads with probability 0.5: Many Coin Flips. To briefly review, if you think of each experiment as being distinct, then there are $\binom{n}{k}$ ways of permuting k successes from n experiments. For any of the mutually exclusive permutations, the probability of that permutation is $p^k \cdot (1-p)^{n-k}$.

The name binomial comes from the term $\binom{n}{k}$ which is formally called the binomial coefficient.

2. Expectation of Binomial

There is an easy way to calculate the expectation of a binomial and a hard way. The easy way is to leverage the fact that a binomial is the sum of Bernoulli random variables. $X = \sum_{i=1}^{n} Y_i$ where $Y_i \sim \text{Bern}(p)$. Since the expectation of the sum of random variables is the sum of expectations, we can add the expectation, $\text{E}[Y_i] = p$, of each of the Bernoulli's:

$$\begin{aligned}
\text{E}[X] &= \text{E}\left[\sum_{i=1}^{n} Y_i\right] && \text{Since } X = \sum_{i=1}^{n} Y_i \\
&= \sum_{i=1}^{n} \text{E}[Y_i] && \text{Expectation of sum} \\
&= \sum_{i=1}^{n} p && \text{Expectation of Bernoulli} \\
&= n \cdot p && \text{Sum } n \text{ times}
\end{aligned}$$

The hard way is to use the definition of expectation:

$$\begin{aligned}
\text{E}[X] &= \sum_{i=0}^{n} i \cdot \text{P}(X = i) && \text{Def of expectation} \\
&= \sum_{i=0}^{n} i \cdot \binom{n}{i} p^i (1-p)^{n-i} && \text{Sub in PMF} \\
&\cdots && \text{Many steps later} \\
&= n \cdot p
\end{aligned}$$

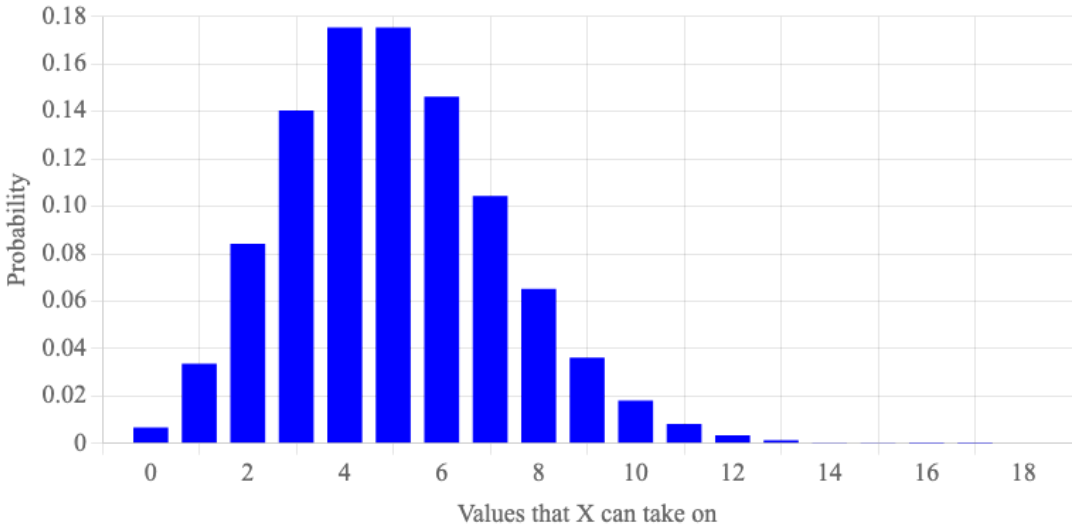
Poisson Distribution

A Poisson random variable gives the probability of a given number of events in a fixed interval of time (or space). It make the Poisson assumption that events occur with a known constant mean rate and independently of the time since the last event.

Poisson Random Variable

Notation:	$X \sim \text{Poi}(\lambda)$
Description:	Number of events in a fixed time frame if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.
Parameters:	$\lambda \in \{0, 1, \ldots\}$, the constant average rate.
Support:	$x \in \{0, 1, \ldots\}$
PMF equation:	$\text{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
Expectation:	$\text{E}[X] = \lambda$
Variance:	$\text{Var}(X) = \lambda$
PMF graph:	

Parameter λ: 5



1. Poisson Intuition

In this section we show the intuition behind the Poisson derivation. It is both a great way to deeply understand the Poisson, as well as good practice with Binomial distributions.

Let's work on the problem of predicting the chance of a given number of events occuring in a fixed time interval — the next minute. For example, imagine you are working on a ride sharing application and you care about the probability of how many requests you get from a particular area. From historical data, you know that the average requests per minute is $\lambda = 5$. What is the probability of getting 1, 2, 3, etc requests in a minute?

💡: We could approximate a solution to this problem by using a binomial distribution! Lets say we split our minute into 60 seconds, and make each second an indicator Bernoulli variable — you either get a request or you don't. If you get a request in a second, the indicator is 1. Otherwise it is 0. Here is a visualization of our 60 binary-indicators. In this example imagine we have requests at 2.75 and 7.12 seconds. the corresponding indicator variables are **blue** filled in boxes:

1 minute

									…	

The total number of requests received over the minute can be approximated as the sum of the sixty indicator variables, which conveniently matches the description of a binomial — a sum of Bernoullis. Specifically define X to be the number of requests in a minute. X is a binomial with $n = 60$ trials. What is the probability, p, of a success on a single trial? To make the expectation of X equal the observed historical average $\lambda = 5$ we should chose p so that $\lambda = \mathrm{E}[X]$.

$$\begin{aligned} \lambda &= \mathrm{E}[X] && \text{Expectation matches historical average} \\ \lambda &= n \cdot p && \text{Expectation of a Binomial is } n \cdot p \\ p &= \frac{\lambda}{n} && \text{Solving for } p \end{aligned}$$

In this case since $\lambda = 5$ and $n = 60$, we should chose $p = 5/60$ and state that $X \sim \mathrm{Bin}(n = 60, p = 5/60)$. Now that we have a form for X we can answer probability questions about the number of requests by using the Binomial PMF:

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

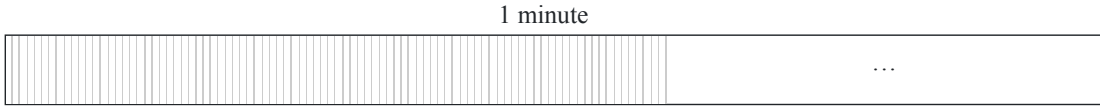
So for example:

$$\mathrm{P}(X = 1) = \binom{60}{1} (5/60)^1 (55/60)^{60-1} \approx 0.0295$$

$$\mathrm{P}(X = 2) = \binom{60}{2} (5/60)^2 (55/60)^{60-2} \approx 0.0790$$

$$\mathrm{P}(X = 3) = \binom{60}{3} (5/60)^3 (55/60)^{60-3} \approx 0.1389$$

Great! But don't forget that this was an approximation. We didn't account for the fact that there can be more than one event in a single second. One way to assuage this issue is to devide our minute into more fine-grained intervals (the choice to split it into 60 seconds was rather arbitrary). Instead lets divide our minute into 600 deciseconds, again with requests at 2.75 and 7.12 seconds:



Now $n = 600$, $p = 5/600$ and $X \sim \mathrm{Bin}(n = 600, p = 6/600)$. We can repeat our example calculations using this better approximation:

$$\mathrm{P}(X = 1) = \binom{600}{1} (5/600)^1 (595/60)^{600-1} \approx 0.0333$$

$$\mathrm{P}(X = 2) = \binom{600}{2} (5/600)^2 (595/600)^{600-2} \approx 0.0837$$

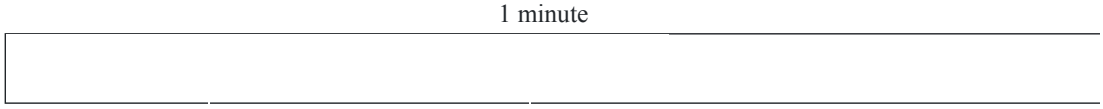
$$\mathrm{P}(X = 3) = \binom{600}{3} (5/600)^3 (595/600)^{600-3} \approx 0.1402$$

Chose any value of n, the number of buckets to divide our minute into: 0

The larger n is, the more accurate the approximation. So what happens when n is infinity? It becomes a Poisson!

2. Poisson, a Binomial in the limit

Or if we really cared about making sure that we don't get two events in the same bucket, we can divide our minute into infinitely small buckets:



Proof: Derivation of the Poisson

What does the PMF of X look like now that we have infinite divisions of our minute? We can write the equation and think about it as n goes to infinity. Recall that p still equals λ/n:

$$\mathrm{P}(X = x) = \lim_{n\to\infty} \binom{n}{x} (\lambda/n)^x (1-\lambda/n)^{n-x}$$

While it may look intimidating, this expression simplifies nicely. This proof uses a few special limit rules that we haven't introduced in this book:

$$\begin{aligned}
\mathrm{P}(X = x) &= \lim_{n\to\infty} \binom{n}{x} (\lambda/n)^x (1-\lambda/n)^{n-x} && \text{Start: binomial in the limit} \\
&= \lim_{n\to\infty} \binom{n}{x} \cdot \frac{\lambda^x}{n^x} \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^x} && \text{Expanding the power terms} \\
&= \lim_{n\to\infty} \frac{n!}{(n-x)!x!} \cdot \frac{\lambda^x}{n^x} \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^x} && \text{Expanding the binomial term} \\
&= \lim_{n\to\infty} \frac{n!}{(n-x)!x!} \cdot \frac{\lambda^x}{n^x} \cdot \frac{e^{-\lambda}}{(1-\lambda/n)^x} && \text{Rule } \lim_{n\to\infty}(1-\lambda/n)^n = e^{-\lambda} \\
&= \lim_{n\to\infty} \frac{n!}{(n-x)!x!} \cdot \frac{\lambda^x}{n^x} \cdot \frac{e^{-\lambda}}{1} && \text{Rule } \lim_{n\to\infty} \lambda/n = 0 \\
&= \lim_{n\to\infty} \frac{n!}{(n-x)!} \cdot \frac{1}{x!} \cdot \frac{\lambda^x}{n^x} \cdot \frac{e^{-\lambda}}{1} && \text{Splitting first term} \\
&= \lim_{n\to\infty} \frac{n^x}{1} \cdot \frac{1}{x!} \cdot \frac{\lambda^x}{n^x} \cdot \frac{e^{-\lambda}}{1} && \lim_{n\to\infty} \frac{n!}{(n-x)!} = n^x \\
&= \lim_{n\to\infty} \frac{\lambda^x}{x!} \cdot \frac{e^{-\lambda}}{1} && \text{Cancel } n^x \\
&= \frac{\lambda^x \cdot e^{-\lambda}}{x!} && \text{Simplify}
\end{aligned}$$

That is a beautiful expression! Now we can calculate the real probability of number of requests in a minute, if the historical average is $\lambda = 5$:

$$\mathrm{P}(X = 1) = \frac{5^1 \cdot e^{-5}}{1!} = 0.03369$$

$$\mathrm{P}(X = 2) = \frac{5^2 \cdot e^{-5}}{2!} = 0.08422$$

$$\mathrm{P}(X = 3) = \frac{5^3 \cdot e^{-5}}{3!} = 0.14037$$

This is both more accurate and much easier to compute!

3. Changing time frames

Say you are given a rate over one unit of time, but you want to know the rate in another unit of time. For example, you may be given the rate of hits to a website per minute, but you want to know the probability over a 20 minute period. You would just need to multiply this rate by 20 in order to go from the "per 1 minute of time" rate to obtain the "per 20 minutes of time" rate.

Continuous Distribution

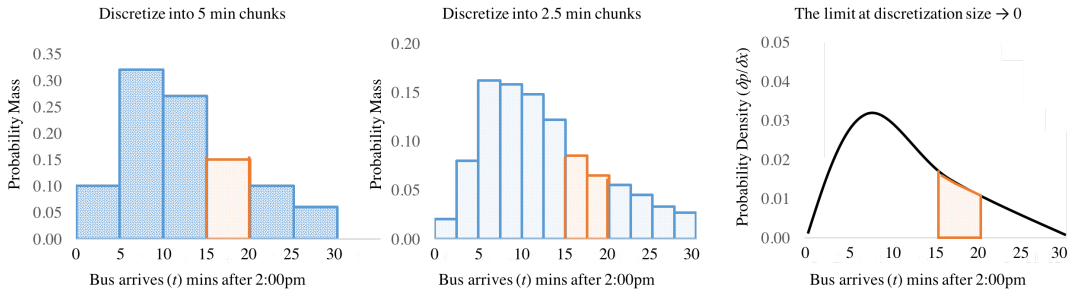
So far, all random variables we have seen have been *discrete*. In all the cases we have seen in CS109 this meant that our RVs could only take on integer values. Now it's time for *continuous* random variables which can take on values in the real number domain ($\mathbb{R}$). Continuous random variables can be used to represent measurements with arbitrary precision (eg height, weight, time).

1. From Discrete to Continuous

To make our transition from thinking about discrete random variable, to thinking about continuous random variables, lets start with a thought experiment: Imagine you are running to catch the bus. You know that you will arrive at 2:15pm but you don't know exactly when the bus will arrive, and want to think of the arrival time in minutes past 2pm as a random variable T so that you can calculate the probability that you will have to wait more than five minutes $P(15 < T < 20)$.

We immediately face a problem. For discrete distributions we would describe the probability that a random variable takes on exact values. This doesn't make sense for continuous values, like the time the bus arrives. As an example, what is the probability that the bus arrives at exactly 2:17pm and 12.12333911102389234 seconds? Similarly, if I were to ask you: what is the probability of a child being born with weight *exactly* equal to 3.523112342234 kilos, you might recognize that question as ridiculous. No child will have precisely that weight. Real values can have infinite precision and as such it is a bit mind boggling to think about the probability that a random variable takes on a specific value.

Instead, let's start by discretizing time, our continuous variable, by breakint it into 5 minute chunks. We can now think about something like, the probability that the bus arrives between 2:00p and 2:05 as an event with some probability (see figure below on the left). Five minute chunks seem a bit coarse. You could imagine that instead, we could have discretized time into 2.5minute chunks (figure in the center). In this case the probability that the bus shows up between 15 mins and 20 mins after 2pm is the sum of two chunks, shown in orange. Why stop there? In the limit we could keep breaking time down into smaller and smaller pieces. Eventually we will be left with a derivative of probability at each moment of time, where the probability that $P(15 < T < 20)$ is the integral of that derivative between 15 and 20 (figure on the right).



2. Probability Density Functions

In the world of discrete random variables, the most important property of a random variable was its probability mass function (PMF) that would tell you the probability of the random variable taking on any value. When we move to the world of continuous random variables, we are going to need to rethink this basic concept. In the continuous world, every random variable instead has a Probability *Density* Function (PDF) which defines the relative likelihood that a random variable takes on a particular value. We traditionally use the symbol f for the probability density function and write it in one of two ways:

$$f(X = x) \quad \text{or} \quad f(x)$$

Where the notation on the right hand side is shorthand where the lowercase x implies that we are talking about the relative likelihood of a continuous random variable which is the upper case X. Like in the bus example, the PDF is the derivative of probability at all points of the random variable. This means that the

PDF has the important property that you can integrate over it to find the probability that the random variable takes on values within a range (a, b).

Definition: Continuous Random Variable

X is a Continuous Random Variable if there is a Probability Density Function (PDF) $f(x)$ that takes in real valued numbers x such that:

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\,dx$$

The following properties must also hold. These preserve the axiom that $P(a \leq X \leq b)$ is a probability:

$$0 \leq \mathrm{P}(a \leq X \leq b) \leq 1$$
$$\mathrm{P}(-\infty < X < \infty) = 1$$

A common misconception is to think of $f(x)$ as a probability. It is instead what we call a probability density. It represents probability/unit of X. Generally this is only meaningful when we either take an integral over the PDF **or** we compare probability densities. As we mentioned when motivating probability densities, the probability that a continuous random variable takes on a specific value (to infinite precision) is 0.

$$\mathrm{P}(X = a) = \int_a^a f(x)\,dx = 0$$

That is pretty different than in the discrete world where we often talked about the probability of a random variable taking on a particular value.

3. Cumulative Distribution Function

Having a probability density is great, but it means we are going to have to solve an integral every single time we want to calculate a probability. To avoid this unfortunate fate, we are going to use a standard called a cumulative distribution function (CDF). The CDF is a function which takes in a number and returns the probability that a random variable takes on a value less than that number. It has the pleasant property that, if we have a CDF for a random variable, we don't need to integrate to answer probability questions!

For a continuous random variable X the Cumulative Distribution Function, written $F(x)$ is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy$$

Why is the CDF the probability that a random variable takes on a value \textbf{less than} the input value as opposed to greater than? It is a matter of convention. But it is a useful convention. Most probability questions can be solved simply by knowing the CDF (and taking advantage of the fact that the integral over the range $-\infty$ to ∞ is 1. Here are a few examples of how you can answer probability questions by just using a CDF:

Probability Query	Solution	Explanation
$\mathrm{P}(X < a)$	$F(a)$	That is the definition of the CDF
$\mathrm{P}(X \leq a)$	$F(a)$	Trick question. $\mathrm{P}(X = a) = 0$
$\mathrm{P}(X > a)$	$1 - F(a)$	$\mathrm{P}(X < a) + \mathrm{P}(X > a) = 1$
$\mathrm{P}(a < X < b)$	$F(b) - F(a)$	$F(a) + \mathrm{P}(a < X < b) = F(b)$

The continuous distribution also exists for discrete random variables, but there is less utility to a CDF in the discrete world as none of our discrete random variables had ``closed form" (eg without any summation) functions for the CDF:

$$F_X(a) = \sum_{i=1}^{a} P(X = i)$$

4. Solving for Constants

Let X be a continuous random variable with PDF:

$$f(x) = \begin{cases} C(4x - 2x^2) & \text{when } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

In this function, C is a constant. What value is C? Since we know that the PDF must sum to 1:

$$\begin{aligned} &\int_0^2 C(4x - 2x^2)\,dx = 1 \\ &C\left(2x^2 - \frac{2x^3}{3}\right)\Big|_0^2 = 1 \\ &C\left(\left(8 - \frac{16}{3}\right) - 0\right) = 1 \\ &C = 3/8 \end{aligned}$$

Now that we know C, what is $\mathrm{P}(X > 1)$?

$$\begin{aligned} \mathrm{P}(X > 1) &= \int_1^\infty f(x)\,dx \\ &= \int_1^2 \frac{3}{8}(4x - 2x^2)\,dx \\ &= \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)\Big|_1^2 \\ &= \frac{3}{8}\left[\left(8 - \frac{16}{3}\right) - \left(2 - \frac{2}{3}\right)\right] = \frac{1}{2} \end{aligned}$$

5. Expectation and Variance of Continuous Variables

For continuous RV X:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\ E[g(X)] &= \int_{-\infty}^{\infty} g(x) f(x) dx \\ E[X^n] &= \int_{-\infty}^{\infty} x^n f(x) dx \end{aligned}$$

For both continuous and discrete RVs:

$$\begin{aligned} &E[aX + b] = aE[X] + b \\ &\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2 \\ &\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X) \end{aligned}$$

Uniform Distribution

The most basic of all the continuous random variables is the uniform random variable, which is equally likely to take on any value in its range (α, β). X is a *uniform random variable* ($X \sim \text{Uni}(\alpha, \beta)$) if it has PDF:

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{when } \alpha \leq x \leq \beta \\ 0 & \text{otherwise} \end{cases}$$

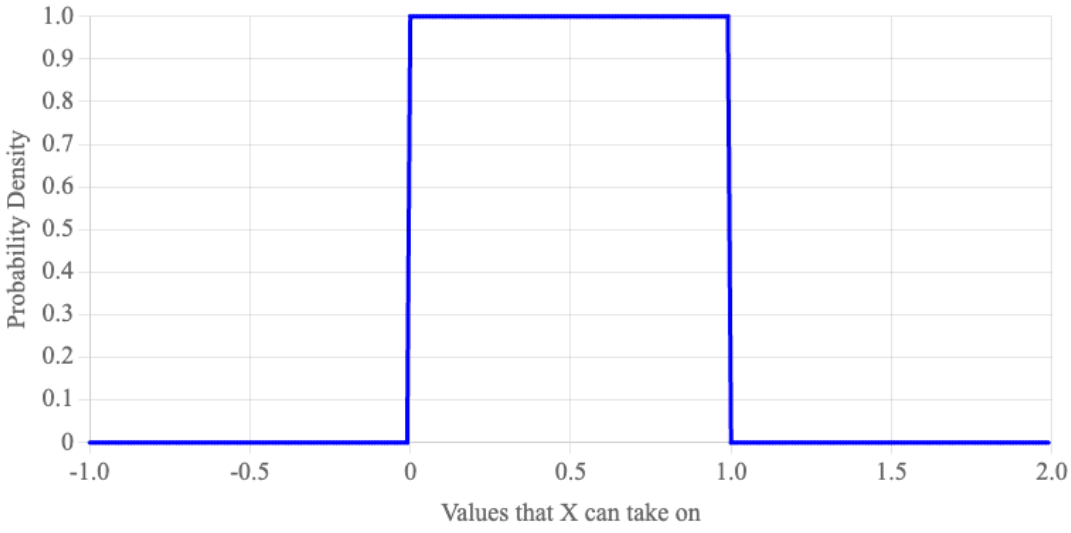
Notice how the density $1/(\beta - \alpha)$ is exactly the same regardless of the value for x. That makes the density uniform. So why is the PDF $1/(\beta - \alpha)$ and not 1? That is the constant that makes it such that the integral over all possible inputs evaluates to 1.

Uniform Random Variable

Notation:	$X \sim \text{Uni}(\alpha, \beta)$
Description:	A continuous random variable that takes on values, with equal likelihood, between α and β
Parameters:	$\alpha \in \mathbb{R}$, the minimum value of the variable. $\beta \in \mathbb{R}$, $\beta > \alpha$, the maximum value of the variable.
Support:	$x \in [\alpha, \beta]$
PDF equation:	$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{else} \end{cases}$
CDF equation:	$F(x) = \begin{cases} \frac{x - \alpha}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x < \alpha \\ 1 & \text{for } x > \beta \end{cases}$
Expectation:	$\text{E}[X] = \frac{1}{2}(\alpha + \beta)$
Variance:	$\text{Var}(X) = \frac{1}{12}(\beta - \alpha)^2$

PDF graph:

Parameter α: 0 Parameter β: 1



Example: You are running to the bus stop. You don't know exactly when the bus arrives. You believe all times between 2 and 2:30 are equally likely. You show up at 2:15pm. What is P(wait < 5 minutes)?

Let T be the time, in minutes after 2p that the bus arrives. Because we think that all times are equally likely in this range, $T \sim \text{Uni}(\alpha = 0, \beta = 30)$. The probability that you wait 5 minutes is equal to the probability that the bus shows up between 2:15 and 2:20. In other words $\text{P}(15 < T < 20)$:

$$
\begin{aligned}
\mathrm{P}(\text{Wait under 5 mins}) &= \mathrm{P}(15 < T < 20) \\
&= \int_{15}^{20} f_T(x) \partial x \\
&= \int_{15}^{20} \frac{1}{\beta - \alpha} \partial x \\
&= \frac{1}{30} \partial x \\
&= \frac{x}{30} \Big|_{15}^{20} \\
&= \frac{20}{30} - \frac{15}{30} = \frac{5}{30}
\end{aligned}
$$

We can come up with a closed form for the probability that a uniform random variable X is in the range a to b, assuming that $\alpha \leq a \leq b \leq \beta$:

$$
\begin{aligned}
\mathrm{P}(a \leq X \leq b) &= \int_a^b f(x)\, dx \\
&= \int_a^b \frac{1}{\beta - \alpha}\, dx \\
&= \frac{b - a}{\beta - \alpha}
\end{aligned}
$$

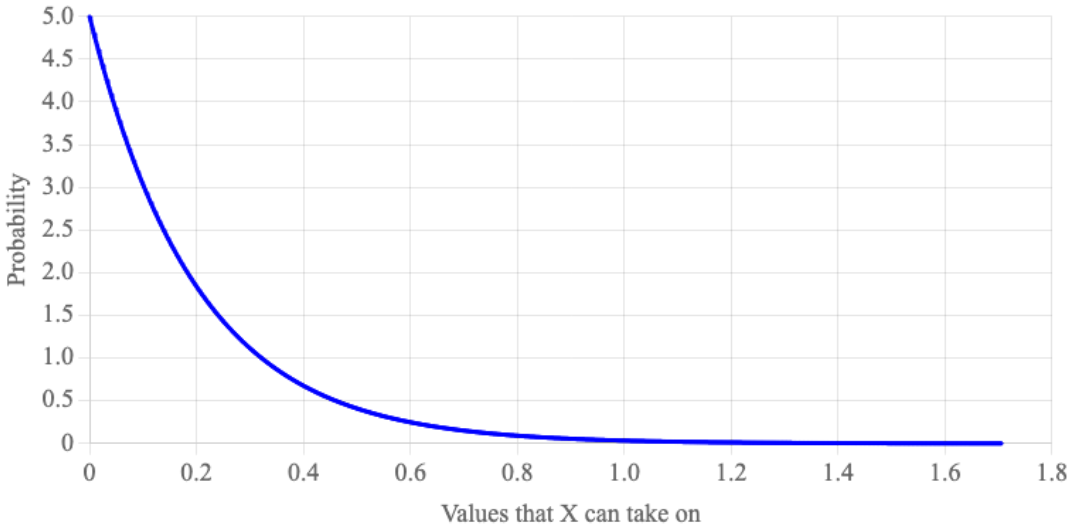
Exponential Distribution

An exponential distribution measures the **amount** of time until a next event occurs. It assumes that the events occur via a poisson process. Note that this is different from the Poisson Random Variable which measures number of events in a fixed amount of time.

Exponential Random Variable

Notation:	$X \sim \text{Exp}(\lambda)$
Description:	Time until next events if (a) the events occur with a constant mean rate and (b) they occur independently of time since last event.
Parameters:	$\lambda \in \{0, 1, \ldots\}$, the constant average rate.
Support:	$x \in \mathbb{R}^+$
PDF equation:	$f(x) = \lambda e^{-\lambda x}$
CDF equation:	$F(x) = 1 - e^{-\lambda x}$
Expectation:	$\text{E}[X] = 1/\lambda$
Variance:	$\text{Var}(X) = 1/\lambda^2$
PDF graph:	

Parameter λ: 5



An exponential distribution is a great example of a continuous distribution where the cumulative distribution funciton (CDF) is much easier to work with as it allows you to answer probability questions without using integrals.

Example: Based on historical data from the USGS, earthquakes of magnitude 8.0+ happen in a certain location at a rate of 0.002 per year. Earthquakes are known to occur via a poisson process. What is the probability of a major earthquake in the next 4 years?

Let Y be the years until the next major earthquake. Because Y measures time *until* the next event it fits the description of an exponential random variable: $Y \sim \text{Exp}(\lambda = 0.002)$. The question is asking, what is $\text{P}(Y < 4)$?

$$\begin{aligned} \text{P}(Y < 4) &= F_Y(4) && \text{The CDF measures } \text{P}(Y < y) \\ &= 1 - e^{-\lambda \cdot y} && \text{The CDF of an Exp} \\ &= 1 - e^{-0.002 \cdot 4} && \text{The CDF of an Exp} \\ &\approx 0.008 \end{aligned}$$

Note that it is possible to answer this question using the PDF, but it will require solving an integral.

1. Exponential is Memoryless

One way to gain intuition for what is meant by the "poisson process" is through the proof that the exponential distribution is "memoryless". That means that the occurence (or lack of occurence) of events in the past does not change our belief as to how long until the next occurence. This can be stated formally. If $X \sim \text{Exp}(\lambda)$ then for an interval of time until the start s, and a proceeding, query, interval of time t:

$$\text{P}(X > s + t | X > s) = \text{P}(X > t)$$

Which is something we can prove:

$$\begin{aligned}
\text{P}(X > s + t | X > s) &= \frac{\text{P}(X > s + t \text{ and } X > s)}{\text{P}(X > s)} && \text{Def of conditional prob.} \\
&= \frac{\text{P}(X > s + t)}{\text{P}(X > s)} && \text{Because } X > s + t \text{ implies } X > s \\
&= \frac{1 - F_X(s + t)}{1 - F_X(s)} && \text{Def of CDF} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} && \text{By CDF of Exp} \\
&= e^{-\lambda t} && \text{Simplify} \\
&= 1 - F_X(t) && \text{By CDF of Exp} \\
&= \text{P}(X > t) && \text{Def of CDF}
\end{aligned}$$

Normal Distribution

The single most important random variable type is the Normal (aka Gaussian) random variable, parametrized by a mean (μ) and variance (σ^2), or sometimes equivalently written as mean and variance (σ^2). If X is a normal variable we write $X \sim N(\mu, \sigma^2)$. The normal is important for many reasons: it is generated from the summation of independent random variables and as a result it occurs often in nature. Many things in the world are not distributed normally but data scientists and computer scientists model them as Normal distributions anyways. Why? Because it is the most entropic (conservative) modelling decision that we can make for a random variable while still matching a particular expectation (average value) and variance (spread).

The Probability Density Function (PDF) for a Normal $X \sim N(\mu, \sigma^2)$ is:

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

Notice the x in the exponent of the PDF function. When x is equal to the mean (μ) then e is raised to the power of 0 and the PDF is maximized.

By definition a Normal has $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

There is no closed form for the integral of the Normal PDF, and as such there is no closed form CDF. However we can use a transformation of any normal to a normal with a precomputed CDF. The result of this mathematical gymnastics is that the CDF for a Normal $X \sim N(\mu, \sigma^2)$ is:

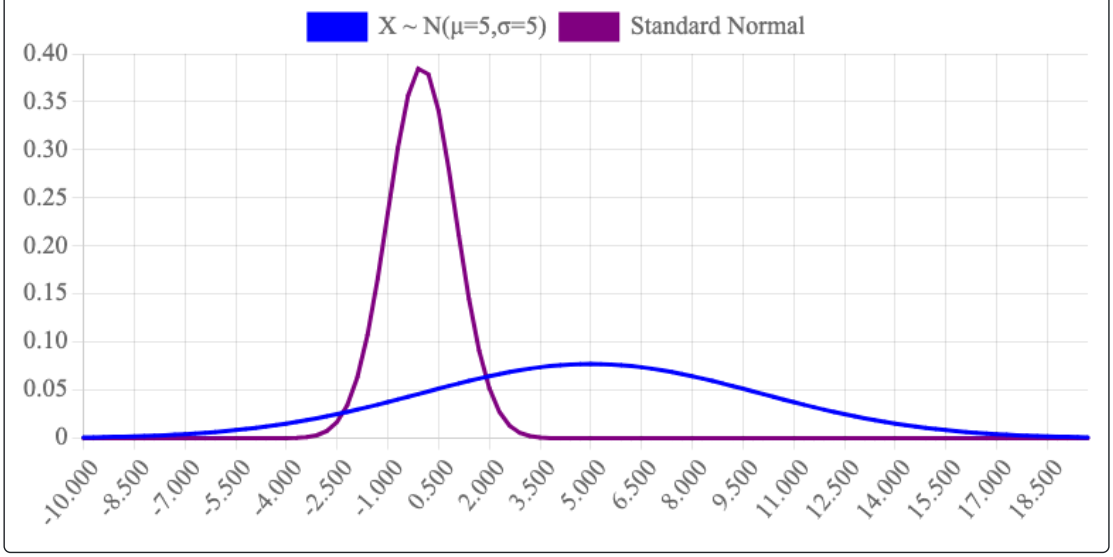
$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Where Φ is a precomputed function that represents that CDF of the Standard Normal.

Normal (aka Gaussian) Random Variable

Notation:	$X \sim \mathrm{N}(\mu, \sigma^2)$
Description:	A common, naturally occuring distribution.
Parameters:	$\mu \in \mathbb{R}$, the mean. $\sigma^2 \in \mathbb{R}$, the variance.
Support:	$x \in \mathbb{R}$
PDF equation:	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF equation:	$F(x) = \phi(\frac{x-\mu}{\sigma})$ Where ϕ is the CDF of the standard normal
Expectation:	$\mathrm{E}[X] = \mu$
Variance:	$\mathrm{Var}(X) = \sigma^2$
PDF graph:	

Parameter μ: 5 Parameter σ: 5



1. Linear Transform

If X is a Normal such that $X \sim N(\mu, \sigma^2)$ and Y is a linear transform of X such that $Y = aX + b$ then Y is also a Normal where:

$$Y \sim N(a\mu + b, a^2\sigma^2)$$

2. Projection to Standard Normal

For any Normal X we can find a linear transform from X to the standard normal $Z \sim N(0, 1)$. Note that Z is the typical notation choice for the standard normal. For any normal, if you subtract the mean (μ) of the normal and divide by the standard deviation (σ) the result is always the standard normal. We can prove this mathematically. Let $W = \frac{X-\mu}{\sigma}$:

$$\begin{aligned}
W &= \frac{X-\mu}{\sigma} && \text{Transform X: Subtract by } \mu \text{ and diving by } \sigma \\
&= \frac{1}{\sigma}X - \frac{\mu}{\sigma} && \text{Use algebra to rewrite the equation} \\
&= aX + b && \text{Linear transform where } a = \frac{1}{\mu}, b = -\frac{\mu}{\sigma} \\
&\sim N(a\mu + b, a^2\sigma^2) && \text{The linear transform of a Normal is another Normal} \\
&\sim N(\frac{\mu}{\sigma} - \frac{\mu}{\sigma}, \frac{\sigma^2}{\sigma^2}) && \text{Substituting values in for } a \text{ and } b \\
&\sim N(0, 1) && \text{The standard normal}
\end{aligned}$$

Using this transform we can express $F_X(x)$, the CDF of X, in terms of the known CDF of Z, $F_Z(x)$. Since the CDF of Z is so common it gets its own Greek symbol: $\Phi(x)$

$$\begin{aligned}
F_X(x) &= P(X \leq x) \\
&= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\
&= P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right)
\end{aligned}$$

The values of $\Phi(x)$ can be looked up in a table. Every modern programming language also has the ability to calculate the CDF of a normal random variable!

Example: Let $X \sim \mathcal{N}(3, 16)$, what is $P(X > 0)$?

$$P(X > 0) = P\left(\frac{X-3}{4} > \frac{0-3}{4}\right) = P\left(Z > -\frac{3}{4}\right) = 1 - P\left(Z \leq -\frac{3}{4}\right)$$
$$= 1 - \Phi(-\frac{3}{4}) = 1 - (1 - \Phi(\frac{3}{4})) = \Phi(\frac{3}{4}) = 0.7734$$

What is $P(2 < X < 5)$?

$$P(2 < X < 5) = P\left(\frac{2-3}{4} < \frac{X-3}{4} < \frac{5-3}{4}\right) = P\left(-\frac{1}{4} < Z < \frac{2}{4}\right)$$
$$= \Phi(\frac{2}{4}) - \Phi(-\frac{1}{4}) = \Phi(\frac{1}{2}) - (1 - \Phi(\frac{1}{4})) = 0.2902$$

Example: You send voltage of 2 or -2 on a wire to denote 1 or 0. Let X = voltage sent and let R = voltage received. $R = X + Y$, where $Y \sim \mathcal{N}(0, 1)$ is noise. When decoding, if $R \geq 0.5$ we interpret the voltage as 1, else 0. What is $P(\text{error after decoding} | \text{original bit} = 1)$?

$$\begin{aligned} P(X + Y < 0.5) &= P(2 + Y < 0.5) \\ &= P(Y < -1.5) \\ &= \Phi(-1.5) \\ &\approx 0.0668 \end{aligned}$$

Example: The 67% rule of a normal within one standard deviation. What is the probability that a normal variable $X \sim \mathrm{N}(\mu, \sigma)$ has a value within one standard deviation of its mean?

$$\begin{aligned} \mathrm{P}(\text{Within one } \sigma \text{ of } \mu) &= \mathrm{P}(\mu - \sigma < X < \mu + \sigma) \\ &= \mathrm{P}(X < \mu + \sigma) - \mathrm{P}(X < \mu - \sigma) && \text{Prob of a range} \\ &= \Phi\Big(\frac{(\mu + \sigma) - \mu}{\sigma}\Big) - \Phi\Big(\frac{(\mu - \sigma) - \mu}{\sigma}\Big) && \text{CDF of Normal} \\ &= \Phi\Big(\frac{\sigma}{\sigma}\Big) - \Phi\Big(\frac{-\sigma}{\sigma}\Big) && \text{Cancel } \mu\text{s} \\ &= \Phi(1) - \Phi(-1) && \text{Cancel } \sigma\text{s} \\ &\approx 0.8413 - 0.1587 \approx 0.683 && \text{Plug into } \Phi \end{aligned}$$

We made no assumption about the value of μ or the value of σ so this will apply to every single normal random variable. Since it uses the Normal CDF this doesn't apply to other types of random variables.

Binomial Approximation

There are times when it is exceptionally hard to numerically calculate probabilities for a binomial distribution, especially when n is large. For example, say $X \sim \text{Bin}(n = 10000, p = 0.5)$ and you want to calculate $\text{P}(X > 5500)$. The correct formula is:

$$\begin{aligned}\text{P}(X > 55) &= \sum_{i=5500}^{10000} \text{P}(X = x) \\ &= \sum_{i=5500}^{10000} \binom{10000}{i} p^i (1-p)^{10000-i}\end{aligned}$$

That is a difficult value to calculate. Luckily there is an easier way. For deep reasons which we will cover in our section on "uncertainty theory" it turns out that a binomial distribution can be very well approximated by both Normal distributions and Poisson distributions if n is large enough.

Use the Poisson approximation when n is large (>20) and p is small (<0.05). A slight dependence between results of each experiment is ok

Use the Normal approximation when n is large (>20), p mid-ranged. Specifically it considered an accurate approximation when the variance is greater then 10, in other words: $np(1-p) > 10$. There are situations where either a Poisson or a Normal can be used to approximate a Binomial. In that situation go with the Normal!

1. Poisson Approximation

When defining the Poisson we proved that a Binomial in the limit as $n \to \infty$ and $p = \lambda/n$ is a Poisson. That same logic can be used to show that a Poisson is a great approximation for a Binomial when the Binomial has extreme values of n and p. A Poisson random variable approximates Binomial where n is large, p is small, and $\lambda = np$ is “moderate”. Interestingly, to calculate the things we care about (PMF, expectation, variance) we no longer need to know n and p. We only need to provide λ which we call the rate. When approximating a Poisson with a Binomial always chose $\lambda = n \cdot p$.

There are different interpretations of "moderate". The accepted ranges are $n > 20$ and $p < 0.05$ or $n > 100$ and $p < 0.1$.

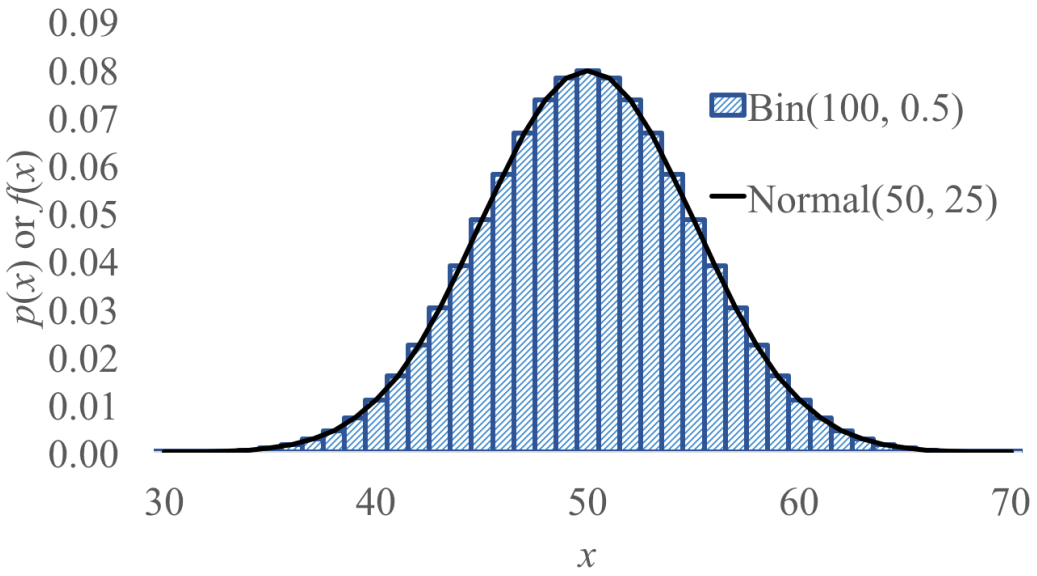
Let's say you want to send a bit string of length $n = 10^4$ where each bit is independently corrupted with $p = 10^{-6}$. What is the probability that the message will arrive uncorrupted? You can solve this using a Poisson with $\lambda = np = 10^4 10^{-6} = 0.01$. Let $X \sim Poi(0.01)$ be the number of corrupted bits. Using the PMF for Poisson:

$$\begin{aligned}P(X = 0) &= \frac{\lambda^i}{i!} e^{-\lambda} \\ &= \frac{0.01^0}{0!} e^{-0.01} \\ &\sim 0.9900498\end{aligned}$$

We could have also modelled X as a binomial such that $X \sim Bin(10^4, 10^{-6})$. That would have been impossible to calculate on a computer but would have resulted in the same number (up to the millionth decimal).

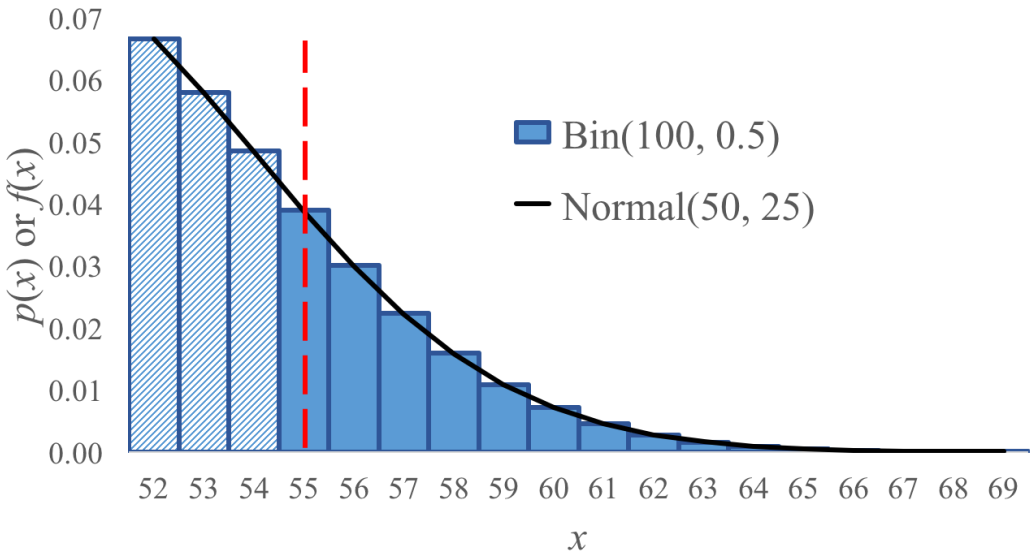
2. Normal Approximation

For a Binomial where n is large and p is mid-ranged, a Normal can be used to approximate the Binomial. Let's take a side by side view of a normal and a binomial:



Lets say our binomial is a random variable $X \sim \text{Bin}(100, 0.5)$ and we want to calculate $P(X \geq 55)$. We could cheat by using the closest fit normal (in this case $Y \sim N(50, 25)$). How did we chose that particular Normal? Simply select one with a mean and variance that matches the Binomial expectation and variance. The binomial expectation is $np = 100 \cdot 0.5 = 50$. The Binomial variance is $np(1-p) = 100 \cdot 0.5 \cdot 0.5 = 25$.

You can use a Normal distribution to approximate a Binomial $X \sim \text{Bin}(n, p)$. To do so define a normal $Y \sim (E[X], Var(X))$. Using the Binomial formulas for expectation and variance, $Y \sim (np, np(1-p))$. This approximation holds for large n and moderate p. That gets you very close. However since a Normal is continuous and Binomial is discrete we have to use a continuity correction to discretize the Normal.



$$P(X = k) \sim P\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right) = \Phi\left(\frac{k - np + 0.5}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - np - 0.5}{\sqrt{np(1-p)}}\right)$$

You should get comfortable deciding what continuity correction to use. Here are a few examples of discrete probability questions and the continuity correction:

Discrete (Binomial) probability question	Equivalent continuous probability question
$P(X = 6)$	$P(5.5 < X < 6.5)$
$P(X \geq 6)$	$P(X > 5.5)$
$P(X > 6)$	$P(X > 6.5)$
$P(X < 6)$	$P(X < 5.5)$
$P(X \leq 6)$	$P(X < 6.5)$

Example: 100 visitors to your website are given a new design. Let X = \# of people who were given the new design and spend more time on your website. Your CEO will endorse the new design if $X \geq 65$. What is $P(\text{CEO endorses change}|\text{it has no effect})$?

$E[X] = np = 50$. $\text{Var}(X) = np(1-p) = 25$. $\sigma = \sqrt{\text{Var}(X)} = 5$. We can thus use a Normal approximation: $Y \sim \mathcal{N}(\mu = 50, \sigma^2 = 25)$.

$$P(X \geq 65) \approx P(Y > 64.5) = P\left(\frac{Y-50}{5} > \frac{64.5-50}{5}\right) = 1 - \Phi(2.9) = 0.0019$$

Example: Stanford accepts 2480 students and each student has a 68\% chance of attending. Let X = \# students who will attend. $X \sim \text{Bin}(2480, 0.68)$. What is $P(X > 1745)$?

$E[X] = np = 1686.4$. $\text{Var}(X) = np(1-p) = 539.7$. $\sigma = \sqrt{\text{Var}(X)} = 23.23$. We can thus use a Normal approximation: $Y \sim \mathcal{N}(\mu = 1686.4, \sigma^2 = 539.7)$.

$$\begin{aligned} P(X > 1745) &\approx P(Y > 1745.5) \\ &\approx P\left(\frac{Y-1686.4}{23.23} > \frac{1745.5-1686.4}{23.23}\right) \\ &\approx 1 - \Phi(2.54) = 0.0055 \end{aligned}$$

100 Binomial Problems

Just for fun (and to give you a lot of practice) I wrote a generative probabilistic program which could sample binomial distribution problems. Here are 100 binomial questions:

Questions

Question 1: Laura is running a server cluster with 50 computers. The probability of a crash on a given server is 0.5. What is the standard deviation of crashes?

Answer 1:
Let X be the number of crashes. $X \sim \text{Bin}(n = 50, p = 0.5)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{50 \cdot 0.5 \cdot (1-0.5)} \\ &= 3.54\end{aligned}$$

Question 2: You are showing an online-ad to 30 people. The probability of an ad ignore on each ad shown is 2/3. What is the expected number of ad clicks?

Answer 2:
Let X be the number of ad clicks. $X \sim \text{Bin}(n = 30, p = 1/3)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 30 \cdot 1/3 \\ &= 10\end{aligned}$$

Question 3: A machine learning algorithm makes binary predictions. The machine learning algorithm makes 50 guesses where the probability of a incorrect prediction on a given guess is 19/25. What is the probability that the number of correct predictions is greater than 0?

Answer 3:
Let X be the number of correct predictions. $X \sim \text{Bin}(n = 50, p = 6/25)$

$$\begin{aligned}\text{P}(X > 0) &= 1 - \text{P}(0 <= X <= 0) \\ &= 1 - \binom{n}{0} p^0 (1-p)^{n-0}\end{aligned}$$

Question 4: Wind blows independently across 50 locations. The probability of no wind at a given location is 0.5. What is the expected number of locations that have wind?

Answer 4:
Let X be the number of locations that have wind. $X \sim \text{Bin}(n = 50, p = 0.5)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 50 \cdot 0.5 \\ &= 25.0\end{aligned}$$

Question 5: Wind blows independently across 30 locations. What is the standard deviation of locations that have wind? the probability of wind at each location is 0.6.

Answer 5:
Let X be the number of locations that have wind. $X \sim \text{Bin}(n = 30, p = 0.6)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{30 \cdot 0.6 \cdot (1-0.6)} \\ &= 2.68\end{aligned}$$

Question 6: You are trying to mine bitcoins. There are 50 independent attempts where the probability of a mining a bitcoin on a given attempt is 0.6. What is the expectation of bitcoins mined?

Answer 6:
Let X be the number of bitcoins mined. $X \sim \text{Bin}(n = 50, p = 0.6)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 50 \cdot 0.6 \\ &= 30.0\end{aligned}$$

Question 7: You are testing a new medicine on 40 patients. What is P(X is exactly 38)? The number of cured patients can be represented by a random variable X. X ~ Bin(40, 3/10).

Answer 7:
Let X be the number of cured patients. $X \sim \text{Bin}(n = 40, p = 3/10)$

$$\begin{aligned}\text{P}(X = 38) &= \binom{n}{38} p^{38}(1-p)^{n-38} \\ &= \binom{40}{38} 3/10^{38}(1-3/10)^{40-38} \\ &< 0.00001\end{aligned}$$

Question 8: You are manufacturing chips and are testing for defects. There are 50 independent tests and 0.5 is the probability of a defect on each test. What is the standard deviation of defects?

Answer 8:
Let X be the number of defects. $X \sim \text{Bin}(n = 50, p = 0.5)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{50 \cdot 0.5 \cdot (1-0.5)} \\ &= 3.54\end{aligned}$$

Question 9: Laura is flipping a coin 12 times. The probability of a tail on a given coin-flip is 5/12. What is the probability that the number of tails is greater than or equal to 2?

Answer 9:
Let X be the number of tails. $X \sim \text{Bin}(n = 12, p = 5/12)$

$$\begin{aligned}\text{P}(X >= 2) &= 1 - \text{P}(0 <= X <= 1) \\ &= 1 - \sum_{i=0}^{1} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 10: You are asking a survey question where responses are "like" or "dislike". There are 30 responses. You can assume each response is independent where the probability of a dislike on a given response is 1/6. What is the probability that the number of likes is greater than 28?

Answer 10:
Let X be the number of likes. $X \sim \text{Bin}(n = 30, p = 5/6)$

$$\begin{aligned}\text{P}(X > 28) &= P(29 <= X <= 30) \\ &= \sum_{i=29}^{30} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 11: A ball hits a series of 50 pins where it can bounce either right or left. The probability of a left on a given pin hit is 0.4. What is the standard deviation of rights?

Answer 11:
Let X be the number of rights. $X \sim \text{Bin}(n = 50, p = 3/5)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{50 \cdot 3/5 \cdot (1 - 3/5)} \\ &= 3.46\end{aligned}$$

Question 12: You are sending a stream of 30 bits to space. The probability of a no corruption on a given bit is 1/3. What is the probability that the number of corruptions is 10?

Answer 12:
Let X be the number of corruptions. $X \sim \text{Bin}(n = 30, p = 2/3)$

$$\begin{aligned}\text{P}(X = 10) &= \binom{n}{10} p^{10}(1-p)^{n-10} \\ &= \binom{30}{10} 2/3^{10}(1 - 2/3)^{30-10} \\ &= 0.00015\end{aligned}$$

Question 13: Wind blows independently across locations. The probability of wind at a given location is 0.9. The number of independent locations is 20. What is the probability that the number of locations that have wind is not less than 19?

Answer 13:
Let X be the number of locations that have wind. $X \sim \text{Bin}(n = 20, p = 0.9)$

$$\begin{aligned}\text{P}(X >= 19) &= P(19 <= X <= 20) \\ &= \sum_{i=19}^{20} \binom{n}{i} p^{i}(1-p)^{n-i}\end{aligned}$$

Question 14: You are sending a stream of bits to space. There are 30 independent bits where 5/6 is the probability of a no corruption on each bit. What is the probability that the number of corruptions is 21?

Answer 14:
Let X be the number of corruptions. $X \sim \text{Bin}(n = 30, p = 1/6)$

$$\begin{aligned}\text{P}(X = 21) &= \binom{n}{21} p^{21}(1-p)^{n-21} \\ &= \binom{30}{21} 1/6^{21}(1 - 1/6)^{30-21} \\ &< 0.00001\end{aligned}$$

Question 15: Cody generates random bit strings. There are 20 independent bits. Each bit has a 1/4 probability of resulting in a 1. What is the probability that the number of 1s is 11?

Answer 15:
Let X be the number of 1s. $X \sim \text{Bin}(n = 20, p = 1/4)$

$$\begin{aligned}\text{P}(X = 11) &= \binom{n}{11} p^{11}(1-p)^{n-11} \\ &= \binom{20}{11} 1/4^{11}(1 - 1/4)^{20-11} \\ &= 0.00301\end{aligned}$$

Question 16: In a restaurant some customers ask for a water with their meal. A random sample of 40 customers is selected where the probability of a water requested by a given customer is 9/20. What is the probability that the number of waters requested is 16?

Answer 16:
Let X be the number of waters requested. $X \sim \text{Bin}(n = 40, p = 9/20)$

$$\begin{aligned}\text{P}(X = 16) &= \binom{n}{16} p^{16}(1-p)^{n-16} \\ &= \binom{40}{16} 9/20^{16}(1-9/20)^{40-16} \\ &= 0.10433\end{aligned}$$

Question 17: A student is guessing randomly on an exam with 12 questions. What is the expected number of correct answers? the probability of a correct answer on a given question is 5/12.

Answer 17:
Let X be the number of correct answers. $X \sim \text{Bin}(n = 12, p = 5/12)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 12 \cdot 5/12 \\ &= 5\end{aligned}$$

Question 18: Laura is trying to mine bitcoins. The number of bitcoins mined can be represented by a random variable X. X ~ Bin(n = 100, p = 1/2). What is P(X is equal to 53)?

Answer 18:
Let X be the number of bitcoins mined. $X \sim \text{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\text{P}(X = 53) &= \binom{n}{53} p^{53}(1-p)^{n-53} \\ &= \binom{100}{53} 1/2^{53}(1-1/2)^{100-53} \\ &= 0.06659\end{aligned}$$

Question 19: You are showing an online-ad to customers. The add is shown to 100 people. The probability of an ad ignore on a given ad shown is 1/2. What is the standard deviation of ad clicks?

Answer 19:
Let X be the number of ad clicks. $X \sim \text{Bin}(n = 100, p = 0.5)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 0.5 \cdot (1-0.5)} \\ &= 5.00\end{aligned}$$

Question 20: You are running a server cluster with 40 computers. 5/8 is the probability of a computer continuing to work on each server. What is the expected number of crashes?

Answer 20:
Let X be the number of crashes. $X \sim \text{Bin}(n = 40, p = 3/8)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 40 \cdot 3/8 \\ &= 15\end{aligned}$$

Question 21: You are hashing 100 strings into a hashtable. The probability of a hash to the first bucket on a given string hash is 3/20. What is the probability that the number of hashes to the first bucket is greater than or equal to 97?

Answer 21:
Let X be the number of hashes to the first bucket. $X \sim \text{Bin}(n = 100, p = 3/20)$

$$\begin{aligned} \text{P}(X >= 97) &= P(97 <= X <= 100) \\ &= \sum_{i=97}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 22: You are running in an election with 50 voters. 6/25 is the probability of a vote for you on each vote. What is the probability that the number of votes for you is less than 2?

Answer 22:
Let X be the number of votes for you. $X \sim \text{Bin}(n = 50, p = 6/25)$

$$\begin{aligned} \text{P}(X < 2) &= P(0 <= X <= 1) \\ &= \sum_{i=0}^{1} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 23: Irina is sending a stream of 40 bits to space. The probability of a corruption on each bit is 3/4. What is the probability that the number of corruptions is 22?

Answer 23:
Let X be the number of corruptions. $X \sim \text{Bin}(n = 40, p = 3/4)$

$$\begin{aligned} \text{P}(X = 22) &= \binom{n}{22} p^{22} (1-p)^{n-22} \\ &= \binom{40}{22} 3/4^{22} (1-3/4)^{40-22} \\ &= 0.00294 \end{aligned}$$

Question 24: You are hashing 100 strings into a hashtable. The probability of a hash to the first bucket on a given string hash is 9/50. What is the probability that the number of hashes to the first bucket is greater than 97?

Answer 24:
Let X be the number of hashes to the first bucket. $X \sim \text{Bin}(n = 100, p = 9/50)$

$$\begin{aligned} \text{P}(X > 97) &= P(98 <= X <= 100) \\ &= \sum_{i=98}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 25: You generate random bit strings. There are 100 independent bits. The probability of a 1 at a given bit is 3/25. What is the probability that the number of 1s is less than 97?

Answer 25:
Let X be the number of 1s. $X \sim \text{Bin}(n = 100, p = 3/25)$

$$\begin{aligned} \text{P}(X < 97) &= 1 - \text{P}(97 <= X <= 100) \\ &= 1 - \sum_{i=97}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 26: You are manufacturing toys and are testing for defects. What is the probability that the number of defects is greater than 1? the probability of a non-defect on a given test is 16/25 and you test 50 objects.

Answer 26:
Let X be the number of defects. $X \sim \text{Bin}(n = 50, p = 9/25)$

$$\begin{aligned}\mathrm{P}(X > 1) &= 1 - \mathrm{P}(0 <= X <= 1) \\ &= 1 - \sum_{i=0}^{1} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 27: Laura is sending a stream of 40 bits to space. The number of corruptions can be represented by a random variable X. X is a Binomial with n = 40 and p = 3/4. What is P(X = 25)?

Answer 27:
Let X be the number of corruptions. $X \sim \mathrm{Bin}(n = 40, p = 3/4)$

$$\begin{aligned}\mathrm{P}(X = 25) &= \binom{n}{25} p^{25} (1-p)^{n-25} \\ &= \binom{40}{25} 3/4^{25} (1 - 3/4)^{40-25} \\ &= 0.02819\end{aligned}$$

Question 28: 100 trials are run. What is the probability that the number of successes is 78? 1/2 is the probability of a success on each trial.

Answer 28:
Let X be the number of successes. $X \sim \mathrm{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\mathrm{P}(X = 78) &= \binom{n}{78} p^{78} (1-p)^{n-78} \\ &= \binom{100}{78} 1/2^{78} (1 - 1/2)^{100-78} \\ &< 0.00001\end{aligned}$$

Question 29: You are flipping a coin. You flip the coin 20 times. The probability of a tail on a given coin-flip is 1/10. What is the standard deviation of heads?

Answer 29:
Let X be the number of heads. $X \sim \mathrm{Bin}(n = 20, p = 0.9)$

$$\begin{aligned}\mathrm{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{20 \cdot 0.9 \cdot (1 - 0.9)} \\ &= 1.34\end{aligned}$$

Question 30: Irina is showing an online-ad to 12 people. 5/12 is the probability of an ad click on each ad shown. What is the probability that the number of ad clicks is less than or equal to 11?

Answer 30:
Let X be the number of ad clicks. $X \sim \mathrm{Bin}(n = 12, p = 5/12)$

$$\begin{aligned}\mathrm{P}(X <= 11) &= 1 - \mathrm{P}(12 <= X <= 12) \\ &= 1 - \binom{n}{12} p^1 2 (1-p)^{n-12}\end{aligned}$$

Question 31: You are flipping a coin 50 times. 19/25 is the probability of a head on each coin-flip. What is the standard deviation of tails?

Answer 31:
Let X be the number of tails. $X \sim \mathrm{Bin}(n = 50, p = 6/25)$

$$\begin{aligned}\mathrm{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{50 \cdot 6/25 \cdot (1 - 6/25)} \\ &= 3.02\end{aligned}$$

Question 32: You are running in an election with 100 voters. The probability of a vote for you on each vote is 1/4. What is the probability that the number of votes for you is less than or equal to 97?

Answer 32:
Let X be the number of votes for you. $X \sim \text{Bin}(n = 100, p = 1/4)$

$$\begin{aligned} \text{P}(X <= 97) &= 1 - \text{P}(98 <= X <= 100) \\ &= 1 - \sum_{i=98}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 33: You are running a server cluster with 40 computers. What is the probability that the number of crashes is less than or equal to 39? 3/4 is the probability of a computer continuing to work on each server.

Answer 33:
Let X be the number of crashes. $X \sim \text{Bin}(n = 40, p = 1/4)$

$$\begin{aligned} \text{P}(X <= 39) &= 1 - \text{P}(40 <= X <= 40) \\ &= 1 - \binom{n}{40} p^4 0 (1-p)^{n-40} \end{aligned}$$

Question 34: Waddie is sending a stream of bits to space. Waddie sends 100 bits. The probability of a corruption on each bit is 1/2. What is the standard deviation of corruptions?

Answer 34:
Let X be the number of corruptions. $X \sim \text{Bin}(n = 100, p = 1/2)$

$$\begin{aligned} \text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 1/2 \cdot (1 - 1/2)} \\ &= 5.00 \end{aligned}$$

Question 35: A student is guessing randomly on an exam with 100 questions. Each question has a 0.5 probability of resulting in a incorrect answer. What is the probability that the number of correct answers is greater than 97?

Answer 35:
Let X be the number of correct answers. $X \sim \text{Bin}(n = 100, p = 1/2)$

$$\begin{aligned} \text{P}(X > 97) &= P(98 <= X <= 100) \\ &= \sum_{i=98}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 36: You are testing a new medicine on patients. 0.5 is the probability of a cured patient on each trial. There are 10 independent trials. What is the expected number of cured patients?

Answer 36:
Let X be the number of cured patients. $X \sim \text{Bin}(n = 10, p = 0.5)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 10 \cdot 0.5 \\ &= 5.0 \end{aligned}$$

Question 37: A ball hits a series of pins where it can either go right or left. The number of independent pin hits is 100. The probability of a right on each pin hit is 0.5. What is the standard deviation of rights?

Answer 37:
Let X be the number of rights. $X \sim \text{Bin}(n = 100, p = 0.5)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 0.5 \cdot (1-0.5)} \\ &= 5.00\end{aligned}$$

Question 38: You are flipping a coin 40 times. The probability of a head on a given coin-flip is 1/2. What is the probability that the number of heads is 38?

Answer 38:
Let X be the number of heads. $X \sim \text{Bin}(n = 40, p = 1/2)$

$$\begin{aligned}\text{P}(X = 38) &= \binom{n}{38} p^{38} (1-p)^{n-38} \\ &= \binom{40}{38} 1/2^{38} (1-1/2)^{40-38} \\ &< 0.00001\end{aligned}$$

Question 39: 100 trials are run and the probability of a success on a given trial is 1/2. What is the standard deviation of successes?

Answer 39:
Let X be the number of successes. $X \sim \text{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 1/2 \cdot (1-1/2)} \\ &= 5.00\end{aligned}$$

Question 40: You are trying to mine bitcoins. There are 40 independent attempts. The probability of a mining a bitcoin on each attempt is 3/10. What is the probability that the number of bitcoins mined is 19?

Answer 40:
Let X be the number of bitcoins mined. $X \sim \text{Bin}(n = 40, p = 3/10)$

$$\begin{aligned}\text{P}(X = 19) &= \binom{n}{19} p^{19} (1-p)^{n-19} \\ &= \binom{40}{19} 3/10^{19} (1-3/10)^{40-19} \\ &= 0.00852\end{aligned}$$

Question 41: 20 trials are run. 0.5 is the probability of a failure on each trial. What is the probability that the number of successes is 6?

Answer 41:
Let X be the number of successes. $X \sim \text{Bin}(n = 20, p = 0.5)$

$$\begin{aligned}\text{P}(X = 6) &= \binom{n}{6} p^{6} (1-p)^{n-6} \\ &= \binom{20}{6} 0.5^{6} (1-0.5)^{20-6} \\ &= 0.03696\end{aligned}$$

Question 42: You are flipping a coin. What is the probability that the number of tails is 0? there are 30 independent coin-flips where the probability of a head on a given coin-flip is 5/6.

Answer 42:
Let X be the number of tails. $X \sim \text{Bin}(n = 30, p = 1/6)$

$$\begin{aligned}\mathrm{P}(X=0) &= \binom{n}{0} p^0 (1-p)^{n-0} \\ &= \binom{30}{0} 1/6^0 (1-1/6)^{30-0} \\ &= 0.00421\end{aligned}$$

Question 43: In a restaurant some customers ask for a water with their meal. A random sample of 20 customers is selected and each customer has a 1/4 probability of resulting in a water not requested. What is the probability that the number of waters requested is 14?

Answer 43:
Let X be the number of waters requested. $X \sim \mathrm{Bin}(n=20, p=3/4)$

$$\begin{aligned}\mathrm{P}(X=14) &= \binom{n}{14} p^{14} (1-p)^{n-14} \\ &= \binom{20}{14} 3/4^{14} (1-3/4)^{20-14} \\ &= 0.16861\end{aligned}$$

Question 44: A student is guessing randomly on an exam. 3/8 is the probability of a incorrect answer on each question. The number of independent questions is 40. What is the probability that the number of correct answers is less than or equal to 37?

Answer 44:
Let X be the number of correct answers. $X \sim \mathrm{Bin}(n=40, p=5/8)$

$$\begin{aligned}\mathrm{P}(X <= 37) &= 1 - \mathrm{P}(38 <= X <= 40) \\ &= 1 - \sum_{i=38}^{40} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 45: You are running in an election with 30 voters. 3/5 is the probability of a vote for you on each vote. What is the standard deviation of votes for you?

Answer 45:
Let X be the number of votes for you. $X \sim \mathrm{Bin}(n=30, p=3/5)$

$$\begin{aligned}\mathrm{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{30 \cdot 3/5 \cdot (1-3/5)} \\ &= 2.68\end{aligned}$$

Question 46: Charlotte is flipping a coin 100 times. The probability of a tail on each coin-flip is 0.5. What is the probability that the number of tails is greater than 2?

Answer 46:
Let X be the number of tails. $X \sim \mathrm{Bin}(n=100, p=0.5)$

$$\begin{aligned}\mathrm{P}(X > 2) &= 1 - \mathrm{P}(0 <= X <= 2) \\ &= 1 - \sum_{i=0}^{2} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 47: You are trying to mine bitcoins. You try 50 times. 3/5 is the probability of a not mining a bitcoin on each attempt. What is the probability that the number of bitcoins mined is 14?

Answer 47:
Let X be the number of bitcoins mined. $X \sim \mathrm{Bin}(n=50, p=2/5)$

$$\begin{aligned}\mathrm{P}(X=14) &= \binom{n}{14} p^{14}(1-p)^{n-14} \\ &= \binom{50}{14} 2/5^{14}(1-2/5)^{50-14} \\ &= 0.02597\end{aligned}$$

Question 48: You are testing a new medicine on 100 patients. The probability of a cured patient on a given trial is 3/25. What is the probability that the number of cured patients is not less than 97?

Answer 48:
Let X be the number of cured patients. $X \sim \mathrm{Bin}(n=100, p=3/25)$

$$\begin{aligned}\mathrm{P}(X >= 97) &= P(97 <= X <= 100) \\ &= \sum_{i=97}^{100} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 49: Wind blows independently across 40 locations. What is the probability that the number of locations that have wind is 40? 11/20 is the probability of no wind at each location.

Answer 49:
Let X be the number of locations that have wind. $X \sim \mathrm{Bin}(n=40, p=9/20)$

$$\begin{aligned}\mathrm{P}(X=40) &= \binom{n}{40} p^{40}(1-p)^{n-40} \\ &= \binom{40}{40} 9/20^{40}(1-9/20)^{40-40} \\ &< 0.00001\end{aligned}$$

Question 50: You are showing an online-ad to 30 people. 1/6 is the probability of an ad click on each ad shown. What is the probability that the number of ad clicks is less than or equal to 28?

Answer 50:
Let X be the number of ad clicks. $X \sim \mathrm{Bin}(n=30, p=1/6)$

$$\begin{aligned}\mathrm{P}(X <= 28) &= 1 - \mathrm{P}(29 <= X <= 30) \\ &= 1 - \sum_{i=29}^{30} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 51: You are flipping a coin. You flip the coin 40 times and 7/8 is the probability of a head on each coin-flip. What is the standard deviation of tails?

Answer 51:
Let X be the number of tails. $X \sim \mathrm{Bin}(n=40, p=1/8)$

$$\begin{aligned}\mathrm{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{40 \cdot 1/8 \cdot (1-1/8)} \\ &= 2.09\end{aligned}$$

Question 52: Cody is sending a stream of bits to space. 2/5 is the probability of a no corruption on each bit and there are 20 independent bits. What is the expectation of corruptions?

Answer 52:
Let X be the number of corruptions. $X \sim \mathrm{Bin}(n=20, p=3/5)$

$$\begin{aligned}\mathrm{E}[X] &= np \\ &= 20 \cdot 3/5 \\ &= 12\end{aligned}$$

Question 53: You are running in an election. There are 12 independent votes and 5/6 is the probability of a vote for you on each vote. What is the probability that the number of votes for you is greater than or equal to 9?

Answer 53:
Let X be the number of votes for you. $X \sim \text{Bin}(n = 12, p = 5/6)$

$$\begin{aligned} \text{P}(X >= 9) &= P(9 <= X <= 12) \\ &= \sum_{i=9}^{12} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 54: You are flipping a coin. The number of tails can be represented by a random variable X. X is a Bin(n = 30, p = 5/6). What is the probability that X = 1?

Answer 54:
Let X be the number of tails. $X \sim \text{Bin}(n = 30, p = 5/6)$

$$\begin{aligned} \text{P}(X = 1) &= \binom{n}{1} p^1 (1-p)^{n-1} \\ &= \binom{30}{1} 5/6^1 (1-5/6)^{30-1} \\ &< 0.00001 \end{aligned}$$

Question 55: In a restaurant some customers ask for a water with their meal. A random sample of 100 customers is selected where 0.3 is the probability of a water requested by each customer. What is the expected number of waters requested?

Answer 55:
Let X be the number of waters requested. $X \sim \text{Bin}(n = 100, p = 0.3)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 100 \cdot 0.3 \\ &= 30.0 \end{aligned}$$

Question 56: You are hashing strings into a hashtable. 30 strings are hashed. The probability of a hash to the first bucket on each string hash is 1/6. What is the expected number of hashes to the first bucket?

Answer 56:
Let X be the number of hashes to the first bucket. $X \sim \text{Bin}(n = 30, p = 1/6)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 30 \cdot 1/6 \\ &= 5 \end{aligned}$$

Question 57: You are flipping a coin 100 times. What is the probability that the number of tails is greater than or equal to 98? 19/20 is the probability of a head on each coin-flip.

Answer 57:
Let X be the number of tails. $X \sim \text{Bin}(n = 100, p = 1/20)$

$$\begin{aligned} \text{P}(X >= 98) &= P(98 <= X <= 100) \\ &= \sum_{i=98}^{100} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

Question 58: Irina is running a server cluster. What is the probability that the number of crashes is less than 99? the server has 100 computers which crash independently and the probability of a computer continuing to work on a given server is 22/25.

Answer 58:
Let X be the number of crashes. $X \sim \text{Bin}(n = 100, p = 3/25)$

$$\begin{aligned}\text{P}(X < 99) &= 1 - \text{P}(99 <= X <= 100) \\ &= 1 - \sum_{i=99}^{100} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 59: You are manufacturing chairs and are testing for defects. You test 100 objects. 1/2 is the probability of a non-defect on each test. What is the probability that the number of defects is not greater than 97?

Answer 59:
Let X be the number of defects. $X \sim \text{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\text{P}(X <= 97) &= 1 - \text{P}(98 <= X <= 100) \\ &= 1 - \sum_{i=98}^{100} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 60: In a restaurant some customers ask for a water with their meal. There are 50 customers. You can assume each customer is independent. 0.2 is the probability of a water requested by each customer. What is the expected number of waters requested?

Answer 60:
Let X be the number of waters requested. $X \sim \text{Bin}(n = 50, p = 0.2)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 50 \cdot 0.2 \\ &= 10.0\end{aligned}$$

Question 61: You are showing an online-ad to 40 people. 1/4 is the probability of an ad ignore on each ad shown. What is the probability that the number of ad clicks is 9?

Answer 61:
Let X be the number of ad clicks. $X \sim \text{Bin}(n = 40, p = 3/4)$

$$\begin{aligned}\text{P}(X = 9) &= \binom{n}{9} p^9 (1-p)^{n-9} \\ &= \binom{40}{9} 3/4^9 (1 - 3/4)^{40-9} \\ &< 0.00001\end{aligned}$$

Question 62: 100 trials are run. Each trial has a 22/25 probability of resulting in a failure. What is the standard deviation of successes?

Answer 62:
Let X be the number of successes. $X \sim \text{Bin}(n = 100, p = 3/25)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 3/25 \cdot (1 - 3/25)} \\ &= 3.25\end{aligned}$$

Question 63: A machine learning algorithm makes binary predictions. There are 12 independent guesses where the probability of a incorrect prediction on a given guess is 1/6. What is the expected number of correct predictions?

Answer 63:
Let X be the number of correct predictions. $X \sim \text{Bin}(n = 12, p = 5/6)$

$$\begin{aligned}\mathrm{E}[X] &= np \\ &= 12 \cdot 5/6 \\ &= 10\end{aligned}$$

Question 64: Waddie is showing an online-ad to customers. 1/2 is the probability of an ad click on each ad shown. The add is shown to 100 people. What is the average number of ad clicks?

Answer 64:
Let X be the number of ad clicks. $X \sim \mathrm{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\mathrm{E}[X] &= np \\ &= 100 \cdot 1/2 \\ &= 50\end{aligned}$$

Question 65: Charlotte is testing a new medicine on 50 patients. The probability of a cured patient on a given trial is 1/5. What is the probability that the number of cured patients is 12?

Answer 65:
Let X be the number of cured patients. $X \sim \mathrm{Bin}(n = 50, p = 1/5)$

$$\begin{aligned}\mathrm{P}(X = 12) &= \binom{n}{12} p^{12}(1-p)^{n-12} \\ &= \binom{50}{12} 1/5^{12}(1-1/5)^{50-12} \\ &= 0.10328\end{aligned}$$

Question 66: You are running in an election. The number of votes for you can be represented by a random variable X. X is a Bin(n = 50, p = 0.4). What is P(X is exactly 8)?

Answer 66:
Let X be the number of votes for you. $X \sim \mathrm{Bin}(n = 50, p = 0.4)$

$$\begin{aligned}\mathrm{P}(X = 8) &= \binom{n}{8} p^{8}(1-p)^{n-8} \\ &= \binom{50}{8} 0.4^{8}(1-0.4)^{50-8} \\ &= 0.00017\end{aligned}$$

Question 67: Irina is flipping a coin 100 times. The probability of a head on a given coin-flip is 1/2. What is the probability that the number of tails is less than or equal to 99?

Answer 67:
Let X be the number of tails. $X \sim \mathrm{Bin}(n = 100, p = 0.5)$

$$\begin{aligned}\mathrm{P}(X <= 99) &= 1 - \mathrm{P}(100 <= X <= 100) \\ &= 1 - \binom{n}{100} p^{1}00(1-p)^{n-100}\end{aligned}$$

Question 68: You are manufacturing airplanes and are testing for defects. You test 30 objects and the probability of a defect on a given test is 5/6. What is the probability that the number of defects is 14?

Answer 68:
Let X be the number of defects. $X \sim \mathrm{Bin}(n = 30, p = 5/6)$

$$\begin{aligned}\mathrm{P}(X = 14) &= \binom{n}{14} p^{14}(1-p)^{n-14} \\ &= \binom{30}{14} 5/6^{14}(1-5/6)^{30-14} \\ &< 0.00001\end{aligned}$$

Question 69: You are flipping a coin 20 times. The number of heads can be represented by a random variable X. X is a Binomial with 20 trials. Each trial is a success, independently, with probability 1/4. What is the standard deviation of X?

Answer 69:
Let X be the number of heads. $X \sim \text{Bin}(n = 20, p = 1/4)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{20 \cdot 1/4 \cdot (1 - 1/4)} \\ &= 1.94\end{aligned}$$

Question 70: You are giving a survey question where responses are "like" or "dislike" to 100 people. What is the probability that X is equal to 4? The number of likes can be represented by a random variable X. X is a Bin(100, 0.5).

Answer 70:
Let X be the number of likes. $X \sim \text{Bin}(n = 100, p = 0.5)$

$$\begin{aligned}\text{P}(X = 4) &= \binom{n}{4} p^4 (1-p)^{n-4} \\ &= \binom{100}{4} 0.5^4 (1 - 0.5)^{100-4} \\ &< 0.00001\end{aligned}$$

Question 71: You are flipping a coin. There are 20 independent coin-flips where the probability of a tail on a given coin-flip is 0.9. What is the standard deviation of tails?

Answer 71:
Let X be the number of tails. $X \sim \text{Bin}(n = 20, p = 0.9)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{20 \cdot 0.9 \cdot (1 - 0.9)} \\ &= 1.34\end{aligned}$$

Question 72: You are flipping a coin. There are 50 independent coin-flips. The probability of a tail on a given coin-flip is 4/5. What is the expectation of heads?

Answer 72:
Let X be the number of heads. $X \sim \text{Bin}(n = 50, p = 1/5)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 50 \cdot 1/5 \\ &= 10\end{aligned}$$

Question 73: You are giving a survey question where responses are "like" or "dislike" to 100 people. What is the standard deviation of likes? the probability of a dislike on each response is 41/50.

Answer 73:
Let X be the number of likes. $X \sim \text{Bin}(n = 100, p = 9/50)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{100 \cdot 9/50 \cdot (1 - 9/50)} \\ &= 3.84\end{aligned}$$

Question 74: In a restaurant some customers ask for a water with their meal. 0.6 is the probability of a water requested by each customer and there are 30 independent customers. What is the expected number of waters requested?

Answer 74:
Let X be the number of waters requested. $X \sim \text{Bin}(n = 30, p = 0.6)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 30 \cdot 0.6 \\ &= 18.0 \end{aligned}$$

Question 75: There are 40 independent trials and 0.5 is the probability of a failure on each trial. What is the expectation of successes?

Answer 75:
Let X be the number of successes. $X \sim \text{Bin}(n = 40, p = 1/2)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 40 \cdot 1/2 \\ &= 20 \end{aligned}$$

Question 76: Imran is showing an online-ad to 30 people. 5/6 is the probability of an ad click on each ad shown. What is the standard deviation of ad clicks?

Answer 76:
Let X be the number of ad clicks. $X \sim \text{Bin}(n = 30, p = 5/6)$

$$\begin{aligned} \text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{30 \cdot 5/6 \cdot (1 - 5/6)} \\ &= 2.04 \end{aligned}$$

Question 77: You are running a server cluster. What is the probability that the number of crashes is 1? the server has 30 computers which crash independently and each server has a 1/3 probability of resulting in a crash.

Answer 77:
Let X be the number of crashes. $X \sim \text{Bin}(n = 30, p = 1/3)$

$$\begin{aligned} \text{P}(X = 1) &= \binom{n}{1} p^1 (1-p)^{n-1} \\ &= \binom{30}{1} 1/3^1 (1 - 1/3)^{30-1} \\ &= 0.00008 \end{aligned}$$

Question 78: Cody is running a server cluster with 40 computers. What is P(X <= 39)? The number of crashes can be represented by a random variable X. X is a Bin(n = 40, p = 3/4).

Answer 78:
Let X be the number of crashes. $X \sim \text{Bin}(n = 40, p = 3/4)$

$$\begin{aligned} \text{P}(X <= 39) &= 1 - \text{P}(40 <= X <= 40) \\ &= 1 - \binom{n}{40} p^4 0 (1-p)^{n-40} \end{aligned}$$

Question 79: You are hashing strings into a hashtable. 5/6 is the probability of a hash to the first bucket on each string hash. There are 30 independent string hashes. What is the probability that the number of hashes to the first bucket is greater than or equal to 29?

Answer 79:
Let X be the number of hashes to the first bucket. $X \sim \text{Bin}(n = 30, p = 5/6)$

$$\begin{aligned}\mathrm{P}(X >= 29) &= P(29 <= X <= 30) \\ &= \sum_{i=29}^{30} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 80: Irina is flipping a coin. Irina flips the coin 30 times and the probability of a head on each coin-flip is 0.4. What is the probability that the number of tails is 19?

Answer 80:
Let X be the number of tails. $X \sim \mathrm{Bin}(n = 30, p = 0.6)$

$$\begin{aligned}\mathrm{P}(X = 19) &= \binom{n}{19} p^{19} (1-p)^{n-19} \\ &= \binom{30}{19} 0.6^{19} (1-0.6)^{30-19} \\ &= 0.13962\end{aligned}$$

Question 81: You are asking a survey question where responses are "like" or "dislike". The probability of a like on a given response is 1/2. You give the survey to 100 people. What is the probability that the number of likes is not less than 2?

Answer 81:
Let X be the number of likes. $X \sim \mathrm{Bin}(n = 100, p = 1/2)$

$$\begin{aligned}\mathrm{P}(X >= 2) &= 1 - \mathrm{P}(0 <= X <= 1) \\ &= 1 - \sum_{i=0}^{1} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Question 82: Wind blows independently across locations. The number of independent locations is 100. The probability of wind at a given location is 3/20. What is the probability that the number of locations that have wind is 93?

Answer 82:
Let X be the number of locations that have wind. $X \sim \mathrm{Bin}(n = 100, p = 3/20)$

$$\begin{aligned}\mathrm{P}(X = 93) &= \binom{n}{93} p^{93} (1-p)^{n-93} \\ &= \binom{100}{93} 3/20^{93} (1-3/20)^{100-93} \\ &< 0.00001\end{aligned}$$

Question 83: You are flipping a coin. 0.9 is the probability of a tail on each coin-flip. You flip the coin 50 times. What is the expected number of heads?

Answer 83:
Let X be the number of heads. $X \sim \mathrm{Bin}(n = 50, p = 0.1)$

$$\begin{aligned}\mathrm{E}[X] &= np \\ &= 50 \cdot 0.1 \\ &= 5.0\end{aligned}$$

Question 84: A machine learning algorithm makes binary predictions. What is the probability that the number of correct predictions is less than or equal to 0? the probability of a incorrect prediction on a given guess is 1/4. The number of independent guesses is 40.

Answer 84:
Let X be the number of correct predictions. $X \sim \mathrm{Bin}(n = 40, p = 3/4)$

$$\begin{aligned}\mathrm{P}(X <= 0) &= P(0 <= X <= 0) \\ &= \binom{n}{0} p^0 (1-p)^{n-0}\end{aligned}$$

Question 85: Wind blows independently across 20 locations. 1/2 is the probability of wind at each location. What is the standard deviation of locations that have wind?

Answer 85:
Let X be the number of locations that have wind. $X \sim \text{Bin}(n = 20, p = 1/2)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{20 \cdot 1/2 \cdot (1 - 1/2)} \\ &= 2.24\end{aligned}$$

Question 86: 7/10 is the probability of a failure on each trial and the number of independent trials is 100. What is the probability that the number of successes is 7?

Answer 86:
Let X be the number of successes. $X \sim \text{Bin}(n = 100, p = 0.3)$

$$\begin{aligned}\text{P}(X = 7) &= \binom{n}{7} p^7 (1-p)^{n-7} \\ &= \binom{100}{7} 0.3^7 (1 - 0.3)^{100-7} \\ &< 0.00001\end{aligned}$$

Question 87: You generate random bit strings. What is the expectation of 1s? there are 100 independent bits and 0.1 is the probability of a 1 at each bit.

Answer 87:
Let X be the number of 1s. $X \sim \text{Bin}(n = 100, p = 0.1)$

$$\begin{aligned}\text{E}[X] &= np \\ &= 100 \cdot 0.1 \\ &= 10.0\end{aligned}$$

Question 88: You are testing a new medicine on patients. 3/5 is the probability of a cured patient on each trial. There are 30 independent trials. What is the probability that the number of cured patients is greater than or equal to 1?

Answer 88:
Let X be the number of cured patients. $X \sim \text{Bin}(n = 30, p = 3/5)$

$$\begin{aligned}\text{P}(X >= 1) &= 1 - \text{P}(0 <= X <= 0) \\ &= 1 - \binom{n}{0} p^0 (1-p)^{n-0}\end{aligned}$$

Question 89: A student is guessing randomly on an exam. 0.9 is the probability of a correct answer on each question and the test has 20 questions. What is the standard deviation of correct answers?

Answer 89:
Let X be the number of correct answers. $X \sim \text{Bin}(n = 20, p = 0.9)$

$$\begin{aligned}\text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{20 \cdot 0.9 \cdot (1 - 0.9)} \\ &= 1.34\end{aligned}$$

Question 90: A student is guessing randomly on an exam with 40 questions. What is the probability that the number of correct answers is 32? 0.5 is the probability of a correct answer on each question.

Answer 90:
Let X be the number of correct answers. $X \sim \text{Bin}(n = 40, p = 0.5)$

$$\begin{aligned} \text{P}(X = 32) &= \binom{n}{32} p^{32}(1-p)^{n-32} \\ &= \binom{40}{32} 0.5^{32}(1-0.5)^{40-32} \\ &= 0.00007 \end{aligned}$$

Question 91: In a restaurant some customers ask for a water with their meal. A random sample of 40 customers is selected where the probability of a water not requested by a given customer is 1/4. What is the standard deviation of waters requested?

Answer 91:
Let X be the number of waters requested. $X \sim \text{Bin}(n = 40, p = 3/4)$

$$\begin{aligned} \text{Std}(X) &= \sqrt{np(1-p)} \\ &= \sqrt{40 \cdot 3/4 \cdot (1-3/4)} \\ &= 2.74 \end{aligned}$$

Question 92: A machine learning algorithm makes binary predictions. The number of correct predictions can be represented by a random variable X. X is a Bin(n = 30, p = 2/5). What is P(X < 27)?

Answer 92:
Let X be the number of correct predictions. $X \sim \text{Bin}(n = 30, p = 2/5)$

$$\begin{aligned} \text{P}(X < 27) &= 1 - \text{P}(27 <= X <= 30) \\ &= 1 - \sum_{i=27}^{30} \binom{n}{i} p^{i}(1-p)^{n-i} \end{aligned}$$

Question 93: Irina is flipping a coin. The probability of a tail on each coin-flip is 3/4. The number of independent coin-flips is 40. What is the probability that the number of tails is greater than 0?

Answer 93:
Let X be the number of tails. $X \sim \text{Bin}(n = 40, p = 3/4)$

$$\begin{aligned} \text{P}(X > 0) &= 1 - \text{P}(0 <= X <= 0) \\ &= 1 - \binom{n}{0} p^{0}(1-p)^{n-0} \end{aligned}$$

Question 94: Waddie is sending a stream of 50 bits to space. The probability of a no corruption on a given bit is 1/2. What is the expectation of corruptions?

Answer 94:
Let X be the number of corruptions. $X \sim \text{Bin}(n = 50, p = 0.5)$

$$\begin{aligned} \text{E}[X] &= np \\ &= 50 \cdot 0.5 \\ &= 25.0 \end{aligned}$$

Question 95: You are hashing strings into a hashtable. There are 30 independent string hashes where the probability of a hash to the first bucket on each string hash is 5/6. What is the probability that the number of hashes to the first bucket is 24?

Answer 95:
Let X be the number of hashes to the first bucket. $X \sim \text{Bin}(n = 30, p = 5/6)$

$$\begin{aligned}\mathrm{P}(X = 24) &= \binom{n}{24} p^{24}(1-p)^{n-24} \\ &= \binom{30}{24} 5/6^{24}(1-5/6)^{30-24} \\ &= 0.16009\end{aligned}$$

Question 96: Charlotte is hashing strings into a hashtable. 100 strings are hashed and the probability of a hash to the first bucket on a given string hash is 1/5. What is the probability that the number of hashes to the first bucket is greater than or equal to 1?

Answer 96:
Let X be the number of hashes to the first bucket. $X \sim \mathrm{Bin}(n = 100, p = 1/5)$

$$\begin{aligned}\mathrm{P}(X >= 1) &= 1 - \mathrm{P}(0 <= X <= 0) \\ &= 1 - \binom{n}{0} p^{0}(1-p)^{n-0}\end{aligned}$$

Question 97: You are flipping a coin. Each coin-flip has a 3/10 probability of resulting in a head and there are 100 coin-flips. You can assume each coin-flip is independent. What is the probability that the number of heads is 0?

Answer 97:
Let X be the number of heads. $X \sim \mathrm{Bin}(n = 100, p = 3/10)$

$$\begin{aligned}\mathrm{P}(X = 0) &= \binom{n}{0} p^{0}(1-p)^{n-0} \\ &= \binom{100}{0} 3/10^{0}(1-3/10)^{100-0} \\ &< 0.00001\end{aligned}$$

Question 98: Chris is sending a stream of 50 bits to space. 16/25 is the probability of a no corruption on each bit. What is the probability that the number of corruptions is greater than or equal to 47?

Answer 98:
Let X be the number of corruptions. $X \sim \mathrm{Bin}(n = 50, p = 9/25)$

$$\begin{aligned}\mathrm{P}(X >= 47) &= P(47 <= X <= 50) \\ &= \sum_{i=47}^{50} \binom{n}{i} p^{i}(1-p)^{n-i}\end{aligned}$$

Question 99: You are flipping a coin 30 times. What is the probability that the number of tails is less than 29? the probability of a tail on a given coin-flip is 2/3.

Answer 99:
Let X be the number of tails. $X \sim \mathrm{Bin}(n = 30, p = 2/3)$

$$\begin{aligned}\mathrm{P}(X < 29) &= 1 - \mathrm{P}(29 <= X <= 30) \\ &= 1 - \sum_{i=29}^{30} \binom{n}{i} p^{i}(1-p)^{n-i}\end{aligned}$$

Question 100: You are manufacturing chips and are testing for defects. There are 40 independent tests. The probability of a non-defect on a given test is 5/8. What is the probability that the number of defects is 10?

Answer 100:
Let X be the number of defects. $X \sim \mathrm{Bin}(n = 40, p = 3/8)$

$$\begin{aligned}
\mathrm{P}(X = 10) &= \binom{n}{10} p^{10}(1-p)^{n-10} \\
&= \binom{40}{10} 3/8^{10}(1-3/8)^{40-10} \\
&= 0.03507
\end{aligned}$$

Jury Selection

In the Supreme Court case: Berghuis v. Smith, the Supreme Court (of the US) discussed the question: "If a group is underrepresented in a jury pool, how do you tell?"

Justice Breyer [Stanford Alum] opened the questioning by invoking the binomial theorem. He hypothesized a scenario involving “an urn with a thousand balls, and sixty are red, and nine hundred forty are green, and then you select them at random… twelve at a time.” According to Justice Breyer and the binomial theorem, if the red balls were black jurors then “you would expect… something like a third to a half of juries would have at least one black person” on them.

Note: What is missing in this conversation is the power of diverse backgrounds when making difficult decisions.

Simulate

1. Simulation:

2. Explination:

Technically, since jurors are selected without replacement, you should represent the number of under-representative jurors as being a Hyper Geometric Random Variable (a random variable we don't look at explicitely in CS109) st

$$X \sim \text{HypGeo}(n = 12, N = 1000, m = 60)$$

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{\binom{60}{0}\binom{940}{12}}{\binom{1000}{12}} \\ &\approx 0.5261 \end{aligned}$$

However Justic Breyer made his case by citing a Binomial distribution. This isn't a perfect use of binomial, because the binomial assumes that each experiment has equal likelihood (p) of success. Because the jurors are selected without replacement, the probability of getting a minority juror changes slightly after each selection (and depending on what the selection was). However, as we will see, because the probabilities don't change too much the binomial distribution is not too far off.

$$X \sim \text{Binomial}(n = 12, p = 60/1000)$$

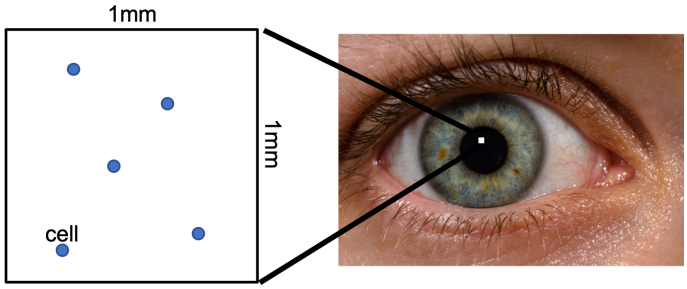
$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \binom{60}{0}(1 - 0.06)^{12} \\ &\approx 0.5241 \end{aligned}$$

Acknowledgements: Problem posed and solved by Mehran Sahami

Grading Eye Inflamation

When a patient has eye inflammation, eye doctors "grade" the inflammation. When "grading" inflammation they randomly look at a single 1 millimeter by 1 millimeter square in the patient's eye and count how many "cells" they see.

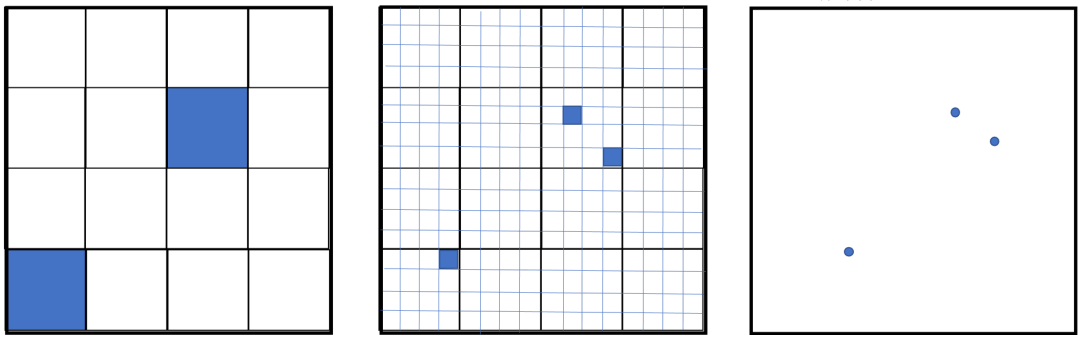
There is uncertainty in these counts. If the true average number of cells for a given patient's eye is 6, the doctor could get a different count (say 4, or 5, or 7) just by chance. As of 2021, modern eye medicine does not have a sense of uncertainty for their inflammation grades! In this problem we are going to change that. At the same time we are going to learn about \Poisson distributions over space.



Why is the number of cells observed in a 1x1 square governed by a \Poisson process?

We can approximate a distribution for the count by discretizing the square into a fixed number of equal sized buckets. Each bucket either has a cell or not. Therefore, the count of cells in the 1x1 square is a sum of Bernoulli random variables with equal p, and as such can be modeled as a binomial random variable. This is an approximation because it doesn't allow for two cells in one bucket. Just like with time, if we make the size of each bucket infinitely small, this limitation goes away and we converge on the true distribution of counts. The binomial in the limit, i.e. a binomial as $n \to \infty$, is truly represented by a \Poisson random variable. In this context, λ represents the average number of cells per 1×1 sample. See Figure 2.

$X \sim \text{Bin}(n = 16, p = \lambda/16)$ $X \sim \text{Bin}(n = 256, p = \lambda/256)$ $X \sim \lim_{n \to \infty} \text{Bin}(n, p = \lambda/n)$



For a given patient the true average rate of cells is 5 cells per 1x1 sample. What is the probability that in a single 1x1 sample the doctor counts 4 cells?

Let X denote the number of cells in the 1x1 sample. We note that $X \sim \text{Poi}(5)$. We want to find $P(X = 4)$.

$$P(X = 4) = \frac{5^4 e^{-5}}{4!} \approx 0.175$$

Multiple Observations

Heads up! This section uses concepts from Part 3. Specifically Independence in Variables

For a given patient the true average rate of cells is 5 cells per 1mm by 1mm sample. In an attempt to be more precise, the doctor counts cells in **two** different, larger **2mm by 2mm** samples. Assume that the occurrences of cells in one 2mm by 2mm samples are independent of the occurrences in any other 2mm by 2mm samples. What is the probability that she counts 20 cells in the first samples and 20 cells in the second?

Let Y_1 and Y_2 denote the number of cells in each of the 2x2 samples. Since there are 5 cells in a 1x1 sample, there are 20 samples in a 2x2 sample since the area quadrupled, so we have that $Y_1 \sim \text{Poi}(20)$ and $Y_2 \sim \text{Poi}(20)$. We want to find $P(Y_1 = 20 \wedge Y_2 = 20)$. Since the number of cells in the two samples are independent, this is equivalent to finding $\text{P}(Y_1 = 20)\,\text{P}(Y_2 = 20)$.

Estimating Lambda

Heads up! This section uses concepts from Part 5. Specifically Maximum A Posteriori

Inflammation prior: Based on millions of historical patients, doctors have learned that the prior probability density function of true rate of cells is:

$$f(\lambda) = K \cdot \lambda \cdot e^{-\frac{\lambda}{2}}$$

Where K is a normalization constant and λ must be greater than 0.

A doctor takes a single sample and counts 4 cells. Give an equation for the updated probability density of λ. Use the "Inflammation prior" as the prior probability density over values of λ. Your probability density may have a constant term.

Let θ be the random variable for true rate. Let X be the random variable for the count

$$\begin{aligned} f(\theta = \lambda | X = 4) &= \frac{P(X = 4 | \theta = \lambda) f(\theta = \lambda)}{P(X = 4)} \\ &= \frac{\frac{\lambda^4 e^{-\lambda}}{4!} \cdot K \cdot \lambda \cdot e^{\lambda/2}}{P(X = 4)} \\ &= \frac{K \cdot \lambda^5 e^{-\frac{3}{2}\lambda}}{4! P(X = 4)} \end{aligned}$$

A doctor takes a single sample and counts 4 cells. What is the Maximum A Posteriori estimate of λ?

Maximize the "posterior" of the parameter calculated in the previous section:

$$\arg\max_\lambda \frac{K \cdot \lambda^5 e^{-\frac{3}{2}\lambda}}{4! P(X = 4)} = \arg\max_\lambda \lambda^5 e^{-\frac{3}{2}\lambda}$$

Take logarithm (preserves argmax, and easier derivative):

$$\begin{aligned} &= \arg\max_\lambda \log\left(\lambda^5 e^{-\frac{3}{2}\lambda}\right) \\ &= \arg\max_\lambda \left(5 \log \lambda - \frac{3}{2}\lambda\right) \end{aligned}$$

Calculate the derivative with respect to the parameter, and set equal to 0

$$\begin{aligned} 0 &= \frac{\partial}{\partial \lambda}\left(5 \log \lambda - \frac{3}{2}\lambda\right) \\ 0 &= \frac{5}{\lambda} - \frac{3}{2} \\ \lambda &= \frac{10}{3} \end{aligned}$$

Explain, in words, the difference between the two estimates of lambda in the two previous parts.

The estimate in the first part is a ``distribution" (also called a soft estimate) whereas the estimate in the second part is a single value (also called a point estimate). The former contains information about confidence.

What is the MLE estimate of λ?

The MLE estimate doesn't use the prior belief. The MLE estimate for a poisson is simply the average of the observations. In this case the average of our single observation is 4. MLE is not a great tool for estimating our parameter from just one datapoint.

A patient comes on two separate days. The first day the doctor counts 5 cells, the second day the doctor counts 4 cells. Based only on this observation, and treating the true rates on the two days as independent, what is the probability that the patient's inflammation has gotten better (in other words, that their λ has decreased)?

Let θ_1 be the random variable for lambda on the first day and θ_2 be the random variable for lambda on the second day.

$$f(\theta_1 = \lambda | X = 5) = K_1 \cdot \lambda^6 e^{-\frac{3}{2}\lambda}$$
$$f(\theta_2 = \lambda | X = 4) = K_2 \cdot \lambda^5 e^{-\frac{3}{2}\lambda}$$

The question is asking what is $P(\theta_1 > \theta_2)$? There are a few ways to calculate this exactly:

$$\begin{aligned}
&\int_{\lambda_1=0}^{\infty} \int_{\lambda_2=0}^{\lambda_1} f(\theta_1 = \lambda_1, \theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} \int_{\lambda_2=0}^{\lambda_1} f(\theta_1 = \lambda_1) \cdot f(\theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} f(\theta_1 = \lambda_1) \int_{\lambda_2=0}^{\lambda_1} f(\theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} K_1 \cdot \lambda^6 e^{-\frac{3}{2}\lambda} \int_{\lambda_2=0}^{\lambda_1} K_2 \cdot \lambda^5 e^{-\frac{3}{2}\lambda}
\end{aligned}$$

Gaussian CDF Calculator

To calculate the Cumulative Density Function (CDF) for a normal (aka Gaussian) random variable at a value x, also writen as $F(x)$, you can transform your distribution to the "standard normal" and look up the corresponding value in the standard normal CDF. However, most programming libraries will provide a normal cdf funciton. This tool replicates said functionality.

1. Calculator

x: 0.25

mu: 1

std: 2

```
norm.cdf(x, mu, std)
```

```
= 0.3538
```

2. Explanation

This function calculates the cumulative density function of a Normal random variable. It is very important in CS109 to understand the difference between a probability density function (PDF), and a cumulative density function (CDF). The CDF of a random variable at point little x is equal to the probability that the random variable takes on a value less than or equal to x. If the random variable is called big X, the CDF can be written as $P(X < x)$ or as $F_X(x)$.

The CDF function of a Normal is calculated by translating the random variable to the Standard Normal, and then looking up a value from the precalculated "Phi" function (Φ), which is the cumulative density function of the standard normal. The Standard Normal, often written Z, is a Normal with mean 0 and variance 1. Thus, $Z \sim N(\mu = 0, \sigma^2 = 1)$.

For your normal $X \sim \mathcal{N}(\mu = 1, \sigma = 2)$

$P(X < 0.25) = P(\frac{X - \mu}{\sigma} < \frac{0.25-\mu}{\sigma}) = P(Z < \frac{0.25 - 1}{2}) = \Phi\left(\frac{0.25 - 1}{2}\right) = 0.3538$

Try different calculations to see different translations to the standard normal!

Grades are Not Normal

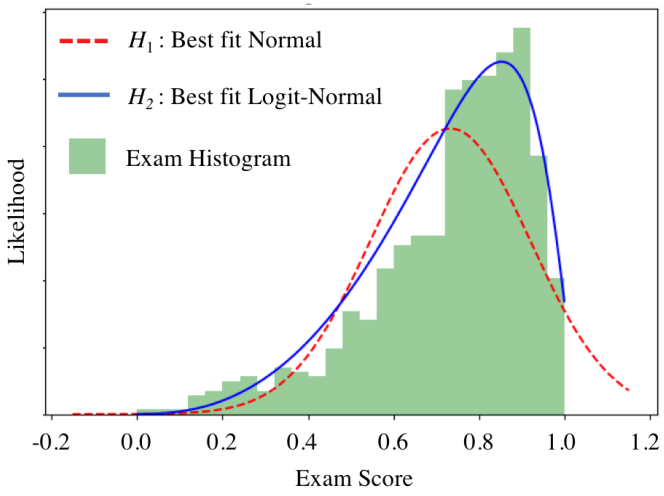
Sometimes you just feel like squashing normals:

Logit Normal
The logit normal is the continuous distribution that results from applying a special "squashing" function to a Normally distributed random variable. The squashing function maps all values the normal could take on onto the range 0 to 1. If $X \sim \text{LogitNormal}(\mu, \sigma^2)$ it has:

$$\begin{aligned}
\text{PDF:} \quad & f_X(x) = \begin{cases} \frac{1}{\sigma(\sqrt{2\pi})x(1-x)} e^{-\frac{(\text{logit}(x)-\mu)^2}{2\sigma^2}} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \\
\text{CDF:} \quad & F_X(x) = \Phi\Big(\frac{\text{logit}(x) - \mu}{\sigma}\Big) \\
\text{Where:} \quad & \text{logit}(x) = \log\Big(\frac{x}{1-x}\Big)
\end{aligned}$$

A new theory shows that the Logit Normal better fits exam score distributions than the traditionally used Normal. Let's test it out! We have some set of exam scores for a test with min possible score 0 and max possible score 1, and we are trying to decide between two hypotheses:

H_1: our grade scores are distributed according to $X \sim \text{Normal}(\mu = 0.7, \sigma^2 = 0.2^2)$.
H_2: our grade scores are distributed according to $X \sim \text{LogitNormal}(\mu = 1.0, \sigma^2 = 0.9^2)$.



Under the normal assumption, H_1, what is $P(0.9 < X < 1.0)$? Provide a numerical answer to two decimal places.

$$P(0.9 < X < 1.0) = \Phi\left(\frac{1.0 - 0.7}{0.2}\right) - \Phi\left(\frac{0.9 - 0.7}{0.2}\right) = \Phi(1.5) - \Phi(1.0) = 0.9332 - 0.8413 = 0.0$$

Under the logit-normal assumption, H_2, what is $P(0.9 < X < 1.0)$?

$$F_X(1.0) - F_X(0.9) = \Phi\Big(\frac{\text{logit}(1.0) - 1.0}{0.9}\Big) - \Phi\Big(\frac{\text{logit}(0.9) - 1.0}{0.9}\Big)$$

Which we can solve for numerically:

$$\Phi\Big(\frac{\text{logit}(1.0) - 1.0}{0.9}\Big) - \Phi\Big(\frac{\text{logit}(0.9) - 1.0}{0.9}\Big) = 1 - \Phi(1.33) \approx 0.91$$

Under the normal assumption, H_1, what is the maximum value that X can take on?

$$\infty$$

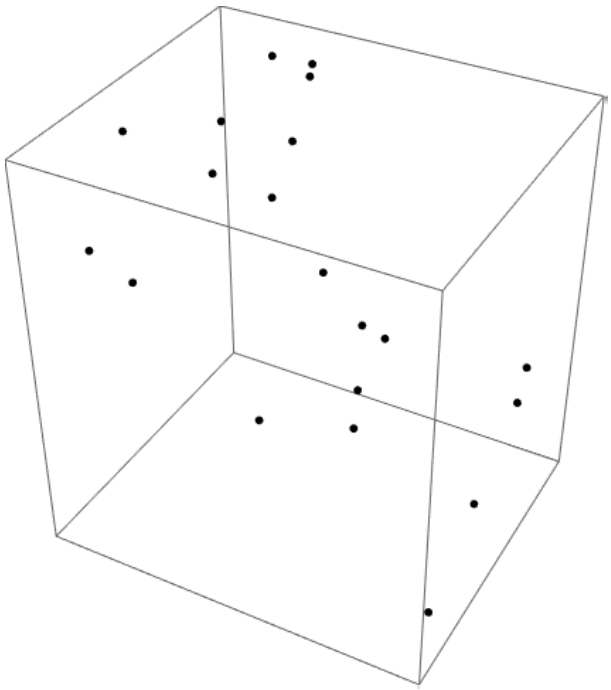
Before observing any test scores, you assume that (a) one of your two hypotheses is correct and (b) that initially, each hypothesis is equally likely to be correct, $P(H_1) = P(H_2) = \frac{1}{2}$. You then observe a single test score, $X = 0.9$. What is your updated probability that the Logit-Normal hypothesis is correct?

$$
\begin{aligned}
P(H_2|X=0.9) &= \frac{f(X=0.9|H_2)P(H_2)}{f(X=0.9|H_2)P(H_2) + f(X=0.9|H_1)P(H_1)} \\
&= \frac{f(X=0.9|H_2)}{f(X=0.9|H_2) + f(X=0.9|H_1)} \\
&= \frac{\frac{1}{\sigma(\sqrt{2\pi})0.9*(1-0.9)}e^{-\frac{(\text{logit}(0.9)-1.0)^2}{2*0.9^2}}}{\frac{1}{\sigma(\sqrt{2\pi})0.9*(1-0.9)}e^{-\frac{(\text{logit}(0.9)-1.0)^2}{2*0.9^2}} + \frac{1}{0.2\sqrt{2\pi}}e^{-\frac{(0.9-0.7)^2}{2*0.2^2}}}
\end{aligned}
$$

Curse of Dimensionality

In machine learning, like many fields of computer science, often involves high dimensional points, and high dimension spaces have some surprising probabilistic properties.

A random *value* X_i is a Uni(0, 1).
A random *point* of dimension d is a list of d random values: $[X_1 \dots X_d]$.



A random *value* X_i is close to an edge if X_i is less than 0.01 **or** X_i is greater than 0.99. What is the probability that a random value is close to an edge?

> Let E be the event that a random value is close to an edge.
> $P(E) = P(X_i < 0.01) + P(X_i > 0.99) = 0.02$

A random *point* $[X_1, X_2, X_3]$ of dimension 3 is close to an edge if *any* of it's values are close to an edge. What is the probability that a 3 dimensional point is close to an edge?

> The event is equivalent to the complement of none of the dimensions of the point is close to an edge, which is: $1 - (1 - P(E))^3 = 1 - 0.98^3 \approx 0.058$

A random *point* $[X_1, \dots X_{100}]$ of dimension 100 is close to an edge if *any* of it's values are close to an edge. What is the probability that a 100 dimensional point is close to an edge?

> Similarly, it is: $1 - (1 - P(E))^{100} = 1 - 0.98^{100} \approx 0.867$

There are many other phenomena of high dimensional points: such as, the euclidean distance between points starts to converge.

Probability and Babies

This demo used to be live. We now know that the delivery happened on Jan 23rd. Lets go back in time to Jan 1st and see what the probability looked like at that point.

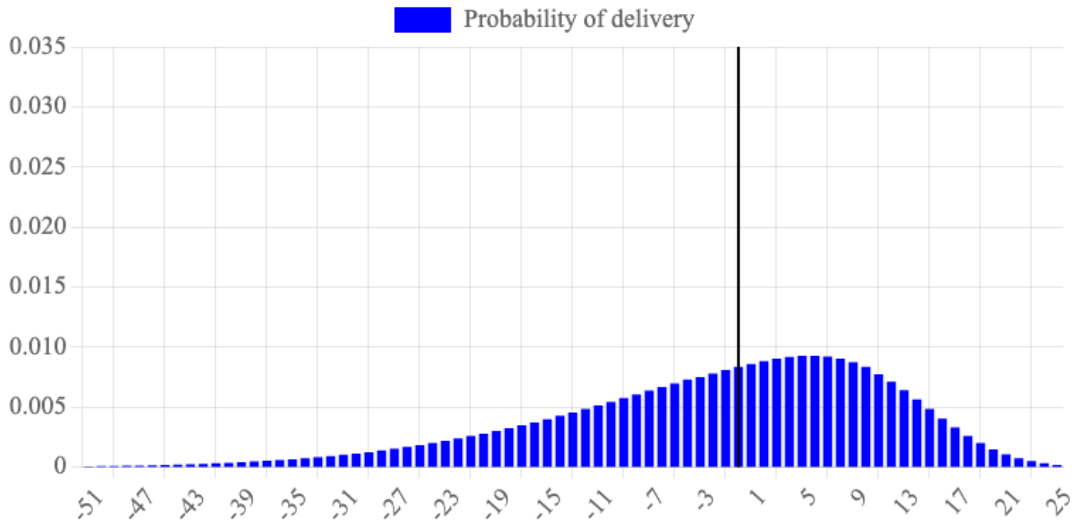
What is the probability that Laura gives birth today (given that she hasn't given birth up until today)?

Today's Date	1/Jan/2021

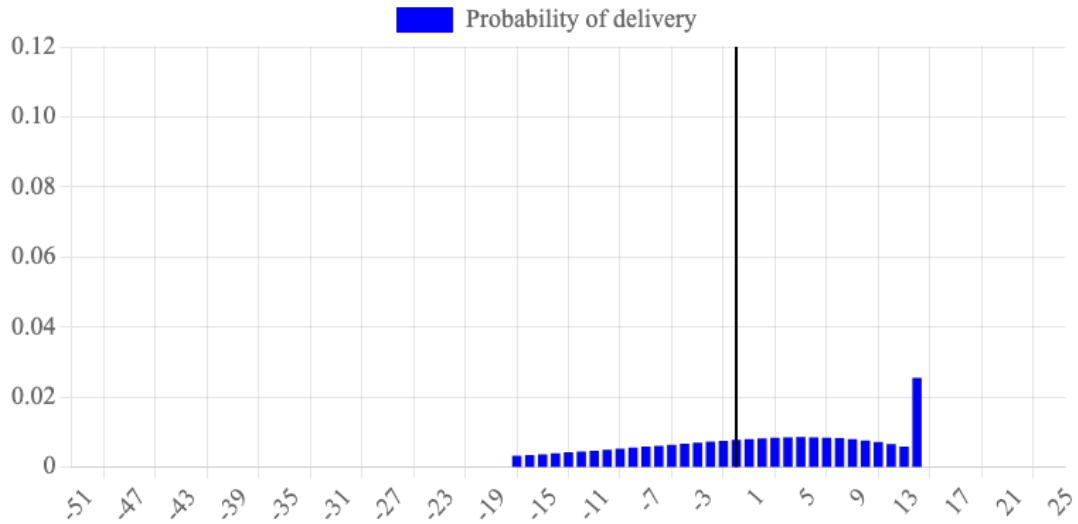
Due Date	18/Jan/2021

Probability of delivery today: **0.014**
Probability of delivery in next 7 days: **0.144**
Current days past due date: **-17** days
Unconditioned probability mass before today: **0.128**

How likely is delivery, in humans, relative to the due date? There have been millions of births which gives us a relatively good picture [1]. The length of human pregnancy varies by quite a lot! Have you heard that it is 9 months? That is a rough, point estimate. The mean duration of pregnancy is 278.6 days, and pregnancy length has a standard deviation (SD) of 12.5 days. This distribution is not normal, but roughly matches a "skewed normal". This is a general probability mass function for the first pregnancy collected from hundreds of thousands of women (this PMF is very similar across demographics, but changes based on whether the woman has given birth before):



Of course, we have more information. Specifically, we know that Laura **hasn't** given birth up until today (we will update this example when that changes). We also know that babies which are over 14 days late are "induced" on day 14. How likely is delivery given that we haven't delivered up until today? Note that the y-axis is scalled differently:



Implementation notes: this calculation was performed by storing the PDF as a list of (day, probability) points. These values are sometimes called weighted samples, or "particles" and are the key component to a "particle filtering" approach. After we observe no-delivery, we set the probability of every point which has a day before today to be 0, and then re-normalize the remaining points (aka we "filter" the "particles"). This is convenient because the "posterior" belief doesn't follow a simple equation -- using particles means we never have to write that equation down in our code.

Three friends have the exact same due date (Really! this isn't a hypothetical) What is the probability that all three couples deliver on the exact same day?

Probability of three couples on the same day: **0.002**

How did we get that number? Let p_i be the probability that one baby is delivered on day i -- this number can be read off the probability mass function. Let D_i be the event that all three babies are delivered on day i. Note that the event D_i is mutually exclusive with the event that all three babies are born on another day (So for example, D_1 is mutually exclusive with D_2, D_3 etc). Let $N = 3$ be the event that all babies are born on the same day:

$$\begin{aligned} \mathrm{P}(N=3) &= \sum_i \mathrm{P}(D_i) && \text{Since days are mutually exclusive} \\ &= \sum_i p_i^3 && \text{Since the three couples are independent} \end{aligned}$$

[1] Predicting delivery date by ultrasound and last menstrual period in early gestation

Acknowledgements: This problem was first posed to me by Chris Gregg.