Basic Mathematical Statistics for 391L

Eric Ingram

2024-10-08

erimingram.github.io

Mathematical Statistics for CS/DSC/AI 391L:

This document is a summary of what I have said in my office hours in respect to the mathematical statistics that you are expected to know for this course. Please watch the office hours recording for more detail.

This document assumes that you have a basic understanding of probability, derivatives/integrals, and mathematical notation. It will try to explain things in a simplified format: forgive the details left out.

Basics of Random Variables

What is a Random Variable?

First, let's examine the concept of a **random variable**. A random variable, commonly denoted by a capital letter such as X represents a numerical value that varies based on the outcome of a random event.

So for example, let's say that X = the number rolled on a fair six-sided dice.

When you roll a fair six-sided dice, you can roll a 1, 2, 3, 4, 5, or 6, each with equal probability (why it is called "fair") of $\frac{1}{6}$. Mathematically, we commonly denote this as $P(X = x) = \frac{1}{6}$: the probability of random variable X being some number is $\frac{1}{6}$. x is just a specific value that the random variable X can take (so 1, 2, 3, 4, 5, or 6 in this scenario).

So when we say that X = the number rolled on a fair six-sided dice, we are saying that X is going to be one of these values based on whatever the roll ends up as but we don't know what value yet: it is up to chance.

Here, since all the probabilities are the same, we can write $P(X = x) = \frac{1}{6}$ since $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = \frac{1}{6}$. This is usually not the case.

Discrete vs. Continuous Random Variables

The dice example is an instance where our random variable is **discrete**. A discrete random variable is one where the possible values of the random variable come from a finite or infinitely countable set.

So, the set of possible values in the dice example was $\{1, 2, 3, 4, 5, 6\}$: a finite set. There are only 6 possible values.

An infinitely countable set is a set that has an infinite number of things but we can count them in order in some way. So, for example, the set of whole numbers ($\{0, 1, 2, 3, \dots\}$) is countably infinite since we can keep counting them in some order even though the set never ends.

In the discrete case, it makes no sense to talk about the probability of values not in the set. So, for example, $P(X = 3.5) = 0$ since X can never equal 3.5 since 3.5 is not in our finite set (it is not a possible dice roll on a fair, six-sided dice).

In contrast, we can look at **continuous** random variables. A continuous random variable can take on any value within an interval of possible values.

So for example, let's look at the example of height. So, X = the height of an adult human.

Looking at historical records, the shortest person ever verified was Nepal's Chandra Dangi at 21.5 inches (54.6 cm) and the tallest person ever verified was the U.S.' Robert Wadlow at 107.1 inches (272.0 cm). So, we can say our interval of possible values is $[21.5, 107.1]$ (this is set notation, so we are saying that our random variable X can be any number between (and including) 21.5 and 107.1).

Since there are infinitely many values between these intervals, it makes no sense to talk about the probability of an exact number since it would be infinitely small. We commonly say $P(X = x) = 0$. Instead, to work with continuous random variables, we talk about probabilities over an interval: more on this later.

Do note that we often see **discretization** (the process of grouping continuous data into finite intervals) with height. Height is a continuous random variable but in practice it is often measured and recorded in discrete values such as "6 feet and 3 inches" or "2.1 meters" when the true height is actually off by some decimals. This is done to simplify data collection and analysis but you end up losing the precision that continuous measurements give.

Expected Value and Variance

We are often interested in understanding certain properties of random variables. For example, what the "average value" of a random variable is or how the values of a random variable tend to "spread out" from the average value.

We define the **Expected Value** of a random variable to be the long-term average outcome we'd get if we were to repeat the random event over and over.

For a discrete random variable X, we define $E(X) = \sum_x x * P(X = x)$. This is functionally a weighted sum of the possible outcomes by their probabilities.

For the dice example, $E(X) = 1 * \frac{1}{6} + 2 * \frac{1}{6} + 3 * \frac{1}{6} + 4 * \frac{1}{6} + 5 * \frac{1}{6} + 6 * \frac{1}{6} = 3.5$.

So, over many dice rolls, we would expect the average dice roll to come out to 3.5. An important clarifying note: if we roll a dice once, it will never come out to 3.5 since it is impossible to roll a 3.5. But if we rolled the dice thousands of times, for example, we'd expect the average value rolled to be 3.5.

In the continuous case, summing up probabilities makes no sense since the probability at a point is infinitely small Let's make a simple connection to calculus here: if the expected value for discrete random variables is like a Riemann sum (where we add up discrete values), what will the expected value for continuous random variables be like?

If you said an integral, you're completely right! For a continuous random variable X, the expected value is found by integrating over all possible values of X, multiplying each value by its contribution to the overall probability. This is written as $E(X) = \int_{-\infty}^{\infty} x * (\text{a function that describes probability}) dx$. We will get into the function that goes here in the next section.

If you know the specific range of possible values for X you can substitute those limits in the integral (e.g., from a to b), but outside of this range, the integral outside that range contributes nothing to the expected value so it is the same thing if you keep the infinities in the integral. The expected value will still come out to be a single number overall with the same interpretation as in the discrete case.

A related statistic is the **variance** of a random variable X, denoted $Var(X)$. Variance tells us how much a random variable's values are spread out from the average (expected value!). If the values are close to the average, the variance is small. If the values are far from the average, the variance is large.

Mathematically, we say $Var(X) = E((X - E(X))^2) = E(X^2) - E(X)^2$. Intuitively, this is finding the average squared distance from the expected value.

A lot of times, it is more useful to use the second form of the variance from above since you will already have $E(X)$ and will only have to calculate $E(X^2)$.

When we have two random variables, X and Y, we need to talk about yet another statistic: the **covariance**, denoted $Cov(X, Y)$. Covariance measures how much two random variables change together. If both variables tend to increase or decrease simultaneously, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

Mathematically, we define covariance as $Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$. Once again, it is often more useful to use the second form from above.

Do note that we kept the form general here: for a continuous distribution, you will have to calculate the expected values with an integral and for a discrete distribution with a summation.

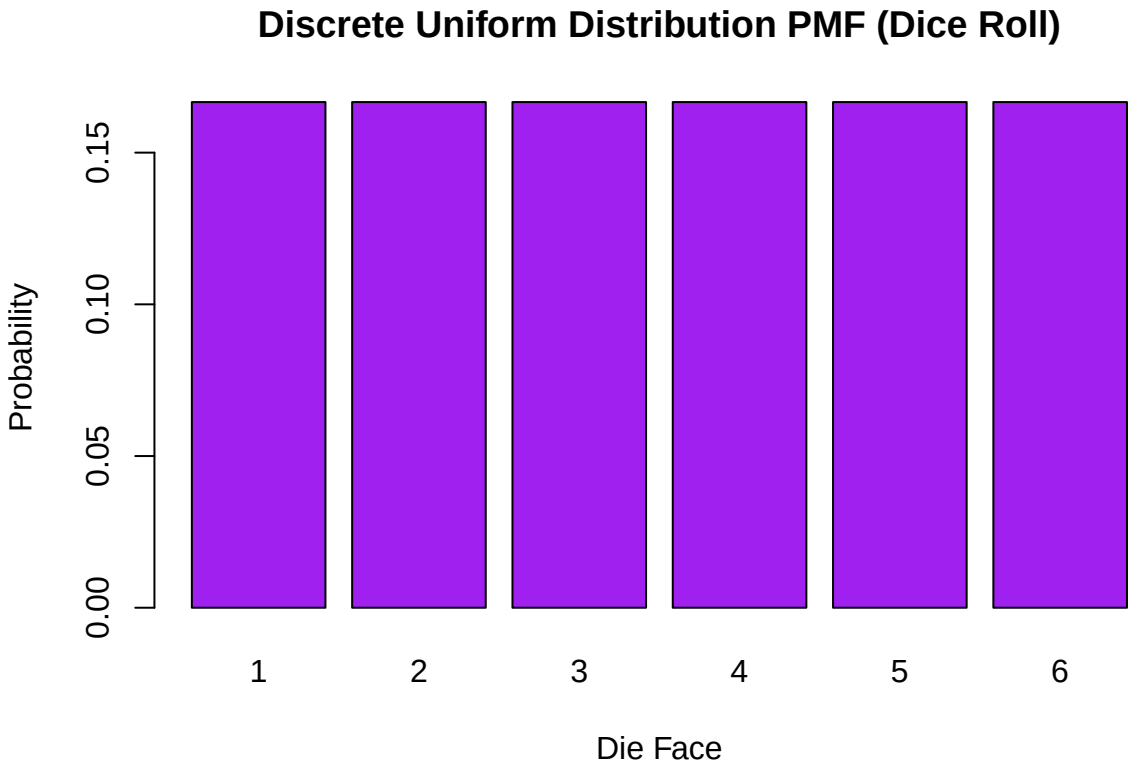
Probability Mass/Density Functions and Cumulative Distribution Functions

To be able to describe how probability is assigned to different values, we describe the probability with a function.

Discrete Case

For a discrete random variable X the **probability mass function**, often abbreviated as the PMF, defines the probability that a discrete random variable is exactly equal to some value.

We have already been doing this with our $P(X = x)$ values from earlier! These were just the values of the pmf at a single value. We can describe all of these with a function. For the dice example from earlier, this is what we would see:



This is an example of a Discrete Uniform Distribution (more on this later).

We can also view this in terms of counts if we wanted to. Imagine that we had a "perfect" experiment where we rolled the dice 12 times. We would have a graph that looked the exact same as above but instead of probability $\frac{1}{6}$, we would have counts of 2.

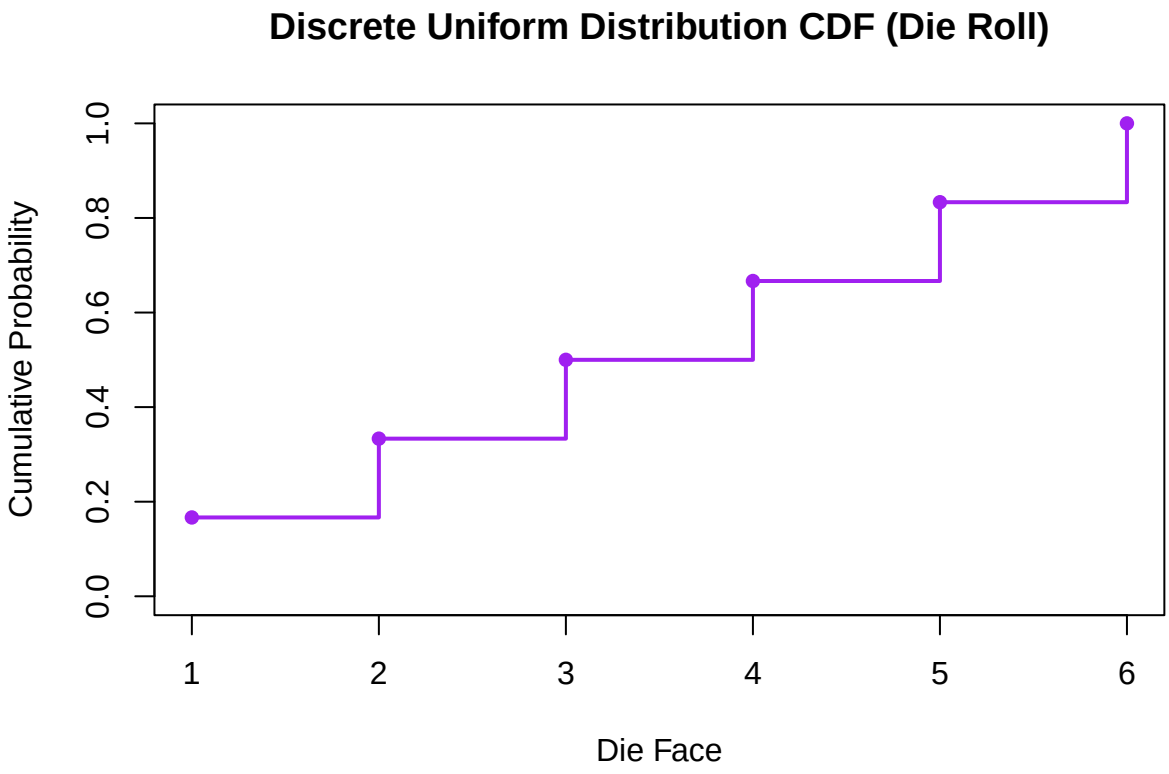
This function is perfect for answering questions like what we have seen earlier such as $P(X = 2)$: you just read it off of the graph/function!

However, for more complicated questions such as $P(X <= 3)$, it isn't enough to just plug in values. For questions such as this, we need the **Cumulative Distribution Function**, often abbreviated as the CDF.

Mathematically we say $F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x)$. So, we are just adding up all the individual probabilities less than or equal to the value that we care about since each probability is an individual, discrete value.

So, $F(3)$ is saying "what is the probability that we roll a value of 3 or less" for the dice example earlier. So, $F(3) = P(X \leq 3) = \sum_{x_i \leq 3} P(X = x_i) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = 1/2$.

Here is how this function would look graphically:



Why does this make sense? Think about it. If we were to ask "what is the probability we roll a 1.5 or less", we can only actually roll 1. So, the value of the CDF at 1.5 will still be $P(X = 1) = \frac{1}{6}$. Note that if we are interested in the opposite idea, $P(X > x)$, it is simple enough to recover with the complement rule: $P(X > x) = 1 - F(X)$.

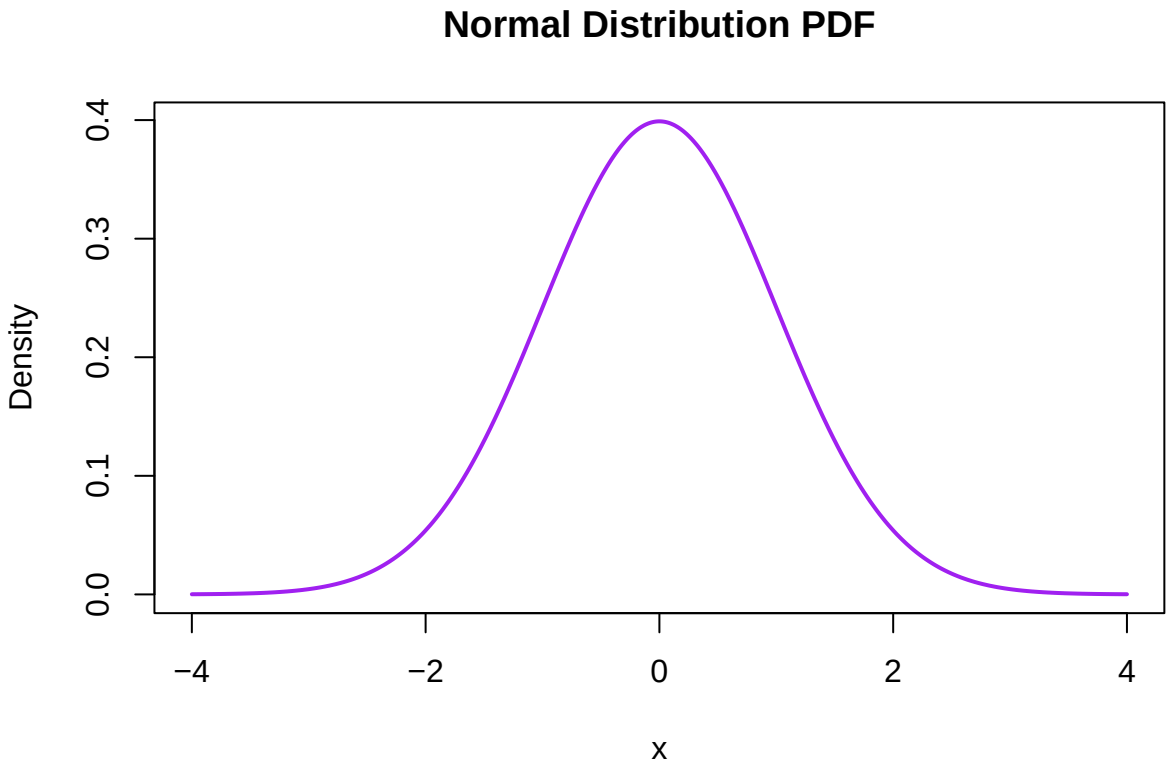
Continuous Case

Eariler, we hand-waved away how to work with continuous random variables. It is finally time to tell you!

As mentioned earlier, $P(X = x) = 0$ for continuous random variables. To talk about continuous random variables, we need to talk about them over an interval.

The **Probability Density Function**, $f(x)$ describes the likelihood of a continuous RV taking on a particular value. The probability that X falls within the interval $[a, b]$ is given by $P(a \leq X \leq b) = \int_a^b f(x)dx$.

Here is a visualization of what a Normal (more on this later) RV with parameters $\mu = 0, \sigma = 1$ looks like:

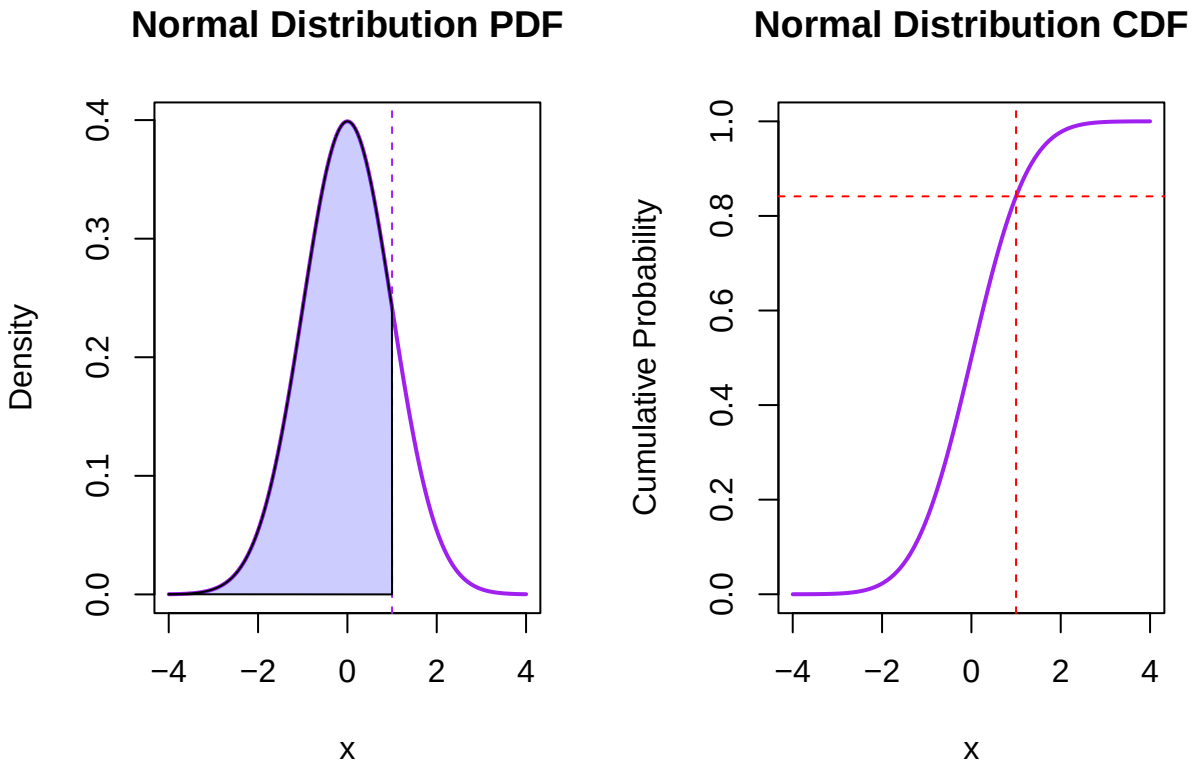


So, logically, when we are finding a probability, we are just looking at the area under the curve from our starting value to our stopping value!

The idea for the CDF for a continuous RV should be pretty clear then: integrate from as far left as you can go to the value you want to stop at.

Mathematically, $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$ where t is a dummy variable of integration.

For a visualization, please see below how the area under the PDF up to x corresponds to the value of the CDF at x.


Normal Distribution PDF
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x
Normal Distribution CDF
Cumulative Probability
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
x

Named Probability Distributions

Often, a distribution comes up often enough or has nice enough properties that we give it a name.

These distributions have **parameters**: numerical values that characterize the distribution and determine its shape and other properties. Parameters allow us to adjust the distribution to model different situations by changing aspects like the mean, variance, or other characteristics.

Here is a quick summary of the most common named probability distributions. Some of these will be useful for this course and some of these will not be.

Discrete Distributions

- Discrete Uniform
  - Interpretation: All events are equally likely. For a finite set of n outcomes, each outcome has probability $\frac{1}{n}$. If parametrized with $a = \min(x), b = \max(x)$, recover n with $n = b - a + 1$.
  - Expected Value: $E(X) = \frac{n+1}{2}$
  - Variance: $Var(X) = \frac{n^2-1}{12}$
  - Probability Mass Function: $P(X = x) = \frac{1}{n}, x = 1, 2, \ldots, n$
  - Cumulative Distribution Function: $F(x) = \frac{\lfloor x \rfloor}{n}, x \in [1, n]$
  - Example: When rolling a six-sided die, each number $(1 - 6)$ is equally likely.
- Bernoulli
  - Interpretation: A single trial with two possible outcomes: success (with probability p) or failure (with probability $1 - p$).
  - Expected Value: $E(X) = p$
  - Variance: $Var(X) = p(1 - p)$
  - Probability Mass Function: $P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$
  - Cumulative Distribution Function: $F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$
  - Example: Flipping a coin where heads is considered a success and tails is considered a failure.
- Binomial
  - Interpretation: Number of successes in n independent Bernoulli trials, each with success probability p.
  - Expected Value: $E(X) = np$
  - Variance: $Var(X) = np(1 - p)$
  - Probability Mass Function: $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \ldots, n$
  - Cumulative Distribution Function: $F(x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1 - p)^{n-k}$. Essentially, we are summing up each individual PMF value.
  - Example: Flipping a coin n times and counting the number of times we get heads.
- Geometric
  - Interpretation: The number of trials needed to get the first success in a series of Bernoulli trials with success probability p.
  - Expected Value: $E(X) = \frac{1}{p}$
  - Variance: $Var(X) = \frac{1-p}{p^2}$
  - Probability Mass Function: $P(X = x) = (1 - p)^{x-1} p, x = 1, 2, \ldots$
  - Cumulative Distribution Function: $F(x) = 1 - (1 - p)^x$
  - Example: Flipping a coin repeatedly until the first heads appears.
- Hypergeometric
  - Interpretation: The number of successes in n draws from a population of size N containing K successes, without replacement.
  - Expected Value: $E(X) = n\frac{K}{N}$

  - Variance: $Var(X) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$
  - Probability Mass Function: $P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}, x \in [0, \min(n, K)]$(but discrete)
  - Cumulative Distribution Function: Usually calculated using software, beyond the scope of this class.
  - Example: Drawing n cards from a deck and counting how many are hearts (successes)
- Negative Binomial
  - Interpretation: Number of trials needed to achieve k successes in a series of Bernoulli trials with success probability p
  - Expected Value: $E(X) = \frac{k}{p}$
  - Variance: $Var(X) = \frac{k(1-p)}{p^2}$
  - Probability Mass Function: $P(X = x) = \binom{x-1}{k-1}p^k(1-p)^{x-k}, x = k, k+1, \ldots$
  - Cumulative Distribution Function: Usually calculated using software, beyond the scope of this class.
  - Example: Counting the number of coin flips needed to get k heads.
- Poisson
  - Interpretation: The number of events occurring in a fixed interval of time, assuming events happen independently and at a constant average rate λ
  - Expected Value: $E(X) = \lambda$
  - Variance: $Var(X) = \lambda$
  - Probability Mass Function: $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1, \ldots$
  - Cumulative Distribution Function: $F(x) = e^{-\lambda}\sum_{k=0}^{x}\frac{\lambda^k}{k!}$. Often calculated using software due to infinite summation.
  - Example: The number of cars passing through a toll booth in an hour

Continuous Distributions

- Continuous Uniform
  - Interpretation: All outcomes within the interval $[a, b]$ are equally likely
  - Expected Value: $E(X) = \frac{a+b}{2}$
  - Variance: $Var(X) = \frac{(b-a)^2}{12}$
  - Probability Density Function: $f(x) = \frac{1}{b-a}, a \leq x \leq b$
  - Cumulative Distribution Function: $F(X) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases}$
  - Example: Choosing a random real number between 0 and 1.
- Normal/Gaussian
  - Interpretation: The bell-shaped distribution that a lot of things naturally happen in. Described by its mean μ and standard deviation σ.
  - Expected Value: $E(X) = \mu$
  - Variance: $Var(X) = \sigma^2$
  - Probability Density Function: $f(x) = \frac{1}{2\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Cumulative Distribution Function: Usually calculated using software, beyond the scope of this class.
  - Example: Heights of adults in a population.
- Multivariate Normal
  - Interpretation: Generalization of the normal distribution to multiple variables. It describes a set of variables where any linear combination of the variables is normally distributed. It is defined by a mean vector μ and a covariance matrix Σ, capturing the relationships between variables.
  - Expected Value: $E(X) = \mu$, a vector of means for each variable
  - Variance: $Var(X) = \Sigma$, a symmetric positive definite matrix with variances between variables on the diagonal and covariances on the off-diagonals.

  - Probability Density Function: $f(x) = \frac{1}{det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{k}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$
  - Cumulative Distribution Function: Probabilities usually calculated using software, beyond the scope of this class.
  - Example: Modeling the joint distribution of students' scores in mathematics and physics exams. Students who score well in math may be more likely to score well in physics, so there is covariance between the individual distributions.
- Beta
  - Interpretation: Describes a random variable in the interval $[0, 1]$. Defined by shape parameters α and β.
  - Expected Value: $E(X) = \frac{\alpha}{\alpha+\beta}$
  - Variance: $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
  - Probability Density Function: $f(x) = [\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}]x^{\alpha-1}(1-x)^{\beta-1}, x \in (0, 1)$ where Γ is the Gamma function.
  - Cumulative Distribution Function: Usually calculated using software, beyond the scope of this class.
  - Example: Modeling the probability of success in a Bernoulli trial when parameters are unknown.
- Exponential
  - Interpretation: Describes the time between events in a Poisson process where events happen independently at a constant rate λ. Can also alternatively parameterize with scale but is most often with rate.
  - Expected Value: $E(X) = \frac{1}{\lambda}$
  - Variance: $Var(X) = \frac{1}{\lambda^2}$
  - Probability Density Function: $f(x) = \lambda e^{-\lambda x}, x \geq 0$
  - Cumulative Distribution Function: $1 - e^{-\lambda x}, x \geq 0$
  - Example: The time until the next customer arrives at a store.
- Gamma
  - Interpretation: Generalization of the exponential distribution: a sum of Exponential random variables. Models the total time until the occurrence of k events in a Poisson process where events happen independently at a constant rate λ.
  - Expected Value: With shape/rate parameterization: $E(X) = \frac{k}{\lambda}$. With shape/scale parameterization: $E(X) = k\theta$. Note that $(\theta = \frac{1}{\lambda})$
  - Variance: With shape/rate parameterization: $Var(X) = \frac{k}{\lambda^2}$. With shape/scale parameterization: $Var(X) = k\theta^2$.
  - Probability Density Function: With shape/rate parameterization: $f(x) = \frac{1}{\Gamma(k)}\lambda^k x^{k-1}e^{-\lambda x}$. With shape/scale parameterization: $f(x) = \frac{1}{\theta^k\Gamma(k)}x^{k-1}e^{\frac{-x}{\theta}}$ with $x \geq 0$.
  - Cumulative Distribution Function: Usually calculated using software, beyond the scope of this class.
  - Example: The total time until the k-th event occurs, such as the time until the third customer arrives at a store.

General Probability Laws

Here are a few general probability laws that you should be aware of for this course.

An **event** is any outcome or a set of outcome in a random experiment. The **sample space**, denoted S is the set containing all possible events.

For an event A in sample space S, $P(A)$ refers to the probability of A. The following three rules must always hold:

- $P(A) \geq 0$
- $P(S) = 1$
- For all independent events E_i in S, $\sum P(E_i) = P(S) = 1$.

For two independent events A, B in a sample space S, $P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$.

For two events A, B in a sample space S, $P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

If A and B are **mutually exclusive**, $P(A \cap B) = \emptyset$.

The conditional probability of event A given that event B has occurred is $P(A|B) = \frac{P(A \cap B)}{P(B)}$ given that $P(B) > 0$.

Two events A and B are said to be **independent** if any one of the following holds. If none hold, the events are dependent.

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \cap B) = P(A)P(B)$

The following equality holds: $P(A \cap B) = P(A) * P(B|A) = P(B) * P(A|B)$.

For event A, the probability of A compliment (not A) is equal to $1 - P(A)$. There are multiple forms of notation such as $P(\bar{A})$, $P(A^C)$, $P(A^{'})$. It is often easier to calculate $P(A^C)$ and then we can find $P(A)$ from that.

For some positive integer k, let the sets $B_1, \ldots, B_k$ be such that $S = \sum_k B_k$ and $B_i \cap B_j = \emptyset$ for $i \neq j$. Then the collection of sets $\{B_1, \ldots, B_k\}$ is called a **partition** of S.

Law of Total Probability: Assume $\{B_1, \ldots, B_k\}$ is a partition of S and A is some subset of S. Then, A can be decomposed as follows: $A = (A \cap B_1) \cup \ldots \cup (A \cap B_k)$. Therefore, if $P(B_i) > 0$, $P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$.

From the law of total probability, we can derive **Bayes' Rule**: $P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$.

- For scalar a and random variables X and Y, $E(aX + Y) = aE(X) + E(Y)$. So, the expected value is linear.
- For independent random variables X and Y, $E(XY) = E(X)E(Y)$.
- For a scalar a, $E(a) = a$.
- For scalar a and independent random variables X and Y, $Var(aX - Y) = a^2 Var(X) + Var(Y)$. So, the variance is not linear.
- For not independent random variables X and Y, $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
- For a scalar a, $Var(a) = 0$.
- For a scalar a and random variable X, $Cov(X, a) = 0$. However, the converse does not hold.
- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- If X and Y are independent, $Cov(X, Y) = 0$. However, the converse does not necessarily hold.
- $Cov(X, Y) = 0 \iff Corr(X, Y) = 0$, given that $Var(X), Var(Y) > 0$.
- $Cov(aW + bX, cY + dZ) = acCov(W, Y) + adCov(W, Z) + bcCov(X, Y) + bdCov(X, Z)$