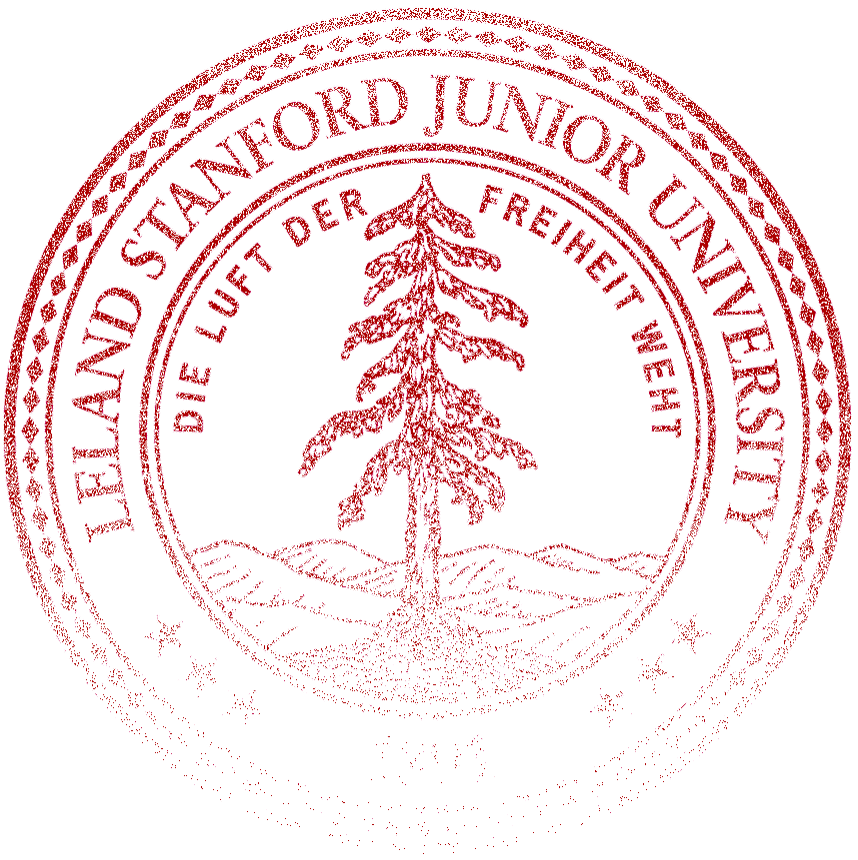
Course Reader for CS109



CS109
Department of Computer Science
Stanford University
Jan 2023
V 0.9

Get Started

Notable Updates Fall 2023:

1. General Inclusion-Exclusion. *Oct 7th 2023*
2. Core Probability Reference. *Oct 7th 2023*

Acknowledgements: This book was written by 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 to contributions. Want to make your mark? Keen to fix a typo? Download the github project and publish a pull request. We will credit all contributors.

Folks who have contributed to editing the book: GitHub Contributors. This includes Logan Bhamidipaty, Jonatan Pérez, Bobby Abraham, Tim Gianitsos, Yogi the Curious, Thanawan "Ly-Ly" Atchariyachanvanit, Kunal

Notation Reference

Core Probability

Notation	Meaning
E or F	Capital letters can denote events
A or B	Sometimes they denote sets
$\|E\|$	Size of an event or set
E^C	Complement of an event or set
EF	And of events (aka intersection)
E and F	And of events (aka intersection)
$E \cap F$	And of events (aka intersection)
E or F	Or of events (aka union)
$E \cup F$	Or of events (aka union)
$\text{count}(E)$	The number of times that E occurs
$\text{P}(E)$	The probability of an event E
$\text{P}(E\|F)$	The conditional probability of an event E given F
$\text{P}(E, F)$	The probability of event E and F
$\text{P}(E\|F, G)$	The conditional probability of an event E given both F and G
$n!$	n factorial
$\binom{n}{k}$	Binomial coefficient
$\binom{n}{r_1, r_2, r_3}$	Multinomial coefficient

Random Variables

Notation	Meaning
x or y or i	Lower case letters denote regular variables
X or Y	Capital letters are used to denote random variables
K	Capital K is reserved for constants
$\text{E}[X]$	Expectation of X

Notation	Meaning
$\text{Var}(X)$	Variance of X
$\text{P}(X = x)$	Probability mass function (PMF) of X, evaluated at x
$\text{P}(x)$	Probability mass function (PMF) of X, evaluated at x
$f(X = x)$	Probability density function (PDF) of X, evaluated at x
$f(x)$	Probability density function (PDF) of X, evaluated at x
$f(X = x, Y = y)$	Joint probability density
$f(X = x \vert Y = y)$	Conditional probability density
$F_X(x)$ or $F(x)$	Cumulative distribution function (CDF) of X
IID	Independent and Identically Distributed

Parametric Distributions

Notation	Meaning
$X \sim \text{Bern}(p)$	X is a Bernoulli random variable
$X \sim \text{Bin}(n, p)$	X is a Binomial random variable
$X \sim \text{Poi}(p)$	X is a Poisson random variable
$X \sim \text{Geo}(p)$	X is a Geometric random variable
$X \sim \text{NegBin}(r, p)$	X is a Negative Binomial random variable
$X \sim \text{Uni}(a, b)$	X is a Uniform random variable
$X \sim \text{Exp}(\lambda)$	X is a Exponential random variable
$X \sim \text{Beta}(a, b)$	X is a Beta random variable

Core Probability Reference

Definition: Empirical Definition of Probability

The probability of any event E can be defined as:

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

Where $\mathrm{count}(E)$ is the number of times that E occured in n experiments.

Definition: Core Identities

For an event E and a sample space S

$0 \leq \mathrm{P}(E) \leq 1$	All probabilities are numbers between 0 and 1.
$\mathrm{P}(S) = 1$	All outcomes must be from the Sample Space.
$\mathrm{P}(E) = 1 - \mathrm{P}(E^{\mathrm{C}})$	The probability of an event from its complement.

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:

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

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

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

Definition: Probability of **or** with Mututally Exclusive Events
If two events E and F are mutually exclusive then the probability of E ***or*** F occurring is:

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

For 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:

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

Definition: General Probability of **or** (Inclusion-Exclusion)
For any two events E and F:

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

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 more than three events see the chapter of probability of or.

Definition: Probability of **and** for Independent Events.
If two events: E, F are independent then the probability of E ***and*** F occurring is:

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

For n events $E_1, E_2, \ldots E_n$ that are independent of one another:

$$\mathrm{P}(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_n) = \prod_{i=1}^{n} \mathrm{P}(E_i)$$

Definition: General Probability of **and** (The Chain Rule)
For any two events E and F:

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

For n events $E_1, E_2, \ldots E_n$:

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

Definition: The Law of Total Probability
For any two events E and F:

$$\begin{aligned}\mathrm{P}(E) &= \mathrm{P}(E \text{ and } F) + \mathrm{P}(E \text{ and } F^{\mathrm{C}}) \\ &= \mathrm{P}(E|F)\,\mathrm{P}(F) + \mathrm{P}(E|F^{\mathrm{C}})\,\mathrm{P}(F^{\mathrm{C}})\end{aligned}$$

For mutually exclusive events: $B_1, B_2, \ldots B_n$ such that every outcome in the sample space falls into one of those events:

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

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

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

Bayes' Theorem combined with the Law of Total Probability:

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

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

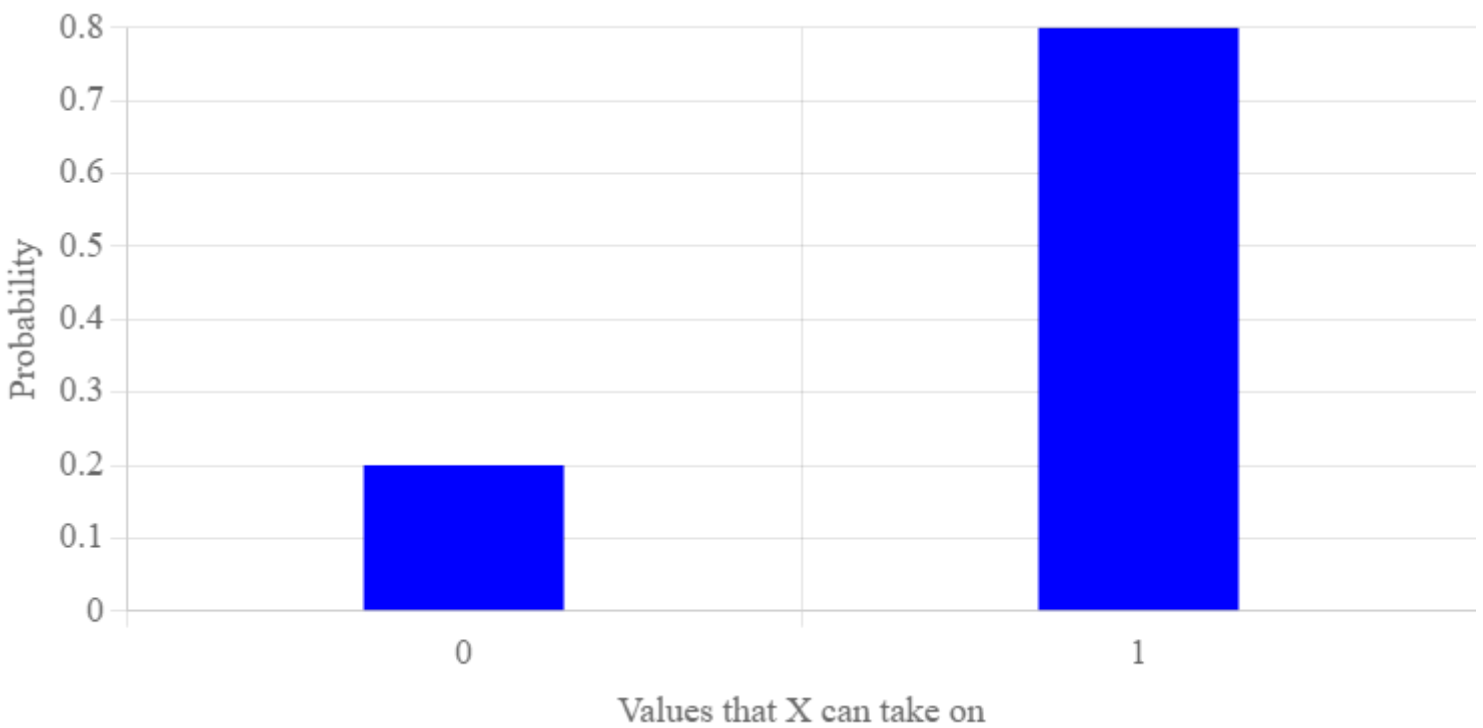
PMF (smooth): $\text{P}(X = x) = p^x (1-p)^{1-x}$

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

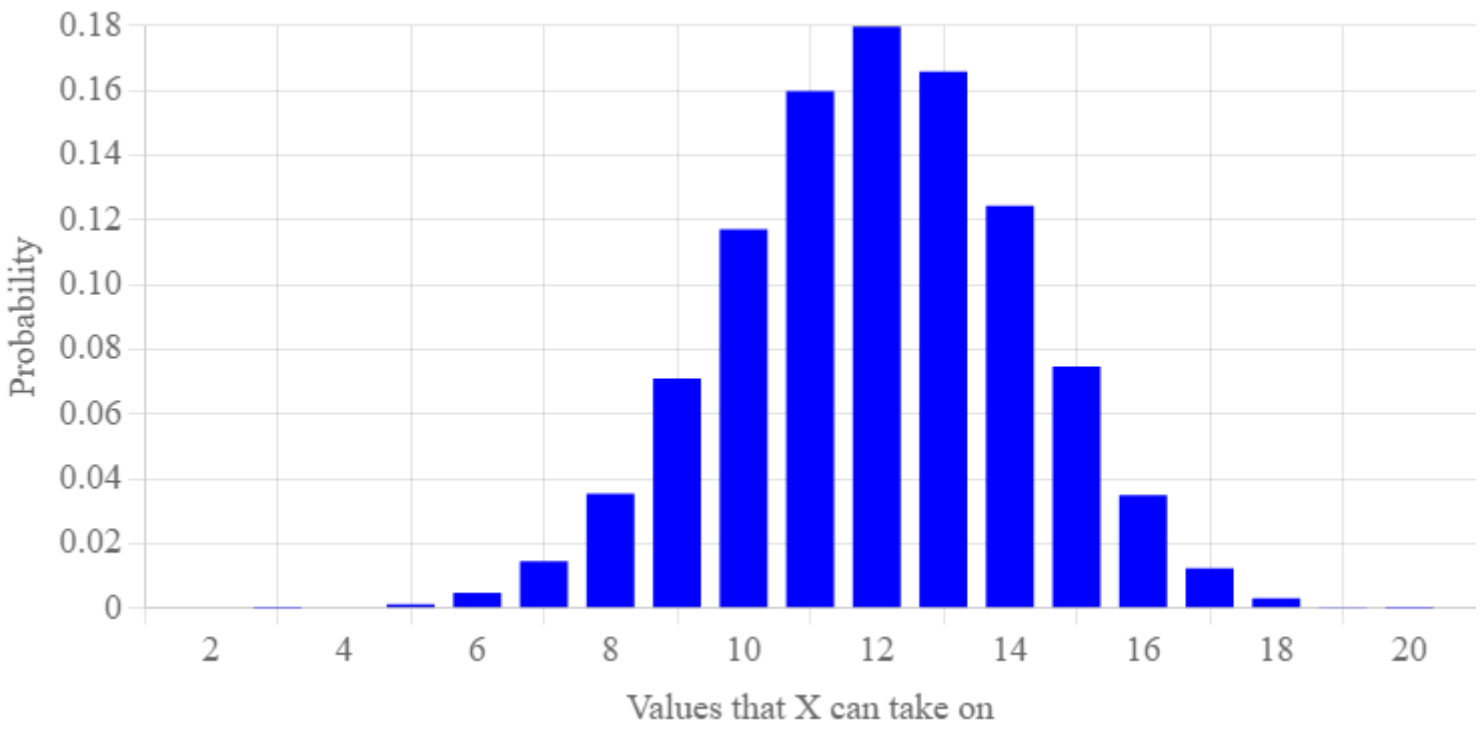
PMF equation: $$\text{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

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

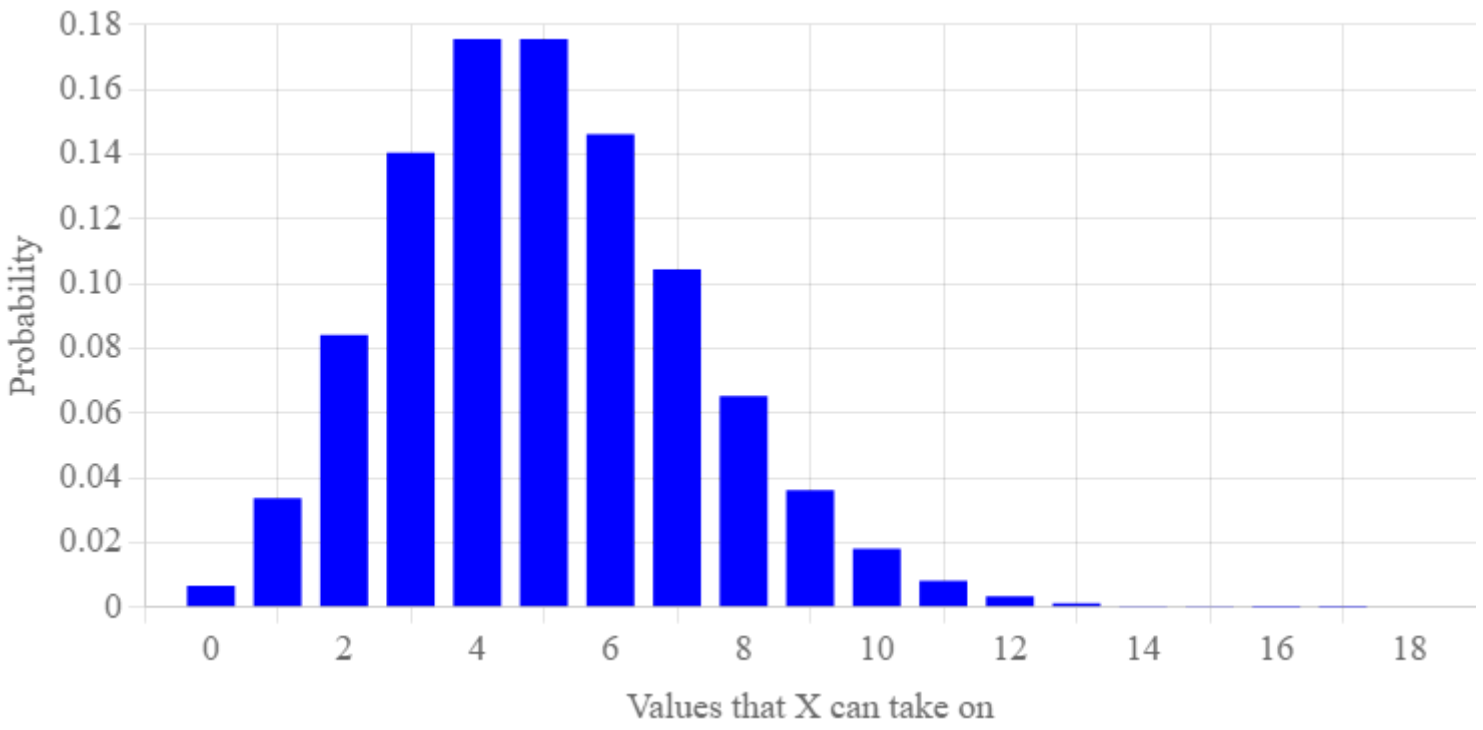
PMF equation: $\text{P}(X = x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

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: $x \in \{1, \ldots, \infty\}$

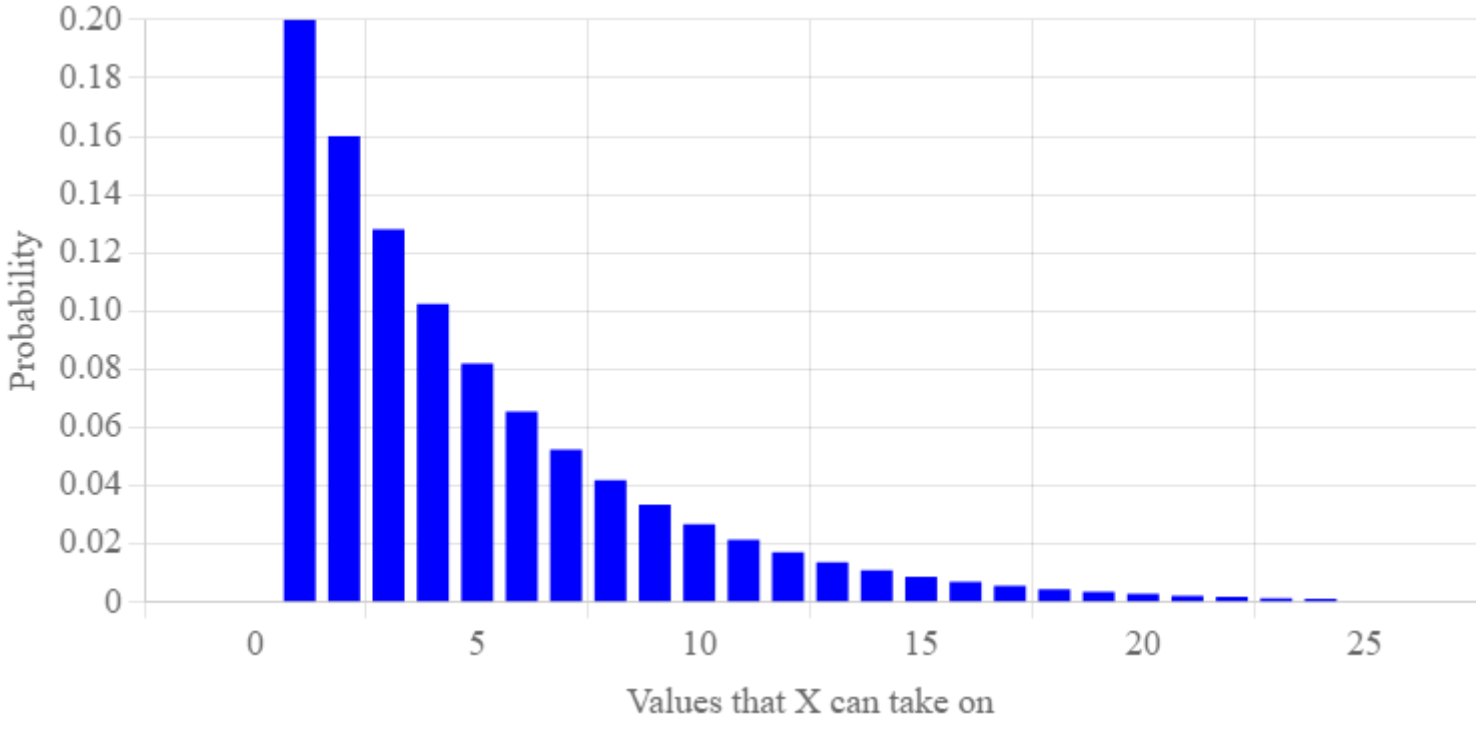
PMF equation: $\text{P}(X = x) = (1 - p)^{x-1} p$

Expectation: $\text{E}[X] = \frac{1}{p}$

Variance: $\text{Var}(X) = \frac{1-p}{p^2}$

PMF graph:

Parameter p: 0.20

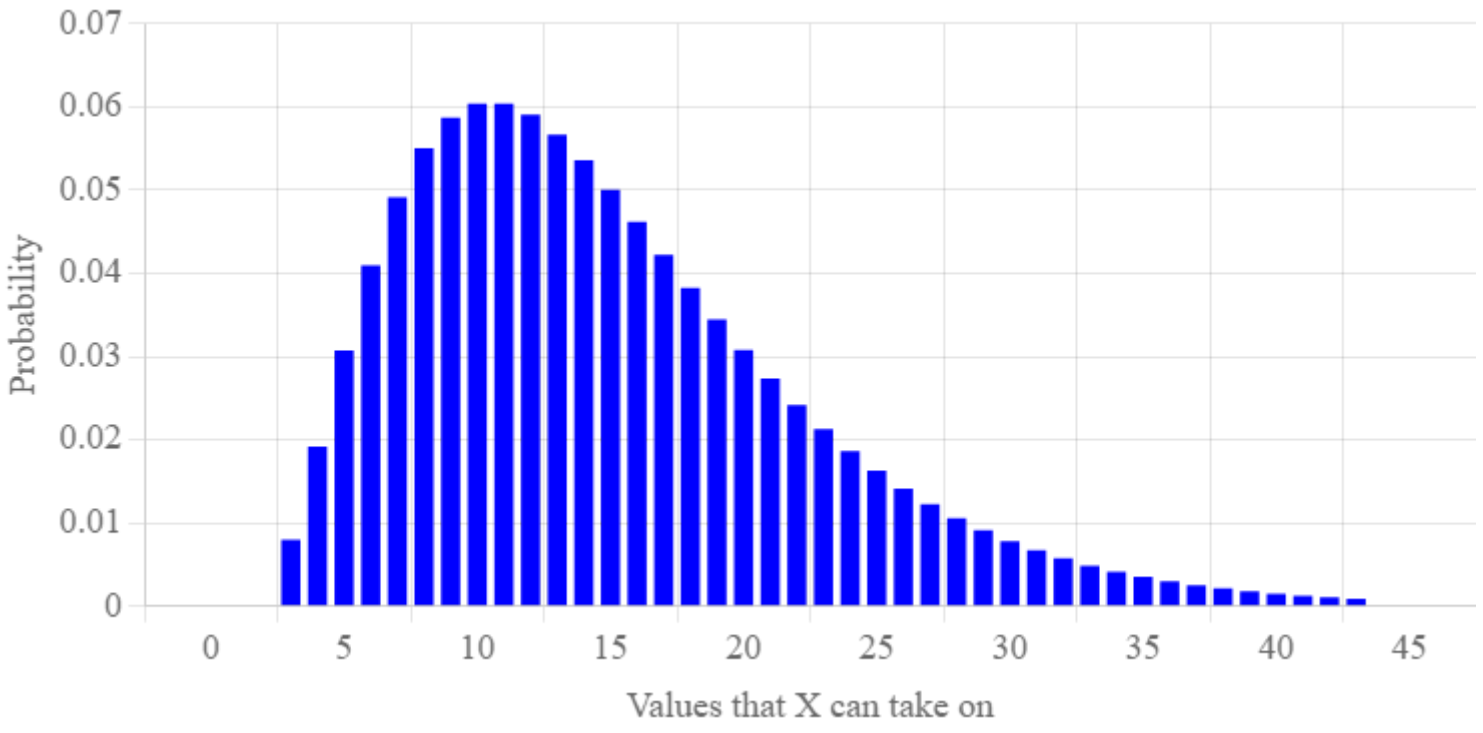


Negative Binomial Random Variable

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:	$x \in \{r, \ldots, \infty\}$
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: $f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{else} \end{cases}$

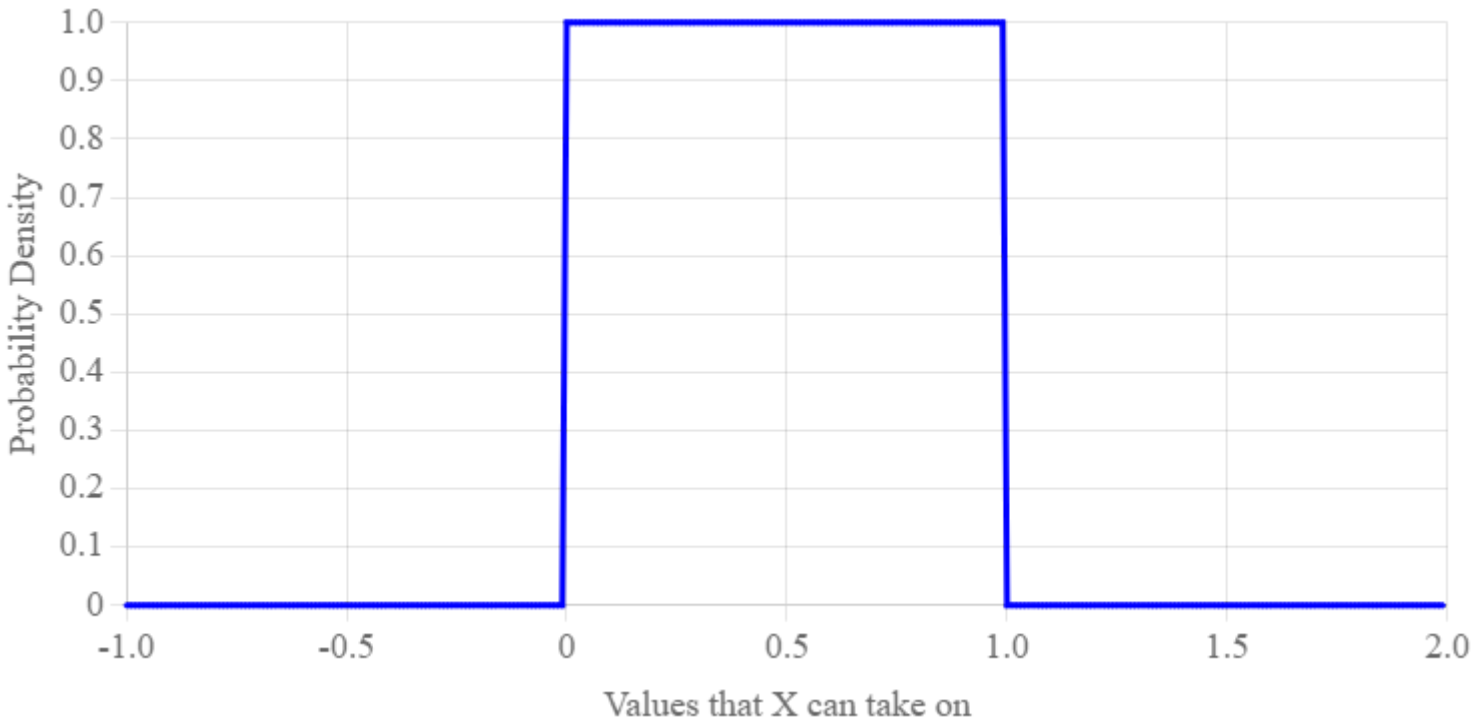
CDF equation: $F(x) = \begin{cases} \frac{x - \alpha}{\beta - \alpha} & \text{for } x \in [\alpha, \beta] \\ 0 & \text{for } x < \alpha \\ 1 & \text{for } x > \beta \end{cases}$

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

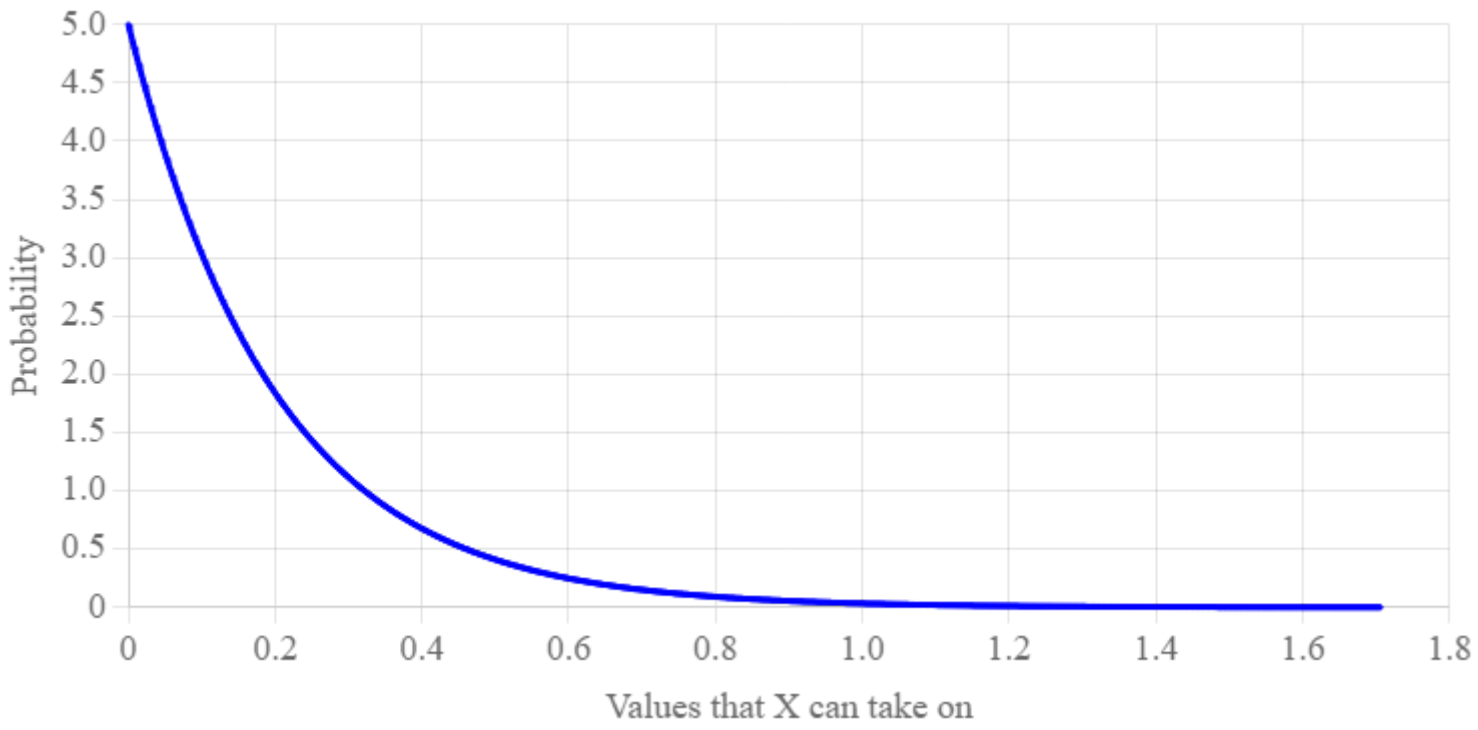
CDF equation: $F(x) = 1 - e^{-\lambda x}$

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 occurring distribution.

Parameters: $\mu \in \mathbb{R}$, the mean.

$\sigma^2 \in \mathbb{R}$, the variance.

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

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

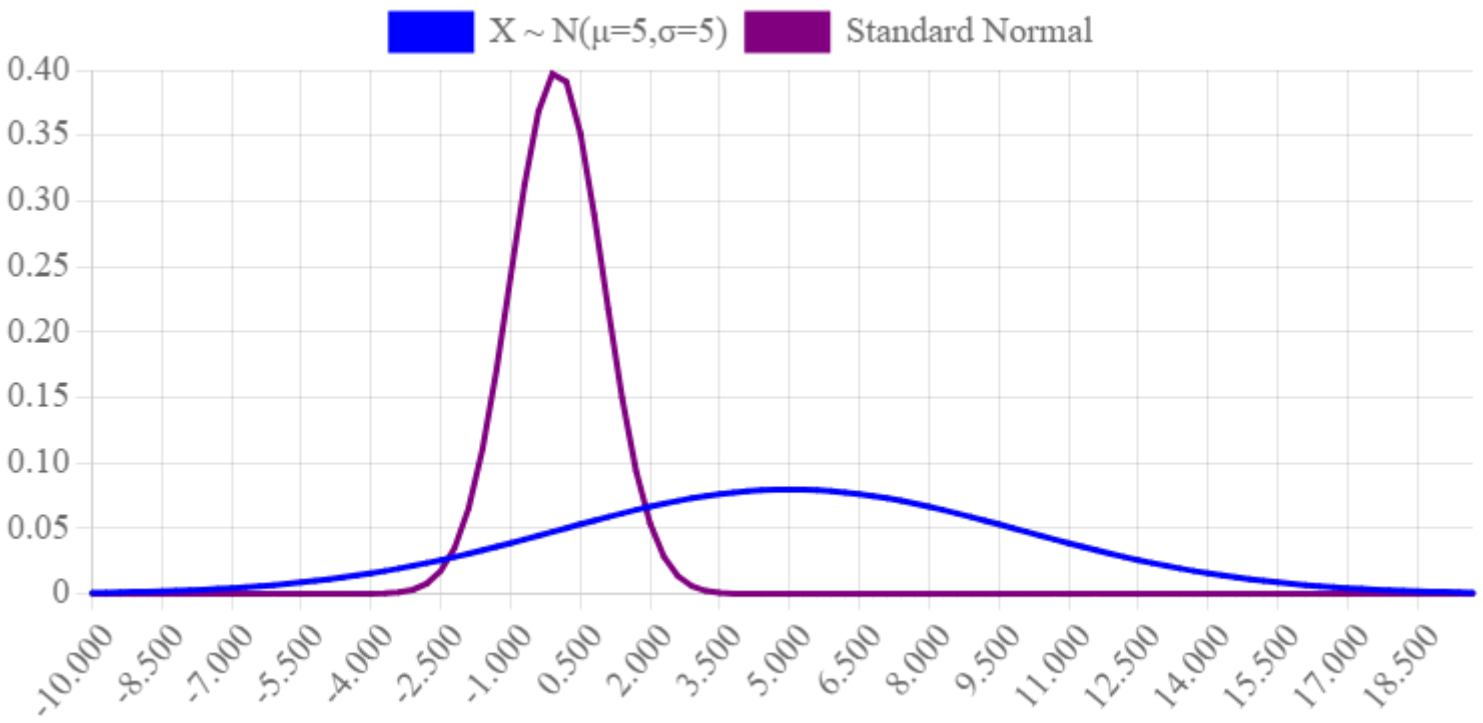
CDF equation: $F(x) = \phi(\frac{x-\mu}{\sigma})$ Where ϕ is the CDF of the standard normal

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

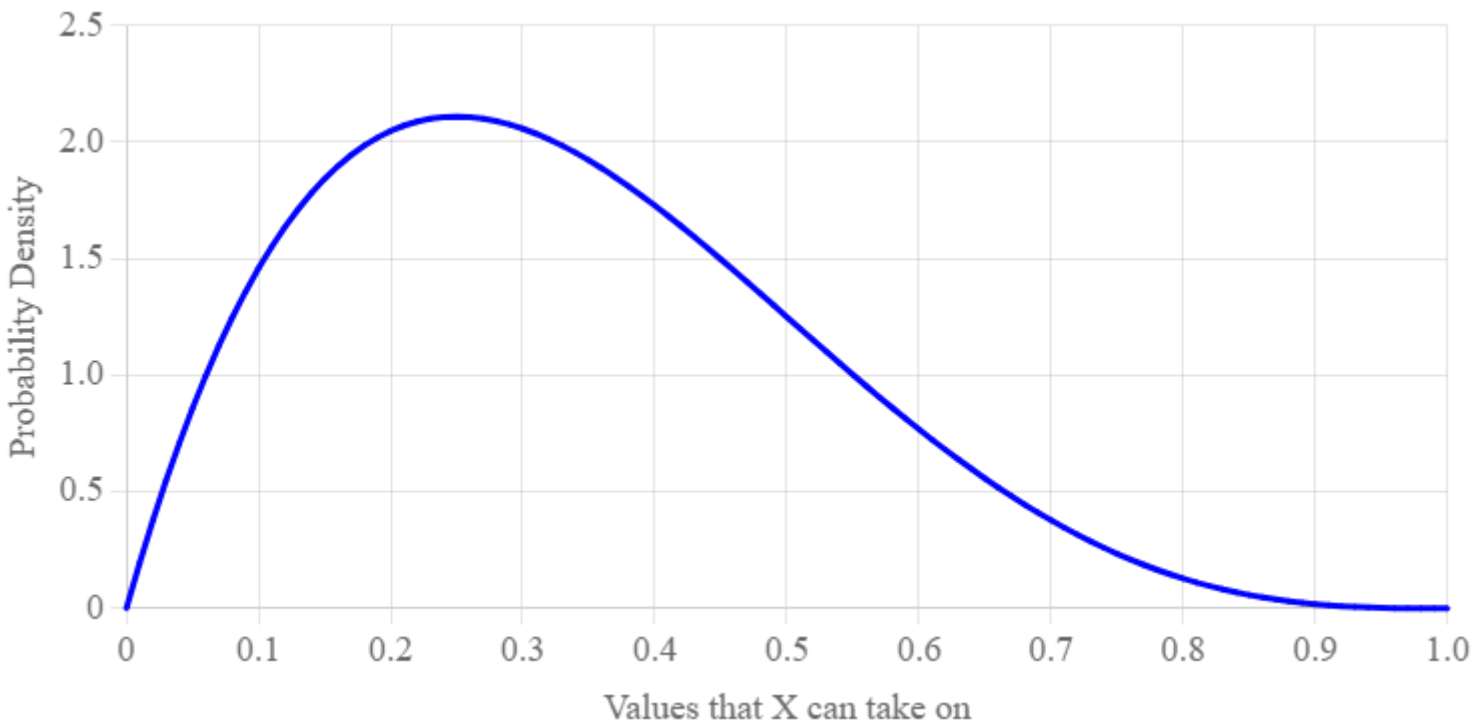
CDF equation: No closed form

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



Python Reference

Factorial

Compute $n!$ as an integer. This example computes 20!:

```
import math
print(math.factorial(20))
```

Choose

As of Python 3.8, you can compute $\binom{n}{m}$ from the math module. This example computes $\binom{10}{5}$:

```
import math
print(math.comb(10, 5))
```

Natural Exponent

Calculate e^x. For example this computes e^3

```
import math
print(math.exp(3))
```

SciPy Stats Library

SciPy is a free and open source library for scientific computing that is built on top of NumPy. You may find it helpful to use SciPy to check the answers you obtain in the written section of your problem sets. NumPy has the capability of drawing samples from many common distributions (type `help(np.random)` in the python interpreter), but SciPy has the added capability of computing the probability of observing events, and it can perform computations directly on the probability mass/density functions.

Binomial

Make a Binomial Random variable X and compute its probability mass function (PMF) or cumulative density function (CDF). We love the scipy stats library because it defines all the functions you would care about for a random variable, including expectation, variance, and even things we haven't talked about in CS109, like entropy. This example declares $X \sim \text{Bin}(n = 10, p = 0.2)$. It calculates a few statistics on X. It then calculates $P(X = 3)$ and $P(X \leq 4)$. Finally it generates a few random samples from X:

```
from scipy import stats
X = stats.binom(10, 0.2) # Declare X to be a binomial random variable
print(X.pmf(3))          # P(X = 3)
print(X.cdf(4))          # P(X <= 4)
print(X.mean())          # E[X]
print(X.var())           # Var(X)
print(X.std())           # Std(X)
print(X.rvs())           # Get a random sample from X
print(X.rvs(10))         # Get 10 random samples form X
```

From a **terminal** you can always use the "help" command to see a full list of methods defined on a variable (or for a package):

```
from scipy import stats
X = stats.binom(10, 0.2) # Declare X to be a binomial random variable
help(X)                  # List all methods defined for X
```

Poisson

Make a Poisson Random variable Y. This example declares $Y \sim \text{Poi}(\lambda = 2)$. It then calculates $P(Y = 3)$:

```
from scipy import stats
Y = stats.poisson(2)  # Declare Y to be a poisson random variable
print(Y.pmf(3))       # P(Y = 3)
print(Y.rvs())        # Get a random sample from Y
```

Geometric

Make a Geometric Random variable X, the number of trials until a success. This example declares $X \sim \text{Geo}(p = 0.75)$:

```
from scipy import stats
X = stats.geom(0.75)  # Declare X to be a geometric random variable
print(X.pmf(3))       # P(X = 3)
print(X.rvs())        # Get a random sample from Y
```

Normal

Make a Normal Random variable A. This example declares $A \sim N(\mu = 3, \sigma^2 = 16)$. It then calculates $f_Y(0)$ and $F_Y(0)$. **Very Important!!!** In class, the second parameter to a normal was the variance (σ^2). In the scipy library, the second parameter is the standard deviation (σ):

```
import math
from scipy import stats
A = stats.norm(3, math.sqrt(16)) # Declare A to be a normal random variable
print(A.pdf(4))                  # f(3), the probability density at 3
print(A.cdf(2))                  # F(2), which is also P(Y < 2)
print(A.rvs())                   # Get a random sample from A
```

Exponential

Make an Exponential Random variable B. This example declares $B \sim \text{Exp}(\lambda = 4)$:

```
from scipy import stats
# `λ` is a common parameterization for the exponential,
#  but `scipy` uses `scale` which is `1/λ`
B = stats.expon(scale=1/4)
print(B.pdf(1))             # f(1), the probability density at 1
print(B.cdf(2))             # F(2) which is also P(B < 2)
print(B.rvs())              # Get a random sample from B
```

Beta

Make an Beta Random variable X. This example declares $X \sim \text{Beta}(\alpha = 1, \beta = 3)$:

```
from scipy import stats
X = stats.beta(1, 3)  # Declare X to be a beta random variable
print(X.pdf(0.5))     # f(0.5), the probability density at 1
print(X.cdf(0.7))     # F(0.7) which is also P(X < 0.7)
print(X.rvs())        # Get a random sample from X
```

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.

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 step rule of counting. An image can be created one pixel at a time, step by step. Each time we choose 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 other words: 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 logarithmic 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 Step 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)

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 Inclusion Exclusion rule.

Definition: Mutually Exclusive Counting

If the outcome of an experiment can either be drawn from set A or set B, where none of the outcomes in set A are the same as 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 route 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 and $B = \emptyset$, the Inclusion-Exclusion Principle gives the same result as the Sum Rule of Counting since $|A \text{ and } B| = 0$.

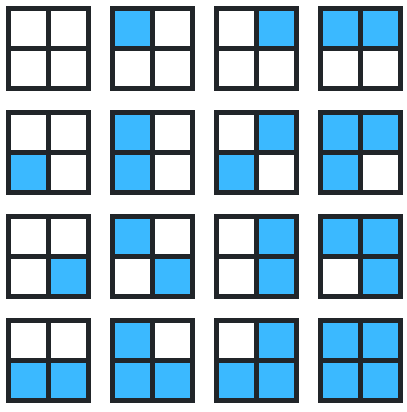
Example: An 8-bit string (one byte) is sent over a network. The valid set of strings recognized by the receiver must either start with "01" or end with "10". How many such strings are there?

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.

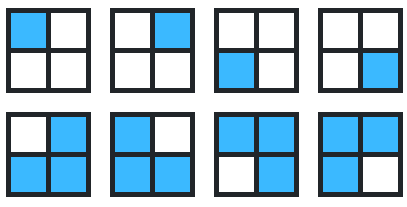
Overcounting and Correcting

One strategy for counting is sometimes to overcount a solution and then 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.

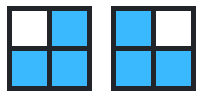
As a simple example to demonstrate the point, lets revisit the problem of generating all images, but this time lets just have 4 pixels (2x2) and each pixel can only be **blue** or white. How many unique images are there? Generating any image is a four step process where you choose each pixel one at a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



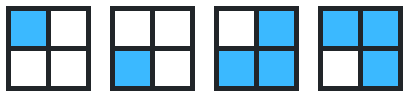
Now lets say we add in new "constraint" that we only want to accept pictures which have an odd number of pixels turned blue. There are two ways of getting to the answer. You could start out with the original 16 and work out that you need to subtract off 8 images that have either 0, 2 or 4 blue pixels (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 (they 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 examples in the set of 8 images with an 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.

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.

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.

$$\begin{array}{ccccc} 0_1 & 1_0 & 1_1 & 0_2 & 0_3 \\ 0_1 & 1_0 & 1_1 & 0_3 & 0_2 \\ 0_2 & 1_0 & 1_1 & 0_1 & 0_3 \\ 0_2 & 1_0 & 1_1 & 0_3 & 0_1 \\ 0_3 & 1_0 & 1_1 & 0_1 & 0_2 \\ 0_3 & 1_0 & 1_1 & 0_2 & 0_1 \end{array}$$

$$\begin{array}{ccccc} 0_1 & 1_1 & 1_0 & 0_2 & 0_3 \\ 0_1 & 1_1 & 1_0 & 0_3 & 0_2 \\ 0_2 & 1_1 & 1_0 & 0_1 & 0_3 \\ 0_2 & 1_1 & 1_0 & 0_3 & 0_1 \\ 0_3 & 1_1 & 1_0 & 0_1 & 0_2 \\ 0_3 & 1_1 & 1_0 & 0_2 & 0_1 \end{array}$$

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"?

Solution: 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 are:

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

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 code 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)}
```

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 from 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.
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. books: There are $\binom{4}{2}$ ways of doing so.
Case 2: Select the 9th Ed. and 2 other non-8th Ed. books: There are $\binom{4}{2}$ ways of doing so.
Case 3: Select 3 books from the 4 remaining books that are neither the 8th nor the 9th 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!

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}.

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 the same (your elements) and $r-1$ of which are the same (the dividers). Thus the total number of outcomes is:

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

The divider method can be derived via the "Stars and Bars" method. This is a creative construction where we consider permutations of indistinguishable items, represented by stars `*`, and dividers between our containers, represented by bars `|`. Any distinct permutation of these stars and bars represents a unique assignments of our items to containers.

Imagine we want to separate 5 indistinguishable objects into 3 containers. We can think of the problem as finding the number of ways to order 5 stars and 2 bars `*****||`. Any permutation of these symbols represents a unique assignment. Here are a few examples:

`**|*|**` represents 2 items in the first bucket, 1 item in the second and 2 items in the third.

`****||*` represents 4 items in the first bucket, 0 item in the second and 1 items in the third.

`||*****` represents 0 items in the first bucket, 0 item in the second and 5 items in the third.

Why are there only 2 dividers when there are 3 buckets? This is an example of a fence-post-problem". With 2 dividers you have created three containers. We already have a method for counting permutations with some indistinct items. For the example above where we have seven elements in our permutation ($n = 5$ stars and $r - 1 = 2$ bars):

$$\text{Number of unique orderings} = \frac{n!}{n_1!n_2!} = \frac{(n+r-1)!}{n!(r-1)!} = \frac{7!}{5!2!} = 21$$

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 $i = 1, 2, 3, 4, 5$.

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 deep 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%"?

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 the 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}
- At least 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 (≥ 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.

Definition of Probability

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

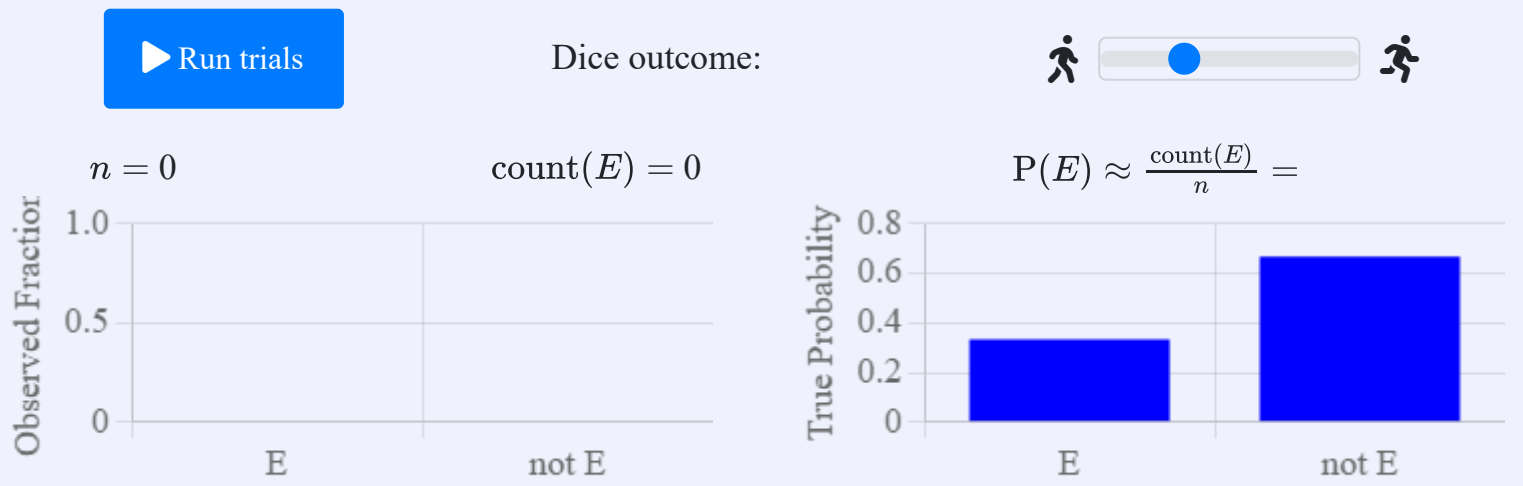
$$\mathrm{P}(\text{Event}) = \lim_{n \to \infty} \frac{\text{count}(\text{Event})}{n}$$

In English this reads: lets say you perform n trials of an "experiment" which could result in a particular "Event" occurring. The probability of the event occurring, $\mathrm{P}(\text{Event})$, 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 converge 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 our belief that an event will happen given our limited knowledge. This interpretation acknowledges 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 total number of trials — will tend to our desired probability:

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

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 guarantee 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.

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 in fact apply to probability. From the foundation provided by this set of axioms mathematicians built the edifice of probability theory.

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* occurring which we write in probability notation as E^{C}, if you know the probability of it occurring -- 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 $\mathrm{P}(E) = \frac{0.4}{1} = 0.4$.

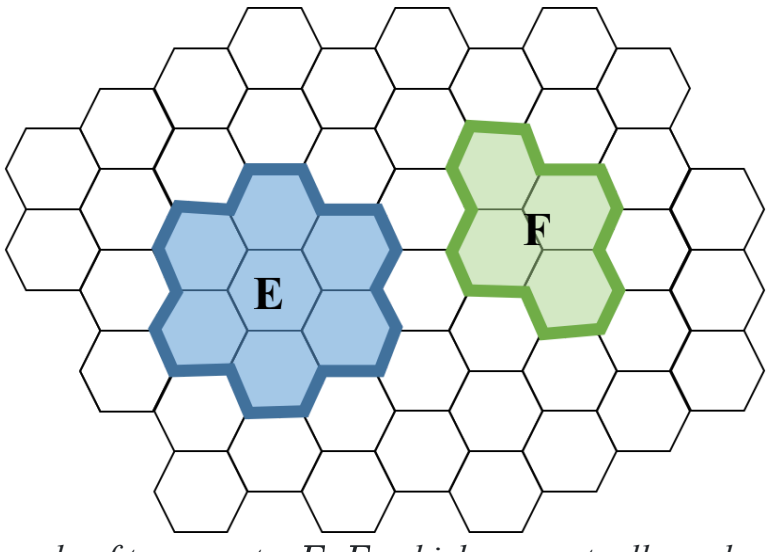
Probability of **or**

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.

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.

Or with 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 occurring 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?

Or with 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 $\mathrm{P}(E) + \mathrm{P}(F)$.

Inclusion-Exclusion with Three Events

What does the inclusion exclusion property look like if we have three events, that are not mutually exclusive, and we want to know the probability of or, $\mathrm{P}(E_1 \text{ or } \mathrm{E}_2 \text{ or } E_3)$?

Recall that if they are mutually exclusive, we simply add the probabilities. If they are not mutually exclusive, you need to use the inclusion exclusion formula for three events:

$$
\begin{aligned}
\mathrm{P}(E_1 \text{or } \mathrm{E}_2 \text{ or } E_3) = \\
&+ \mathrm{P}(E_1) \\
&+ \mathrm{P}(E_2) \\
&+ \mathrm{P}(E_3) \\
&- \mathrm{P}(E_1 \text{ and } E_2) \\
&- \mathrm{P}(E_1 \text{ and } E_3) \\
&- \mathrm{P}(E_2 \text{ and } E_3) \\
&+ \mathrm{P}(E_1 \text{ and } E_2 \text{ and } E_3)
\end{aligned}
$$

In words, to get the probability of three events, you: (1) add the probability of the events on their own. (2) Then you need to subtract off the probability of every *pair* of events co-occuring. (3) Finally, you add in the probability of all three events co-occuring.

Inclusion-Exclusion with n Events

Before we explore the general formula, lets look at one more example. Inclusion-exclusion with four events:

$$
\begin{aligned}
\mathrm{P}(E_1 \text{ or } E_2 \text{ or } E_3 \text{ or } E_4) = \\
&+ \mathrm{P}(E_1) \\
&+ \mathrm{P}(E_2) \\
&+ \mathrm{P}(E_3) \\
&+ \mathrm{P}(E_4) \\
&- \mathrm{P}(E_1 \text{ and } E_2) \\
&- \mathrm{P}(E_1 \text{ and } E_3) \\
&- \mathrm{P}(E_1 \text{ and } E_4) \\
&- \mathrm{P}(E_2 \text{ and } E_3) \\
&- \mathrm{P}(E_2 \text{ and } E_4) \\
&- \mathrm{P}(E_3 \text{ and } E_4) \\
&+ \mathrm{P}(E_1 \text{ and } E_2 \text{ and } E_3) \\
&+ \mathrm{P}(E_1 \text{ and } E_2 \text{ and } E_4) \\
&+ \mathrm{P}(E_1 \text{ and } E_3 \text{ and } E_4) \\
&+ \mathrm{P}(E_2 \text{ and } E_3 \text{ and } E_4) \\
&- \mathrm{P}(E_1 \text{ and } E_2 \text{ and } E_3 \text{ and } E_4)
\end{aligned}
$$

Do you see the pattern? For every possible subset of events from our n events, we calculate the probability of the "and" of the subset. If the subset has an odd number of events, we add its probability mass, otherwise we subtract the probability. This can be written up mathematically — but it is a rather hard pattern to express in notation:

$$
\begin{aligned}
\mathrm{P}(E_1 \text{ or } E_2 \text{ or } \cdots \text{ or } E_n) &= \sum_{r-1}^{n} (-1)^{r+1} Y_r \\
\text{s.t. } Y_r &= \sum_{1 \leq i_1 < \cdots < i_r \leq n} \mathrm{P}(E_{i_1} \text{ and } \cdots \text{ and } E_{i_r})
\end{aligned}
$$

The notation for Y_r is especially hard to parse. Y_r sums over all ways of selecting a subset of r events. For each selection of r events, calculate the probability of the intersection of those events. $(-1)^{r+1}$ is saying: alternate between addition and subtraction, starting with addition.

It is not especially important to follow the math notation here. The main take away is that the general inclusion exclusion principle gets incredibly complex with multiple events. Often, the way to make progress in this situation is to find a way to solve your problem using another method.

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

Conditional Probability

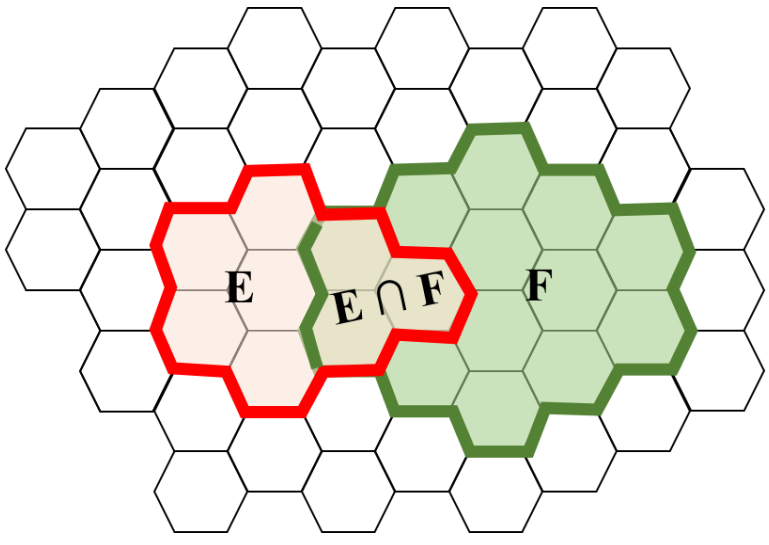
In English, a conditional probability states "what is the chance of an event E happening given that I have already observed some other event F". It is a critical idea in machine learning and probability because it allows us to update our probabilities in the face of new evidence.

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!

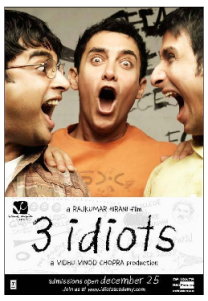
Conditional Probability Example

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 the 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:

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



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

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

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

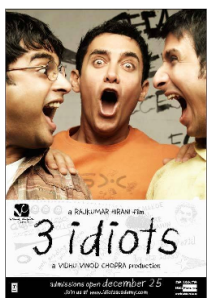
$\mathrm{P}(E) = 0.03$

Now for a more interesting question. What is the 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!

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	$\mathrm{P}(E^{\mathrm{C}}) = 1 - \mathrm{P}(E)$	$\mathrm{P}(E^{\mathrm{C}}\|G) = 1 - \mathrm{P}(E\|G)$

Conditioning on Multiple Events

The conditional paradigm also applies to the definition of conditional probability! Again if we consistently condition on some event G occurring, 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 occurring, given that both F and G have occurred". This equation states that the definition for conditional probability of $E|F$ still applies in the universe where G has occurred. 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.

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

Alternative Definition

Another definition of independence can be derived by using an equation called the chain rule, which we will learn about later, in the context where two events are independent. Consider two indepedent events A and B:

$$\begin{aligned} \mathrm{P}(A, B) &= \mathrm{P}(A) \cdot \mathrm{P}(B|A) && \text{Chain Rule} \\ &= \mathrm{P}(A) \cdot \mathrm{P}(B) && \text{Independence} \end{aligned}$$

Independence is Symmetric

This definition is symmetric. If E is independent of F, then F is independent of E. We can prove that $\mathrm{P}(F|E) = \mathrm{P}(F)$ implies $\mathrm{P}(E|F) = \mathrm{P}(E)$ starting with a law called Bayes' Theorem which we will cover shortly:

$$\begin{aligned} \mathrm{P}(E|F) &= \frac{\mathrm{P}(F|E) \cdot \mathrm{P}(E)}{\mathrm{P}(F)} && \text{Bayes Theorem} \\ &= \frac{\mathrm{P}(F) \cdot \mathrm{P}(E)}{\mathrm{P}(F)} && \mathrm{P}(F|E) = \mathrm{P}(F) \\ &= \mathrm{P}(E) && \text{Cancel} \end{aligned}$$

Generalized Independence

Events $E_1, E_2, \ldots, E_n$ are independent if for every subset with r elements (where $r \leq n$):

$$\mathrm{P}(E_{1'}, E_{2'}, \ldots, E_{r'}) = \prod_{i=1}^{r} \mathrm{P}(E_i')$$

As an example, consider the probability of getting 5 heads on 5 coin flips where we assume that each coin flip is independent of one another.

Let H_i be the event that the ith coin flip is a heads:

$$
\begin{aligned}
&\mathrm{P}(H_1, H_2, H_3, H_4, H_5) \\
&\quad = \mathrm{P}(H_1) \cdot \mathrm{P}(H_2) \cdots \mathrm{P}(H_5) && \text{Independence} \\
&\quad = \prod_{i=1}^{5} \mathrm{P}(H_i) && \text{Product notation} \\
&\quad = \prod_{i=1}^{5} \frac{1}{2} \\
&\quad = \frac{1}{2}^5 \\
&\quad = 0.03125
\end{aligned}
$$

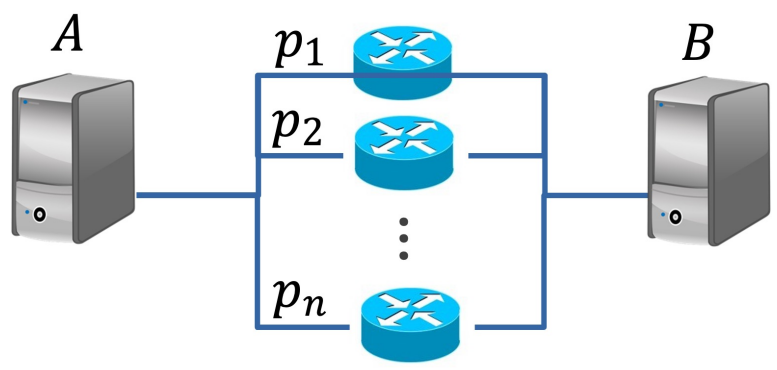
How to Establish Independence

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 independent. 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 discrepancy 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 negligible). Let's work through an example to better understand.

Example: Parallel Networks

Over networks, such as the internet, computers can send information. Often there are multiple paths (mediated by routers) between two computers and as long as one path is functional, information can be sent. Consider the following parallel network with n ***independent*** routers, each with probability p_i of functioning (where $1 \leq i \leq n$). Let E be the event that there is a functional path from A to B. What is $\mathrm{P}(E)$?



A simple network that connects two computers, A and B.

Let F_i be the event that router i fails. Note that the problem states that routers are indepenent, and as such we assume that the events F_i are all independent of one another.

$$
\begin{aligned}
\mathrm{P}(E) &= \mathrm{P}(\text{At least one router works}) \\
&= 1 - \mathrm{P}(\text{All routers fail}) \\
&= 1 - \mathrm{P}(F_1 \text{ and } F_2 \text{ and } \ldots \text{ and } F_n) \\
&= 1 - \prod_{i=1}^{n} \mathrm{P}(F_i) && \text{Independence of } F_i \\
&= 1 - \prod_{i=1}^{n} 1 - p_i
\end{aligned}
$$

Where p_i is the probability that router i is functional.

Independence and Compliments

Given independent events A and B, we can prove that A and B^C are independent. Formally we want to show that: $\mathrm{P}(AB^C) = \mathrm{P}(A)\,\mathrm{P}(B^C)$. This starts with a rule called the Law of Total Probability which we will cover shortly.

$$\begin{aligned}
\mathrm{P}(AB^C) &= \mathrm{P}(A) - \mathrm{P}(AB) && \text{LOTP} \\
&= \mathrm{P}(A) - \mathrm{P}(A)\,\mathrm{P}(B) && \text{Independence} \\
&= \mathrm{P}(A)[1 - \mathrm{P}(B)] && \text{Algebra} \\
&= \mathrm{P}(A)\,\mathrm{P}(B^C) && \text{Identity 1}
\end{aligned}$$

Conditional Independence

We saw earlier that the laws of probability still held if you consistently conditioned on an event. As such, the definition of independence also transfers to the universe of conditioned events. We use the terminology "conditional independence" to refer to events that are independent when consistently conditioned. For example if someone claims that events E_1, E_2, E_3 are conditionally independent given event F. This implies that

$$\mathrm{P}(E_1, E_2, E_3 | F) = P(E_1|F) \cdot P(E_2|F) \cdot P(E_3|F)$$

Which can be written more succinctly in product notation

$$\mathrm{P}(E_1, E_2, E_3 | F) = \prod_{i=1}^{3} P(E_i|F)$$

Warning: While the rules of probability stay the same when conditioning on an event, the independence *property* between events might change. Events that were dependent can become independent when conditioning on an event. Events that were independent can become dependent. For example, if events E_1, E_2, E_3 are conditionally independent given event F it is not necessarily true that

$$\mathrm{P}(E_1, E_2, E_3) = \prod_{i=1}^{3} P(E_i)$$

As we are no longer conditioning on F.

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 probability of the ***and*** of events.

And with 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 occurring 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

And with Dependent Events

Events which are not independent are called ***dependent*** events. How can you calculate the probability of 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 later 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:

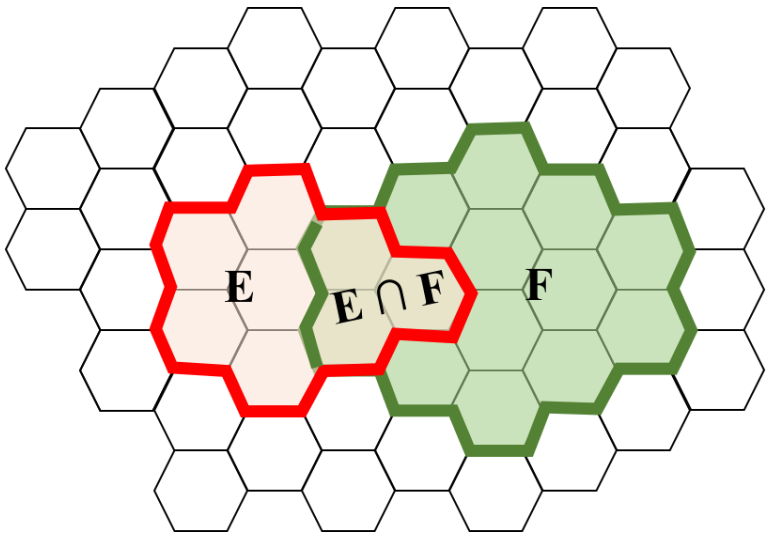
$$\mathrm{P}(E \text{ and } F) = \mathrm{P}(F \text{ and } E) = \mathrm{P}(F|E) \cdot \mathrm{P}(E)$$

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 saw for conditional probability:



that event E can be thought of as having two parts, the part that is in F, (E and F), and the part that isn't, (E and F^{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.

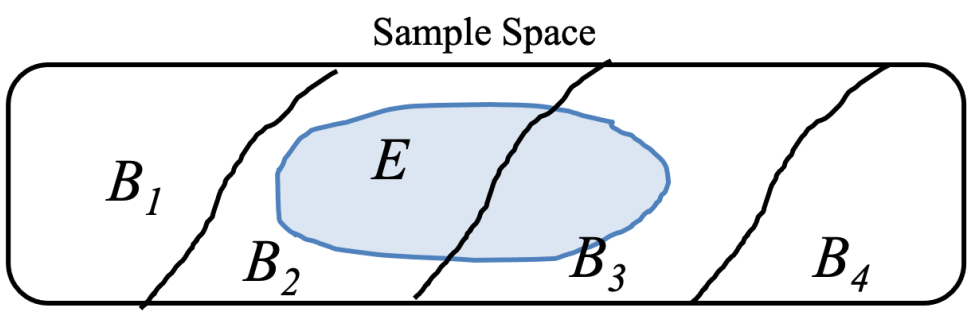
The Law of Total Probability
If we combine our above observation with the chain rule, we get a very useful formula:

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

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 the sample space falls 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 $\mathrm{P}(E)$ to be equal to $\mathrm{P}\left[(E \,\text{and}\, B_1) \text{ or } (E \,\text{and}\, B_2) \ldots\right]$. 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 expand 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 it 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 known. 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.

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 the illness 0.92

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

The numbers in this example are from the Mammogram test for breast cancer. The seriousness of cancer underscores the potential for bayesian probability to be applied to important contexts. The natural occurrence of breast cancer is 8%. The mammogram 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 recalculate.

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:

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

Natural Frequency Intuition

One way to build intuition for Bayes Theorem is to think about "natural frequences". Let's take another approach to 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 plausible 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).

Bayes with the General Law of Total Probability

A classic challenge when applying Bayes' theorem 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:

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

Unknown Normalization Constant

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 solution that we used above was the law of total probability: $\mathrm{P}(E) = \mathrm{P}(E|B)\,\mathrm{P}(B) + \mathrm{P}(E|B^{\mathrm{C}})\,\mathrm{P}(B^{\mathrm{C}})$. This allows us to calculate $\mathrm{P}(E)$.

Here is another strategy for dealing with an unknown $\mathrm{P}(E)$. We can make it 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:

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

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.

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

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:

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

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:

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

Total number of heads: 8

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

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: $p^k \cdot (1-p)^{n-k}$

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, $\mathrm{P}(E_i) = p^k \cdot (1-p)^{n-k}$. Now we are ready to calculate the probability of exactly k heads:

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

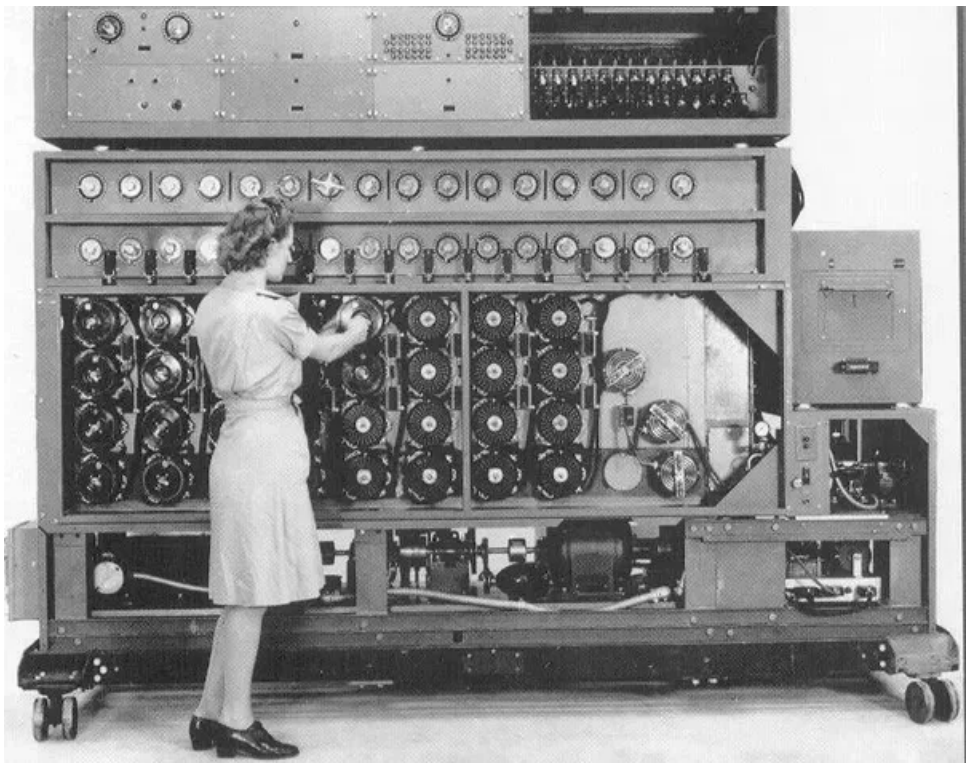
More than k heads

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 Nazis would choose a new configuration and if the 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?

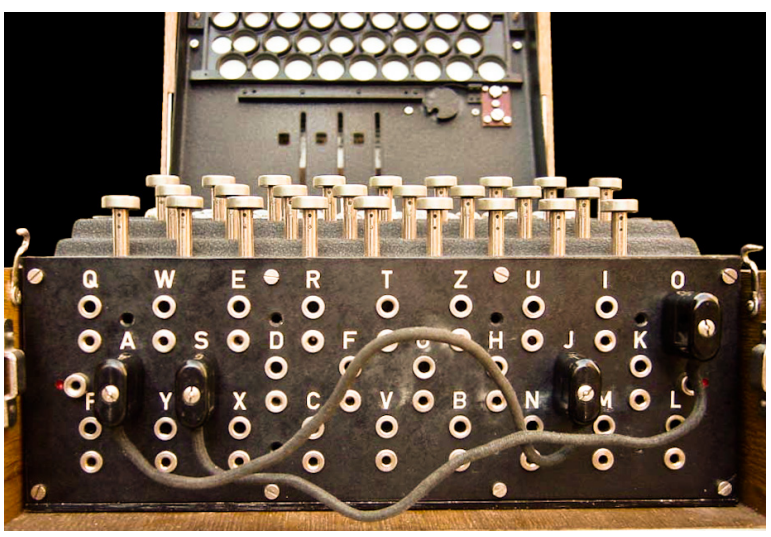


The WW2 machine built to search different enigma configurations.

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 electrical 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? It's 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 choose 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 rotors, 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 went on to build a machine which could help test many configurations -- a predecessor 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:

Random Shuffles

Here is a suprising claim. If you shuffle a standard deck of cards seven times, with almost total certainty you can claim that the exact ordering of cards has never been seen! Wow! Let's explore. We can ask this question formally as: What is the probability that in the n shuffles seen since the start of time, yours is unique?

Orderings of 52 Cards

Our adventure starts with a simple observation: there are **very** many ways to order 52 cards. But exactly how many unique orderings of a standard deck are there?

There are 52! ways to order a standard deck of cards. Since each card is unique (each card has a unique suit, value combination) then we can apply the rule for Permutations of Distinct Objects:

$$\text{Num. Unique Orderings} = 52!$$

That is a humongous number. 52! equals
80658175170943878571660636856403766975289505440883277824000000000000.
That is over $8 \cdot 10^{67}$. Recall it is estimated that there are around 10^{82} atoms in the observable universe

Number of Shuffles Ever Seen

Of course we don't know what the value of n is — nobody has been counting how many times humans have shuffled cards. We can come up with a reasonable overestimate. Assume k = 7 billion people have been shuffling cards once a second since cards were invented. Playing cards may have been invented as far back as the Tang dynasty in the 9th century. To the best of my knowledge the oldest set of 52 cards is the Topkapı deck of cards in Istanbul around the 15th century ad. That is about s = 16,472,828,422 seconds ago. As such our overestimate is $n = s \cdot k \approx 10^{20}$.

Next let's calculate the probability that none of those n historical shuffles matches your particular ordering of 52 cards. There are two valid approaches: using equally likely outcomes, and using independence.

Equally Likely Outcomes

One way to the probability that your ordering of 52 cards is unique in history is to use Equally Likely Outcomes. Consider the sample space of all the possible ordering of all the cards ever dealt. Each outcome in this set will have n card decks each with their own ordering. As such the size of the sample space is $|S| = (52!)^n$. Note that all outcomes in the sample space are equally likely — we can convince ourselves of this by symmetry — no ordering is more likely than any other. Out of that sample space we want to count the number of outcomes where none of the orderings matches yours. There are $52! - 1$ ways to order 52 cards that are not yours. We can construct the event space by steps: for each of the n shuffles in history select any one of those $52! - 1$ orderings. Thus $|E| = (52! - 1)^n$.

Let U be the event that your particular ordering of 52 cards is unique

$$
\begin{aligned}
\mathrm{P}(U) &= \frac{|E|}{|S|} && \text{Equally Likely Outcomes} \\
&= \frac{(52!-1)^n}{(52!)^n} \\
&= \frac{(52!-1)^{10^{20}}}{(52!)^{10^{20}}} && n = 10^{20} \\
&= \left(\frac{52!-1}{52!}\right)^{10^{20}}
\end{aligned}
$$

In theory that is the correct answer, but those numbers are so big, its not clear how to evaluate it, even when using a computer. One good idea is to first compute the log probability:

$$
\begin{aligned}
\log \mathrm{P}(U) &= \log\left[\left(\frac{52!-1}{52!}\right)^{10^{20}}\right] \\
&= 10^{20} \cdot \log\left(\frac{52!-1}{52!}\right) \\
&= 10^{20} \cdot \Big[\log(52!-1) - \log(52!)\Big] \\
&= 10^{20} \cdot (-1.24 \times 10^{-68}) \\
&= -1.24 \times 10^{-48}
\end{aligned}
$$

Now if we undo the log (and use the fact that e^{-x} is very close to $1-x$ for small values of x):

$$
\begin{aligned}
\mathrm{P}(U) &= e^{-1.24 \times 10^{-48}} \\
&\approx 1 - 1.24 \times 10^{-48}
\end{aligned}
$$

So the probability that your particular ordering is unique is very close to 1, and the probability that someone else got the same ordering, $1 - P(U)$, is a number with 47 zeros after the decimal point. It is safe to say your ordering is unique.

In python, you can use a special library called decimal to compute very small probabilities. Here is an example of how to compute $\log \frac{52!-1}{52!}$:

```python
from decimal import *
import math

n = math.pow(10, 20)
card_perms = math.factorial(52)
denominator = card_perms
numerator = card_perms - 1

# decimal library because these are tiny numbers
getcontext().prec = 100 # increase precision
log_numer = Decimal(numerator).ln()
log_denom = Decimal(denominator).ln()
log_pr = log_numer - log_denom

# approximately -1.24E-68
print(log_pr)
```

We can also check our result using the binomial approximation.

For small values of x, the value $(1-x)^n$ is very close to $1 - nx$, and this gives us another way to compute $P(U)$:

$$
\begin{aligned}
\mathrm{P}(U) &= \frac{(52!-1)^n}{(52!)^n} \\
&= \left(1 - \frac{1}{52!}\right)^{10^{20}} && n = 10^{20} \\
&\approx 1 - \frac{10^{20}}{52!} \\
&\approx 1 - 1.24 \times 10^{-48}
\end{aligned}
$$

This agrees with the result we got using python's decimal library.

Independence

Another approach is to define events D_i that the ith card shuffle is different than yours. Because we assume each shuffle is independent, then $P(U) = \prod_i P(D_i)$. What is the probability of (D_i)? If you think of the sample space of D_i, it is 52! ways of ordering a deck cards. The event space is the 52! - 1 outcomes which are not your ordering.

$$
\begin{aligned}
\mathrm{P}(U) &= \prod_{i=1}^{n} \mathrm{P}(D_i) \\
\log \mathrm{P}(U) &= \sum_{i=1}^{n} \log \mathrm{P}(D_i) \\
&= n \cdot \log \mathrm{P}(D_i) \\
&= 10^{20} \cdot \log \frac{52! - 1}{52!} \\
&\approx 10^{20} \cdot -1.24 \times 10^{-68}
\end{aligned}
$$

Which is the same answer we got with the other approach for $\log \mathrm{P}(U)$

How Random is your Shuffle?

A final question we can look into. How do you get a truly random ordering of cards? Dave Bayer and Persi Diaconis in 1992 worked through this problem and published their results in the article Trailing the Dovetail Shuffle to its Lair. They showed that if you shuffle a deck of cards seven times using a riffle shuffle also known as the dovetail shuffle, you are almost garunteed a random ordering of cards. The methodology used paved the way for studying psuedo random numbers produced by computers.

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 mutation. By the 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 giving 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!