Probability

based on *A First Course in Probability* by Sheldon Ross

Kanyes Thaker

Last updated: February 19, 2023

Note to the Reader

These notes aim to provide a high level, but complete, introduction to probability. The contents range from basic definitions, examples of distributions, to sampling methods. Probability is a fundamental area of mathematics, one that powers almost every analytical discipline, from the pure sciences to economics and finance. An understanding of the fundamentals is vital to be able to do important work in any field. Probability theory is one area which is so core to virtually all aspects of life that it deserves such a complete treatment.

Contents

❖ Counting

The study of probability is complemented by the study of counting, more broadly known as **combinatorics**. If we perform an experiment, we may note the number of possible results, known as **outcomes**. The most fundamental principle of counting is that if one experiment has n outcomes, and a second experiment has m, there are mn total possible outcomes between the two of them. In general, for r experiments where each experiment has n_i outcomes, there are $\prod n_i$ total possible outcomes. Here we are going to explore some of the most common, simple classes of counting problems.

1.1 Permutation

This problem of ordering objects is called a **permutation**. We can decompose this into a series of experiments. The first experiment is to select one of n objects. The second experiment is to select one of $n-1$ remaining objects. Doing this for all n objects, we conclude there are $n! = n(n-1)...1$ possible orderings. If all the objects are *not* distinct, we are *overcounting*. For example, there are two ts in the word "permutation." These identical ts could be in any of 2! possible arrangements. To adjust for this, if object i appears n_i times, we divide $n!$ by $n_i!$, the number of orderings of object i, so the total number is

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

1.2 Combination

Another problem is the act of selecting r items out of n possible items, called a **combination**. In this case, the order of the selected options doesn't matter. There are n ways to choose one object, $n-1$ to choose the second, and so on, with $n-r+1$ ways to choose the last one ($n!/(n-r)!$). We then divide this by the number of ways to arrange those r objects, $r!$ (since order doesn't matter). We write the number of combinations of r items out of a pool of n as

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

If the order of the selection *did* matter, we can reframe this permutation problem, where we want to order n items where the $n-r$ unselected items can be thought of as being "identical."

The values $\binom{n}{r}$ are sometimes referred to as **binomial coefficients** due to their prevalence in the well-known **binomial theorem**.

Binomial theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

There is a recursive definition for $\binom{n}{k}$. We can fix one of the n objects. Then there are $\binom{n-1}{r-1}$ collections of size r that contain that object, and $\binom{n-1}{r}$ collections of size r that do not contain that object. So $\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$. This definition can be used to power a proof by induction of the binomial theorem.

1.3 Multinomials

The last class of combinatorial problems involves counting the number of ways of partitioning a set into r subgroups, where subgroup i has size n_i, and $\sum n_i = n$. There are $\binom{n}{n_1}$ ways of picking n_1 items for group 1. Then there are $\binom{n-n_1}{n_2}$ ways of picking n_2 items for group 2. Multiplying all these together yields

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

total possible configurations. The values $\binom{n}{n_1,\ldots,n_r}$ are known as **multinomial coefficients**. Note that the above looks like a permutation, and we can think of it as one. Instead of assigning objects to categories, we can consider having in our hands a set of labels – n_1 labels that say "1", n_2 labels that says "2", etc. Then lining up all of our objects, the problem is equivalent to the number of orderings of labels we can assign, which is a permutation problem.

1.4 Balls and Bins

The final problem we will consider in this brief introduction to combinatorics is the famous "stars-and-bars" or "balls-and-bins" problem; this is a formulation that appears countless times in computer science. Imagine we have n distinct balls that we need to allocate into r distinct bins. There are r^n possible allocations – for each ball, there are r possible bins to put it in. The problem becomes more complex if we consider that the balls are now *in*distinguishable.

$$(\circ\ \circ\ \circ)(\circ\ \circ)(\circ\ \circ\ \circ\ \circ)(\circ)$$

If each bin must have at least one ball, we can reformulate this in a way where we only care about the divisions between groups of balls. We can picture having n *stars*, which we must separate into r groups by placing $r-1$ *bars* between them. There are $n-1$ places we can put these bars, hence the total number of allocations is $\binom{n-1}{r-1}$.

$$\star\star\star\,|\,\star\star\,|\,\star\star\star\star\,|\,\star$$

If not every bin needs a ball, we can think of the problem as having a total of $n + r - 1$ empty "slots," which we must fill with $r - 1$ bars and n stars. Then our answer is the multinomial $\binom{n+r-1}{n,r-1} = \binom{n+r-1}{r-1} = \binom{n+r-1}{n}$.

$$\_\_\_\_\_ + \star\star\star\star\star\star\star\star\star\star\star\star\star\star\star \rightarrow |\star\star|\star\star\star|\star\star\star\star\star$$

❖ Fundamentals

For a particular experiment, we denote with the symbol Ω the set of all possible outcomes. A subset $\mathcal{E} \subset \Omega$ is called an **event**. These sets are subject to the common rules of set algebra; namely, **union** ($\cup$) **intersection** ($\cap$), the **empty set** ($\varnothing$), **subset** ($\subset$), **superset** ($\supset$) and complement ($\mathcal{E}^c$) are defined as they would be for any general set. We also abide by the general commutative, associative, and distributive laws, where union can be though as an analogue to addition and intersection can be treated like multiplication.

DeMorgan's Laws

$$\left(\bigcup_{i=1}^{n} \mathcal{E}_i\right)^c = \left(\bigcap_{i=1}^{n} \mathcal{E}_i^c\right)$$

$$\left(\bigcap_{i=1}^{n} \mathcal{E}_i\right)^c = \left(\bigcup_{i=1}^{n} \mathcal{E}_i^c\right)$$

The most intuitive way to think about probability is in terms of *relative frequency*. If $n(\mathcal{E})$ is the number of times event $\mathcal{E}$ occurs in n trials, we can define the **probability of** $\mathcal{E}$ as

$$P(\mathcal{E}) = \lim_{n\to\infty} \frac{n(\mathcal{E})}{n}.$$

We make a big assumption here, and that is that this limit exists and is well-defined. We typically define this "frequentist" probability in terms of more digestible, atomic axioms, known as the **Kolmogorov axioms**:

1. $0 \leq P(\mathcal{E}) \leq 1$
2. $P(\Omega) = 1$
3. $P\left(\bigcup_{i=1}^{\infty} \mathcal{E}_i\right) = \sum_{i=1}^{\infty} P(\mathcal{E}_i)$ where each $\mathcal{E}_i$ is mutually exclusive with every other.

Here we operate under the assumption that $P(\mathcal{E})$ is defined for all $\mathcal{E}$, which is only true for *measurable* sets. We do not cover a measure-theoretic approach to probability here. There are several corrolaries to the above, the most important being $P(\mathcal{E}^c) = 1 - P(\mathcal{E})$, and if $\mathcal{E} \subset \mathcal{F}$ then $P(\mathcal{E}) \leq P(\mathcal{F})$.

In particular, we note axiom (3), which is true only when the $\mathcal{E}_i$ are mutually exclusive. In the event that they are *not*, we have a looser inequality.

Generalized inclusion-exclusion

Property (3) does not generally hold when mutual exclusivity is violated. Instead, we have **Boole's Inequality**:

$$P\left(\bigcup_{i=1}^{\infty}\mathcal{E}_i\right) \leq \sum_{i=1}^{\infty} P(\mathcal{E}_i).$$

In particular, $P(\mathcal{E}_1 \cup \mathcal{E}_2) = P(\mathcal{E}_1) + P(\mathcal{E}_2) - P(\mathcal{E}_1\mathcal{E}_2)$. This is because for two events, we "double count" their intersection. For three events, removing the pairwise intersection "overcorrects" for the intersection of all three events, so $P(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) = P(\mathcal{E}_1)+P(\mathcal{E}_2)+P(\mathcal{E}_3)-P(\mathcal{E}_1\mathcal{E}_2)-P(\mathcal{E}_2\mathcal{E}_3)-P(\mathcal{E}_1\mathcal{E}_3)+P(\mathcal{E}_1\mathcal{E}_2\mathcal{E}_3)$:

$$P\left(\bigcup_{i=1}^{n}\mathcal{E}_i\right) = \sum_{k=1}^{n}(-1)^{k+1}\left(\sum_{1\leq i_1<\cdots<i_k\leq n} P\left(\bigcap_{j=1}^{k}\mathcal{E}_{i_j}\right)\right).$$

In some cases, it is safe to assume all events are equally likely. Then finding the probability is as simple as counting the size of $\mathcal{E}$ and dividing by the size of Ω. As an example, suppose we want to find the number of ways we can deal a full house from a deck of cards. There are $\binom{52}{5}$ possible hands in a deck. There are $\binom{4}{2}$ ways to choose two cards of the same rank, of 13 ranks to choose from. Then there are $\binom{4}{3}$ ways to choose three cards of the same rank for each of the remaining 12 ranks. So the total number of ways to deal a full house is $(12 \cdot 13\binom{4}{2}\binom{4}{3})/\binom{52}{5}$.

What we have discussed so far is a *frequentist* view of probability, wherein we model probability as the outcome of a series of repeatable experiments. This is not always possible – for example, the probability that a stock will go up or down based on the current market is not a repeatable task. These kinds of problems require a framing wherein probability represents a measure of *belief* in a particular outcome versus a measurement of an experiment. This is the foundation for the **Bayesian** interpretation.

❖ Conditional Probability and Independence

Here we discuss the idea that the probability of an experiment changes when we have partial information available. For example, consider the probability that the sum of two die is 8, given that one of them is 3. There are 36 possible die rolls we can have – but because of the partial information we are given, the probability of 30 of those outcomes (where the first die isn't 3) is 0. Of the remaining 6 options, exactly one satisfies our condition, so the probability is 1/6.

In this way we can think of these **conditional probabilities** as restrictions of the sample space. Suppose we are given an event $\mathcal{F}$. For $\mathcal{E}$ to occur, it needs to be in $\mathcal{E} \cap \mathcal{F}$. Additionally, since $\mathcal{F}$ has occurred, our potential sample space is restricted to $\mathcal{F}$. So

our conditional probability is

$$P(\mathcal{E}|\mathcal{F}) = \frac{P(\mathcal{E} \cap \mathcal{F})}{P(\mathcal{F})} \iff P(\mathcal{E}\mathcal{F}) = P(\mathcal{E}|\mathcal{F})P(\mathcal{F})$$

The latter formulation generalizes to the **multiplication rule**:

$$P\left(\bigcap_{i=1}^{n} \mathcal{E}_i\right) = P(\mathcal{E}_1)P(\mathcal{E}_2|\mathcal{E}_1)(\mathcal{E}_3|\mathcal{E}_1\mathcal{E}_2)... = \prod_{i=1}^{n} P\left(\mathcal{E}_i \middle| \bigcap_{j=1}^{i-1} \mathcal{E}_j\right)$$

As an illustration, let's find the probability that a dealing of 52 cards into 13 hands has exactly one ace in each hand. This is the intersection four events:

$\mathcal{E}_1$ The ace of spaces is in a pile.

$\mathcal{E}_2$ The ace of spades and ace of hearts are in different piles.

$\mathcal{E}_3$ Spades, hearts, and diamonds are in different piles.

$\mathcal{E}_4$ All four cards are in different piles.

Trivially $P(\mathcal{E}_1) = 1$. Then there are 51 cards left to choose from, of which 12 must go to the ace-of-spades hand (that do *not* include the ace of hearts). So $P(\mathcal{E}_2|\mathcal{E}_1)$ is the number of possible spots for the ace of hearts, which is $(51 - 13)/50 = 39/51$. Then there are 26 spots left for the ace of diamonds where it would be in a different hand, of 50 remaining spots to choose from – so $P(\mathcal{E}_3|\mathcal{E}_2\mathcal{E}_1) = 26/50$. Likewise $P(\mathcal{E}_4|\mathcal{E}_3\mathcal{E}_2\mathcal{E}_1) = 13/49$. So the end result is the product of these four probabilities, which is roughly 0.105. Note that conditional probabilities still satisfy all the Kolmogorov axioms, meaning conditional distributions are valid probability distributions themselves.

3.1 Bayes' Theorem

The above ideas give rise to a very powerful notion – since $\mathcal{E} = (\mathcal{E} \cap \mathcal{F}) \cup (\mathcal{E} \cap \mathcal{F}^C)$, and $P(\mathcal{E} \cap \mathcal{F}) = P(\mathcal{E}|\mathcal{F})P(\mathcal{F})$, we may be able to determine the probability of an event by "conditioning" using another – so that if we can't measure $\mathcal{E}$ exactly, we can use its interactions with $\mathcal{F}$ to determine $P(\mathcal{E})$.

$$P(\mathcal{E}) = P(\mathcal{E}|\mathcal{F})P(\mathcal{F}) + P(\mathcal{E}|\mathcal{F}^C)(1 - P(\mathcal{F}))$$

As an example, suppose a certain medical diagnostic test correctly identifies the presence of a disease 95% of the time when the patient actually has the disease (this is the **true positive rate**). When the patent does not have the disease, the test correctly returns negative 99% of the time (the **true negative rate**). Note this also means that in the first case, the test is wrong 5% of the time (**false negative**) and in the second, it is wrong 1% of the time (**false positive**). If the disease actually presents in 0.5% of the population,

we can use the above to determine the probability that a patient has the disease given the test returns positive.

We are looking for $P(D|+)$ where D is the event that the patient has the disease, and $+$ is the event of a positive test result. So:

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{P(D \cap +)}{P(+|D)P(D) + P(+|D^C)P(D^C)} = \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = 0.32$$

Surprisingly, if you have a positive result on the test, you only have a 32% chance of actually having the disease!

The central idea here is that we have a certain **hypothesis** that we are trying to defend, whose true probability is unknown. Accruing **evidence** can help either increase or decrease the probability that the hypothesis is true. The evidence supports our hypothesis whenever $P(\mathcal{H}|\mathcal{E}) \geq P(\mathcal{H})$. This happens whenever $P(\mathcal{E}|\mathcal{H}) \geq P(\mathcal{E}|\mathcal{H}^C)$. Equivalently, the **odds** of the hypothesis are being true $\frac{P(\mathcal{E}|\mathcal{H})}{P(\mathcal{E}|\mathcal{H}^C)}$ are greater than one. This means that our evidence supports the hypothesis whenever the *evidence* is made more probable when the hypothesis is true versus when it is false.

Bayes' Theorem

Let $\mathcal{H}$ be our hypothesis. Suppose we are given some evidence $\mathcal{E}$ and want to calculate how our current (or **prior**) estimate of $P(\mathcal{H})$ is changed by knowing $\mathcal{E}$ (the **posterior** $P(\mathcal{H}|\mathcal{E})$). We calculate the posterior distribution by multiplying prior by the **likelihood** that the hypothesis explains the evidence $P(\mathcal{E}|\mathcal{H})$ and dividing by the **marginal** probability of $\mathcal{E}$, $P(\mathcal{E})$.

$$P(\mathcal{H}|\mathcal{E}) = \frac{P(\mathcal{E}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{E})}$$

3.2 Independence

Having information about a second event doesn't necessarily mean the probability of the first event changes. In the case where $P(\mathcal{E}|\mathcal{F}) = P(\mathcal{E})$ we say that $\mathcal{E}$ and $\mathcal{F}$ are **independent**. This has the further implication that $P(\mathcal{E} \cap \mathcal{F}) = P(\mathcal{E})P(\mathcal{F})$.

❖ Random Variables

Often times we are not explicitly concerned with the direct outcome of an experiment, but rather with some real-valued function of the output. For example, many games will depend on the value of the sum of a roll of two die, regardless of what the actual events are. A function which maps events in the output space to real numbers, $X : \Omega \mapsto \mathbb{R}$, is known as a **random variable**. Just as we can assign a probability to the events in Ω, we can equivalently assign probabilities to the values that a random variable can assume.

For a random variable X, the function F defined by $F(x) = P(X \leq x)$ is known as the **cumulative distribution function** (CDF) of X. Due to the fact that probability distributions are strictly positive, $F(x)$ is a strictly increasing function.

4.1 Discrete Random Variables

If a random variable X can take on a finite number of values, it is called a **discrete random variable.** We can define a function p whose domain is the set of values X can assume. This function, $p(a) = P(X = a)$ is known as the **probability mass function**. It has the property that $p(x_i) \geq 0$ if X can possibly be x_i, and 0 otherwise. Additionally, $\sum_{x_i} p(x_i) = 1$ (this is so that we adhere to the Kolmogorov axioms). We then write the cumulative distribution as

$$F(a) = \sum_{x \leq a} p(x).$$

For discrete random variables, $F(a)$ looks like a *step function* when plotted.

4.2 First and Second Moments

Often times it is sufficient to ignore the overall probability distribution (mass function) if we have access to certain values which concisely summarize the information represented by that distribution. One such measure is the **expected value** or **expectation**.

Expected Value

For X distributed according to $p(X)$, the expected value is

$$\mathbb{E}[X] = \sum_{x:p(x)>0} xp(x).$$

The expectation is a weighted average of the possible values of X, where the weights are the probabilities of each value. In a frequentist setting, this can be thought of the average outcome per trial if the experiment is repeated infinitely. Expected value is an analogous concept to the center of gravity of an object, i.e. the point in an object about which all torques sum to zero.

Indicator Random Variables

A random variable $\mathbf{1}_A$ is an *indicator* of event A if

$$\mathbf{1}_A = \begin{cases} 1 & A \\ 0 & A^C \end{cases}$$

As a result $\mathbb{E}[\mathbf{1}_A] = p(A)$.

We can also think of the expected value of a *function* of a random variable, $g(X)$. The expectation $\mathbb{E}[g(X)]$ can be thought of as a weighted average of the values of $g(X)$, weighted by the probabilities of X, meaning

$$\mathbb{E}[g(X)] = \sum_{x:p(x)>0} g(x)p(x).$$

This has the consequence that for $g(X) = aX+b$, we have $\mathbb{E}[aX+b] = a\mathbb{E}[X]+b$ (**linearity of expectation**).

The quantity $\mathbb{E}[X]$ is also referred to as the **first moment** of X. The quantities $\mathbb{E}[X^n]$ are the n**th moments** of X. Moments are quantities that capture some essential quality about the shape of a function, in this case the "center."

The expected value gives us limited information. It tells us the central value for the random variable, but does not give us any information on how the random variable behaves at locations other than the mean. For that, we try to measure how much the function *varies*, which we can think of as determining the distance (more typically, the *squared* Euclidean distance) X typically is from its mean:

$$\begin{aligned}\mathbb{E}[(X-\mu)^2] &= \sum_x (x-\mu)^2 p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

This quantity is the **variance** of X, written $\mathrm{Var}(X)$, with the property that $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$. Variance is also the (mean-adjusted) second moment, known in physics as the **moment of inertia**. We sometimes instead use the square root of variance, $\sigma(X) = \sqrt{\mathrm{Var}(X)}$, known as the **standard deviation**.

4.3 Bernoulli and Binomial

A random variable that outputs 1 with probability p and 0 with probability $(1-p)$ is known as a **Bernoulli** random variable. If we repeat n Bernoulli experiments, the total number of successes is a **binomial** random variable $X \sim Binom(n,p)$. The probability mass function is the probability of getting i successes of n trials. The probability of exactly i successes and $n-i$ failures is $p^i(1-p)^{n-i}$, and there are $\binom{n}{i}$ ways to arrange the successes and failures.

$$p(i) = \binom{n}{i} p^i (1-p)^{n-i}.$$

We can also calculate the moments of X.

$$\mathbb{E}[X^k] = \sum x^k p(x) = \sum i^k \binom{n}{i} p^i (1-p)^{n-i} = np \sum (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-j-1}$$

where $j = i - 1$ and we take advantage of the fact that $i\binom{n}{i} = n\binom{n-1}{i-1}$. Then $\mathbb{E}[X^k] = np\mathbb{E}[(Y+1)^{k-1}]$ where $Y \sim Binom(n-1, p)$. Plugging in $k = 1$ gives us $\mathbb{E}[X] = np$, and plugging in $k = 2$ gives us $\mathbb{E}[X^2] = np(n-1)p + 1$, meaning $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np(1-p)$.

4.4 Poisson

The binomial mass function can become infeasible compute when n is extremely large or p is extremely small. In these cases – where np is "medium sized," we might try to approximate the result. Letting $\lambda = np$:

$$\begin{aligned} p(i) &= \binom{n}{i} p^i (1-p)^{n-i} \\ &= \frac{n!}{(n-1)!i!} p^i (1-p)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{i!} \frac{\lambda^i}{n^i} \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \frac{\lambda^i}{i!} \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} \\ &\simeq e^{-\lambda} \frac{\lambda^i}{i!}. \end{aligned}$$

The final approximation comes from the fact that, for large n and $i \ll n$, $\frac{n(n-1)\cdots(n-i+1)}{n^i} \simeq 1$ and $(1-\lambda/n)^i \simeq 1$. Also, for λ sufficiently greater than zero $(1-\lambda/n)^n \simeq e^{-\lambda}$. This is the probability mass function for the **Poisson** distribution with parameter λ, $X \sim Pois(\lambda)$. Since $\lambda = np$ and $\mathbb{E}[Y \sim Binom(n,p)] = np$, we might anticipate that $\mathbb{E}[X] = \lambda$, and likewise that $\text{Var}(X) = \text{Var}(Y) = np(1-p) \approx np = \lambda$ for small p. In general, the Poisson distribution appears when a large number of experiments occur where the probability of each of those events is small, and the experiments are either fully independent or only have "weak" dependence, i.e. $P(E_i|E_j) \simeq P(E_i)$ for large n.

4.5 Other Discrete Distributions

1. Suppose an experiment succeeds with probability p and fails with probability $1-p$ Then the random variable X describing the iteration in which the first success happens is a **geometric** random variable $X \sim Geom(p)$ with $p(n) = (1-p)^{n-1}p$. If we let $q = 1-p$ then

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} n(1-p)^{n-1}p = p\sum \frac{d}{dq}q^n = p\frac{d}{dq}\sum q^n = p\frac{d}{dq}\left(\frac{1}{1-q}\right) = \frac{1}{p}.$$

   The variance is then $(1-p)/p^2$.

2. Suppose an experiment that succeeds with probability p is performed until r experiments succeed. The random variable X describing the total number of

trials required is a **negative binomial** random variable with parameters (p, r), $X \sim NB(p, r)$.

$$p(n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}, \quad n \geq r.$$

Its mean and variance are $\mathbb{E}[X] = rp/(1-p)$ and $\text{Var}(X) = rp/(1-p)^2$; we can think of the negative binomial as r independent geometric random variables.

3. Suppose we select n objects from a collection of size N, of which m belong to a category and $N - m$ belong to another. The number of objects in our selection belonging to the first class represents a **hypergeometric** random variable,

$$p(i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}, \quad i \leq m.$$

Its mean is nm/N and its variance is $np(1-p)\left(1 - \frac{n-1}{N-1}\right)$.

4. A random variable with the probability mass function

$$p(k)\frac{C}{k^{\alpha+1}}$$

where $\alpha > 0$ and C is a normalizing coefficient is known as the **zeta** or **Zipf** distribution. It receives its name from its relationship to the Riemann zeta function $\zeta(s) = \sum \left(\frac{1}{n}\right)^s$.

❖ Continuous Random Variables

A random variable X that can take on an uncountably infinite set of values are known as **continuous** random variables if there exists a nonnegative function f over $\mathbb{R}$ such that, for any measurable set $B \subset \mathbb{R}$, $P[X \in B] = \int_B f(x)dx$. f is the **probability density function** of X. As in the discrete case, the $\int_{\mathbb{R}} f(x)dx = 1$. It doesn't make sense to query a single point in a continuous distribution, as any point has probability zero. Instead, we query an interval $[a, b]$ whose probability is then $\int_a^b f(x)dx$. The cumulative distribution function for a continuous random variable is simply $F(a) = \int_{-\infty}^a f(x)dx$, meaning $F'(a) = f(a)$. For continuous random variables, the expectation is $\mathbb{E}[X] = \int_{\mathbb{R}} xf(x)dx$, and the variance is defined similarly. As in the discrete case, $\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x)f(x)dx$. Also like the discrete case, $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, and $\text{Var}(aX + b) = a^2\text{Var}(X)$.

A random variable is said to be **uniformly distributed** over $[0, 1]$ if $f(x) = 1$ for $0 < x < 1$ and $f(x) = 0$ otherwise. More generally, a random variable is uniform over $[a, b]$ if its density function is $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and zero otherwise. The uniform random variable has mean $\frac{b-a}{2}$ and variance $\frac{(b-a)^2}{12}$.

5.1 The Gaussian Distribution

A random variable X is **normal** or **Gaussian** with mean μ and variance σ^2 if the density function of X is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

The density function is a bell-shaped symmetric curve about μ, and was formulated as a way to approximate the binomial for large n. The normal distribution appears frequently in the natural world due to the central limit theorem (later in the text), and is one of the most essential distributions in many sciences.

It is traditional to denote the cumulative distribution function of a **standard normal** (zero-mean, unit-variance) Gaussian with Φ, as

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{y^2}{2}\right\}$$

One of the most critical results in probability theory is the idea that when n is large, the binomial distribution with parameters n and p will have approximately the same distribution as a Gaussian distribution with the same mean and variance.

The DeMoivre-Laplace Limit Theorem

If S_n is the number of successes for n independent trials, succeeding with probability p, then for $a < b$:

$$\lim_{n\to\infty} P\left\{a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right\} = \Phi(b) - \Phi(a).$$

5.2 Other Continuous Distributions

Here is a short list of other vital continuous distributions that often appear in the sciences.

1. A continuous random variable with density function (with parameter λ) $\lambda e^{-\lambda x}$ for $x \geq 0$ and 0 otherwise is an **exponential** random variable $X \sim Expo(\lambda)$. It has mean $1/\lambda$ and variance $1/\lambda^2$. It is typically used with respect to arrival times; while the Poisson distribution can be used to model the number of (rare) events occurring in a time frame, the exponential distribution models the amount of time until the next occurrence. The exponential distribution is **memoryless**, meaning that the probability of an event occurring in the next t time is consistent no matter what the current time is.

2. The **Gamma distribution**, $X \sim \Gamma(\alpha, \lambda)$ has density function (for nonnegative x)

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}.$$

The Gamma distribution represents the wait time until the αth event occurs; the exponential distribution is a special case of the Gamma distribution when $\alpha = 1$. In literature, if we restrict α to positive integers, we may alternatively call the distribution the n**-Erlang** distribution.

3. A **Weibull** random variable, $X \sim Weib(\alpha, \beta, \nu)$, is one whose *cumulative* density function (for $x > \nu$) is
$$1 - \exp\left\{-\left(\frac{x-\nu}{\alpha}\right)^{\beta}\right\}.$$
This distribution is prominently used in reliability engineering, in the process of predicting the probability of a system failure.

4. The **Cauchy** distribution, $X \sim Cauchy(\theta)$, also known as the **Lorentz** or **Cauchy-Lorentz** distribution, is a well-known distribution in physics describing resonance patterns, and is given by the density function
$$f(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}.$$
It can be thought of as the location where a beam of light shone from $(0, 1)$ hits the x-axis, assuming that beam makes an angle θ with the y-axis.

5. A random variable has a **Beta** distribution if its density, for $0 < x < 1$, is
$$f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1},$$
where B is the Beta function
$$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$
The Beta distribution is useful in modeling the probability distribution of a *probability*. One good example (not my own): suppose a baseball batter bats exceptionally poorly their first three games of the season. We might give them a batting average of 0. But we know this is nonsensical, since the average batting average is around 0.25. So it's far more likely that their batting average is actually just a bit below 0.25, not all the way down at 0. The Beta distribution helps model this prior knowledge.

❖ Jointly Distributed Random Variables

The discussion so far concerns distributions of single random variables. However, we are often curious about the probabilities associated with multiple random variables at once. In the case of two events, the **joint cumulative distribution function** is
$$F(a,b) = P(X \le a, Y \le b).$$

The individual distributions of X and Y, $F_X(a) = F(a, \infty)$ and $F_Y(b) = F(\infty, b)$, are known as **marginal** distributions. For discrete distributions, the joint mass function is $p(x, y) = P(X = x, Y = y)$, where the individual marginal probabilities can be retrieved by summing over the other variable, i.e.

$$p_X(x) = \sum_y P(X = x, Y = y).$$

We call X and Y **jointly continuous** if there is some $f(x, y) : \mathbb{R}^2 \mapsto [0, 1]$ (a **joint density function**) such that, for all subsets $C \subset \mathbb{R}^2$,

$$P((X, Y) \in C) = \iint_{(x,y)\in C} f(x, y)dxdy.$$

As in the discrete case, we can marginalize the distribution, this time with integration, as in

$$f_X(x) = \int_{\mathbb{R}} f(x, y)dy.$$

These definitions naturally extend to n random variables (beyond just the two we use for illustration). As an example of the above concepts, suppose we are trying to find the distribution of some continuous random variable X/Y, given $f(x, y)$. We can set up the appropriate integral using the definition of the joint cumulative distribution function, and then differentiate to find the density:

$$F_{X/Y}(a) = P(X/Y \leq a) = \iint_{x/y\leq a} f(x, y)dxdy = \int_0^\infty \int_0^{ay} f(x, y)dxdy \rightarrow$$

$$f_{X/Y}(a) = \frac{d}{da} \int_0^\infty \int_0^{ay} f(x, y)dxdy.$$

6.1 Independent Random Variables

X and Y are independent if for any sets of numbers $\mathcal{A}$ and $\mathcal{B}$,

$$P(X \in \mathcal{A}, Y \in \mathcal{B}) = P(X \in \mathcal{A})P(Y \in \mathcal{B}).$$

As such, a necessary and sufficient condition for independence is that the joint probability density/mass function $f(x, y)$ cleanly factors into exactly two terms, one of which is exclusively dependent on x and the other exclusively dependent on y, i.e. $f(x, y) = h(x)g(y)$. Independence is a symmetric relation, meaning that $X \perp\!\!\!\perp Y \iff Y \perp\!\!\!\perp X$.

Suppose we wish find the distribution of a sum of independent random variables, $X+Y$. Then we could use the same method as we did above to find

$$\begin{aligned} F_{X+Y}(a) = P(X + Y \leq a) &= \iint_{x+y\leq a} f(x, y)dxdy \\ &= \int_{-\infty}^\infty \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy = \int_{\mathbb{R}} F_X(a - y)f_Y(y)dy. \end{aligned}$$

This is called performing a **convolution** on the cumulative distribution functions F_X and F_Y. In general, adding two independent random variables amounts to performing a convolution on their cumulative distribution functions and differentiating the result. The sum of uniform distributions is not uniform, but the sum of normal distributions *is*, with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. We omit the other various distributions.

6.2 Conditional Distributions

If X and Y are discrete random variables, the conditional probability mass function is

$$p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{p(x, y)}{p(y)}.$$

The conditional distribution function is then likewise

$$F_{X|Y}(x|y) = \sum_{a \leq x} p_{X|Y}(a|y).$$

In the continuous case, the conditional density function of $x|y$ is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

Note that we have a workable expression for the continuous case, even though the probability of the event in question $P(Y = y) = 0$. This idea works even if X and Y are not jointly continuous or jointly discrete, i.e. if X is continuous and Y is discrete.

6.3 Order Statistics

If $X_1, ..., X_n$ are i.i.d. continuous random variables, and $X_{(1)}$ is the smallest among them, $X_{(2)}$ is the next smallest, etc., then the total partial ordering $X_{(1)} \leq ... \leq X_{(n)}$ is known as the **order statistics** of the set of variables. The joint density of the order statistics can be thought of a permutation problem – if any of the $X_i = x_j$ then we are satisfied. So the joint density function is simply $n! \prod f(x_i)$. If each X_i is jointly distributed with Y_i, then the corresponding ordering induced on Y by the order statistic is called a **concomitant**.

❖ Expectation

Recall that the expectation can be thought as a weighted sum of the possible values of a random variable, where the weights are the probabilities of those events.

Suppose that we now want to find the expectation of a function $g(X, Y)$ of *two* random variables, X and Y, where X and Y have a joint density f (we omit the discrete case, though it looks almost identical, just with sums instead of the integral).

$$\mathbb{E}[g(X, Y)] = \iint_{\mathbb{R}^2} g(x, y) f(x, y) dx dy$$

When $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are both finite, then using the above we get the result:

$$\mathbb{E}[X+Y] = \iint_{\mathbb{R}^2} (x+y)f(x,y)dxdy = \int_{\mathbb{R}} x f_X(x)dx + \int_{\mathbb{R}} y f_Y(y)dy = \mathbb{E}[X] + \mathbb{E}[Y].$$

Indeed, for any $X_1, ..., X_n$ where each X_i is finite, $\mathbb{E}\left[\sum X_i\right] = \sum \mathbb{E}[X_i]$. This is known as **linearity of expectation**, and is true regardless of whether the X_i are independent.

7.1 The Coupon Collector's Problem

This is a brief explanation of a common probability problem. Suppose a brand is doing a promotion where every item sold contains one of N possible types of coupons, where each type appears with equal probability. What is the expected number of coupons that will need to be collected before one of each type is acquired?

We can model the total number of coupons X as the sum of $X_0, ..., X_{N-1}$, where X_i is the number of additional coupons that need to be obtained to get the $i+1$th unique coupon after i unique coupons have already been collected. The probability $P(X_i = k)$ is $\frac{N-i}{N}\left(\frac{i}{N}\right)^{k-1}$ – probability $p = \frac{N-i}{N}$ that we get the next unique coupon on the kth coupon and $1-p$ for each of the $k-1$ coupons before it. This is a geometric random variable with parameter p. Then since $\mathbb{E}[X_i] = \frac{N}{N-i}$, the total expected number of coupons is

$$\mathbb{E}[X] = N\left[1 + \ldots + \frac{1}{N-1} + \frac{1}{N}\right].$$

7.2 Covariance and Correlations

For independent variables X and Y, the expectation of the product of functions of X and Y is the product of their expectations:

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)].$$

The **covariance** between two random variables gives us information about the relationship between those variables – namely, how much those to variables vary *together*. It is defined as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

If X and Y are independent, varying X tells us nothing about varying Y, so $\mathrm{Cov}(X,Y) = 0$ (though the converse is not generally true – dependent variables can also have zero covariance). The covariance of a variable with itself is its variance.

We can also think about the variance of sums of random variables. The pairwise covariances of a set of random variables $(X_1, ..., X_n)$ form a **covariance matrix**. The

variance of the sum of all variables in that set is equivalent to summing over the entire covariance matrix:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2\sum_{i<j}\sum \text{Cov}(X_i, X_j).$$

Note that this explains the previous formula for the variance of a sum of *independent* random variables – if they are independent, the covariances are all zero.

The **correlation**, $\rho(X, Y)$, of two random variables is defined (for positive $\text{Var}(X)\text{Var}(Y)$) as

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

Correlation and covariance are two quantities often confused in statistics. Intuitively, covariance can be thought of measuring the *direction* of the relationship between two random variables, but is a poor measure of the exact strength of that relationship, because it is impacted by scale. Correlation, a scaled function of covariance, allows us to determine the exact *degree of linearity* (strength of the relationship), ranging from $[-1, 1]$. 1 indicates, for example, that there is a perfectly linear relationship $Y = aX + b$, $a > 0$. Note that both of these ideas only measure the *linear* relationship between random variables, meaning they give little to no information in non-linear cases.

7.3 Conditional Expectation

The conditional probability for X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

Likewise, the conditional expectation can be defined as

$$\mathbb{E}[X|Y = y] = \int_x x f_{X|Y}(x|y) dy$$

We can think of the conditional expectation as $Y = y$ the equivalent of calculating the expected value of $f(x, y)$ with the sample space reduced to $Y = y$.

An important property of expectations is that $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. Note that $\mathbb{E}[X|Y]$ is itself a random variable. We can show this by expanding out the $\mathbb{E}$'s:

Law of Iterated Expectations

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_y \int_x f(x|y) f(y) dx y = \int_y \int_x x f(x, y) dx dy = \int_x x f(x) dx = \mathbb{E}[X].$$

To intuit this a bit better: to calculate $\mathbb{E}[X]$, we may take a weighted average of $\mathbb{E}[X|Y]$ over all possible values of Y, where the weights are the probabilities of each $Y = y$.

This result lets us readily compute expected values when the target variable is difficult to measure directly, but we have access to a conditional probability conditioned on an event with known probability. This is known as the **law of iterated expectations**.

As an example, we may calculate the expected number of uniform random variables in $(0, 1)$ that must be drawn until their sum exceeds 1. We generalize to finding $N(x)$, the number of uniform random variables drawn until their sum exceeds x. Let the drawn variables be listed as $U_1, U_2, ...$ The density function of $Uniform(0, 1)$ is $f(x) = 1$, $x \in (0, 1)$. Then the average number $m(x) = \mathbb{E}[N(x)]$ is

$$m(x) = \int_0^1 \mathbb{E}[N(x)|U_1 = y]dy.$$

If $y > x$ then $m(x)$ is simply 1 (we are already done) – otherwise we can recursively say that $\mathbb{E}[N(x)|U_1 = y], y \leq x$ is $1 + m(x - y)$. So then our equation becomes

$$m(x) = 1 + \int_0^x m(x - y)dy = 1 + \int_0^x m(u)du. \implies m'(x) = m(x)$$

where in the last step we differentiate both sides. Then solving the resulting homogenous differential equation reveals that $m(x) = ke^x$, where $k = 1$ since $m(0) = 1$. Then $m(1)$ is the number of uniform random variables required for their sum to exceed 1, and is $m(1) = 1e^1 = e$.

We can also calculate conditional variance in the same way:

$$\text{Var}(X|Y) = \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y].$$

And just as we can compute the expectation $\mathbb{E}[X]$ from conditional expectations, we can calculate the conditional variance from conditional variances:

$$\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y]).$$

Finally, consider a case where we try to estimate Y by some function $g(X)$, so that when we observe $X = x$ we predict that $Y \simeq g(x)$. This is a problem that appears frequently in statistics, optimization theory, and computer science. Ideally, $g(X)$ is close to Y. In fact, the best estimator is provably $g(x) = \mathbb{E}[X|Y]$, or

$$\mathbb{E}[(Y - g(X))^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y|X])^2].$$

Instead of a rigorous proof, here's an intuitive example – the quantity $\mathbb{E}[(Y - c)^2]$ is minimized by $c = \mathbb{E}[Y]$. If we want to predict Y when we have no way of directly measuring Y, the value that will minimize the mean-squared error is $\mathbb{E}[Y]$. In this case, where we have access to $X = x$, our best estimator is still $\mathbb{E}[Y]$, but just with the additional information present that $X = x$, hence $\mathbb{E}[Y|X = x]$ is our best guess.

7.4 Moment Generating Functions

Consider the function $M(x) = \mathbb{E}[e^{tX}]$. If we take the nth derivative of this function and let $t = 0$, we get:

$$\left.\frac{d^n \mathbb{E}[e^{tX}]}{dt^n}\right|_{t=0} = \mathbb{E}\left[\left.\frac{d^n e^{tX}}{dt^n}\right|_{t=0}\right] = \mathbb{E}[X^n e^{0X}] = \mathbb{E}[X^n].$$

These values are, as previously introduced, the moments of X, and so we call $M(x)$ a **moment-generating function**. In the above we exchange the expectation with the derivative; note that this is not generally possible, but it is for all distributions in this text. For more information, refer to my notes on measure theory for a discussion on the dominated convergence theorem. An important property of the moment generating functions is that the MGF of a sum of random variables is the product of the individual MGFs of each random variable in the sum.

We may also define the joint MGF of multiple random variables:

$$M(t_1, \ldots, t_n) = \mathbb{E}[e^{t_1 X_1 + \ldots + t_n X_n}].$$

The joint moment generating function uniquely determines the joint distribution for the collection $X_1, \ldots, X_n$; the proof of this fact can again be found in a measure-theoretic reference for probability theory.

❖ Limit Theorems

Probabilistic limit theorems fall under two main categories – *laws of large numbers* (limits concerning a sequence of random variables exhibiting some convergence property) or *central limit theorems* (limits concerning the behavior of a sum of a large number of random variables).

8.1 Weak Law of Large Numbers

Let us begin by considering a simple inequality, concerning bounding the probability $P(X \geq a)$ for $X \geq 0$. Let $I = \mathbf{1}_{X \geq a}$. Since $X/a \geq I$ for all values of X, We must have $P(X \geq a) = \mathbb{E}[I] \leq \mathbb{E}[X]/a$:

Markov's Inequality

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

As a corollary, we may evaluate the bound for $P(|X - \mu| \geq k)$ (or equivalently, $P((X-\mu)^2 \geq k^2)$) to analyze the same bound for a zero-mean random variable:

Chebyshev's Inequality

$$P(|X - \mu| \geq k) \leq \frac{\text{Var}(X)}{k^2}$$

We can take this one step further by considering a sequence of i.i.d. random variables $X_1, ..., X_n$. We might consider what the behavior of the mean of this sequence might be as n gets large. Since $\frac{1}{n}\mathbb{E}[\sum X_i] = \mu$ and $\text{Var}\left(\frac{\sum X_i}{n}\right) = \frac{1}{n}\sigma^2$, we can use Chebyshev's inequality to show that

$$P\left(\left|\frac{\sum X_i}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Weak Law of Large Numbers (Khinchin's Law)

$$\lim_{n\to\infty} P\left(\left|\frac{\sum X_i}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

As n grows large, the mean of a sample of n random variables grows infinitely close to the theoretical expected value with high probability.

8.2 Central Limit Theorem

Central Limit Theorem

Let $X_1, ..., X_n$ be a sequence of i.i.d. random variables, with shared mean μ and variance σ^2. Then the distribution of $\frac{1}{\sigma\sqrt{n}}(\sum_i (X_i - \mu))$ is approximately standard-normal in large n. That is:

$$\lim_{n\to\infty} P\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} e^{-x^2/2}dx.$$

8.3 Strong Law of Large Numbers (Kolmogorov's Law)

The strong law of large numbers states that a sequence of independent random variables with common distribution will converge to its mean with probability 1.

Strong Law of Large Numbers

If $X_1, X_2, ...$ is a sequence of i.i.d. random variables, each with finite mean μ, then

$$P\left(\lim_{n\to\infty} \frac{X_1 + ... + X_n}{n} = \mu\right) = 1.$$

Although the two formulations look similar, the WLLN makes a looser assertion than the SLLN. The weak law asserts that, as the number of samples tends to infinity, the probability that the sample mean deviates from the true mean by any significant margin

approaches zero. However, this convergence only happens in *probability*. We make no assumptions about the actual events in the limit – for up to infinitely many values of n, it may be the case that $\left|\frac{1}{n}\sum X_i - \mu\right| \geq \varepsilon$.

The SLLN asserts that this deviation *almost surely* will not occur. The SLLN asserts that the limit of sample means *is equal to* the true mean, meaning that the number of "deviations" is finite. In other words, there is a 0 probability that you could take the limit of the sample mean and find a value not equal to μ. The SLLN in this way *implies* the WLLN, though the other is not true. The WLLN, however, is still applicable in scenarios where the first moment is not known, and as such relies on a weaker set of initial assumptions.

8.4 Chernoff Bounds

Let $M(t)$ be the moment generating function for X. Then $P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \mathbb{E}[e^{tX}]e^{-ta}$ by the Markov inequality. So then we have the following inequalities, referred to as the **Chernoff bounds**:

Chernoff Bounds

$$P(X \geq a) \leq e^{-ta}M(t), \quad t > 0$$
$$P(X \leq a) \leq e^{-ta}M(t), \quad t < 0$$

❖ Processes, Entropy

9.1 Poisson Process

Stochastic processes are among the most widely applicable mathematical objects in probability. A stochastic process is a collection of random variables on a shared probability space, i.e. a model of a system that evolves over time that is probabilistic in nature. A collection of random variables $N(t)$ is called a **Poisson process with parameter** λ if:

- The process begins at time 0 $N(0) = 0$.
- The number of events in disjoint time intervals are independent
- The process is **stationary**, meaning the number of events in an interval only depends on the length of that interval.
- $P(N(t) = 1) = \lambda h + O(h)$ and $P(N(t) \geq 2) = O(h)$.

The random variable $N(t)$ is itself a Poisson random variable with mean λt:

$$P(N(t) = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

Instead of looking at the number of events in an interval, we can instead create a sequence of the *times* when the events occur. This is a sequence of **interarrival times**. For a Poisson process with parameter λ, the interarrival times are independent exponential random variables with mean $1/\lambda$.

9.2 Markov Chains

A sequence of random variables X_n is a **Markov chain** if the probability of $X_t = i$ **transitioning** to $X_{t_1} = j$ is a fixed value P_{ij} *regardless* of all the values X_k had *before* time t. This is the **Markov property**. The probabilities P_{ij} can be arranged into a square array (or **matrix**) where the entry at i, j is the probability of transitioning from state i to state j. By using a superscript, as in $P_{ij}^{(n)}$, we can also talk about the probability that the chain will start in state i and end in state j after n steps (the n-**step transition probabilities**). For Markov chains that have the property where $P_{ij}^{(n)} > 0$ for all i, j (**ergodicity**), The $P_{ij}^{(n)}$ converge in n to some fixed value π_j that depends only on the destination state j (these π_j are called **stationary probabilities**). Intuitively, the stationary probabilities model the fraction of time the system spends in state j over a long period of time.

A common example of a Markov chain is the modeling of a particle as it moves along a one-dimensional axis, where the particle will move either to the left or right with probabilities p and $1 - p$ (the **one-dimensional random walk**). The random walk can be written as a Markov chain, and the n-step transition probabilities reveal information the position of the particle after n steps of the walk.

9.3 Entropy

Information theory deals heavily with the notion of *surprised* – for any system, how much important information do we gain as observers every time a new observation comes in? It makes intuitive sense to tie the idea of information to probability, as seeing a low-probability event occur gives more information than one that is inspected. We use a logarithm to quantify this idea, since we desire a function $h(p)$ such that a high probability event carries no information ($h(1) = 0$), information increases as probability decreases $q > p \implies h(p) > h(q)$, and the amount of surprise is additive ($h(pq) = h(p) = h(q)$).

Entropy

Suppose we have X which can take on any of the values x_i with corresponding probability $p(x_i)$. Then the **entropy** of X is

$$H(X) = -\sum_{i=1}^{n} p(x_i) \ln p(x_i).$$

We leave a more thorough discussion of entropy to a course on information theory.

❖ Simulations and Sampling

For practical applications of probability, we require systems that let us *simulate* sampling from distributions when we do not have access to a natural system that yields that distribution naturally. There are two general methods for sampling from a continuous distribution.

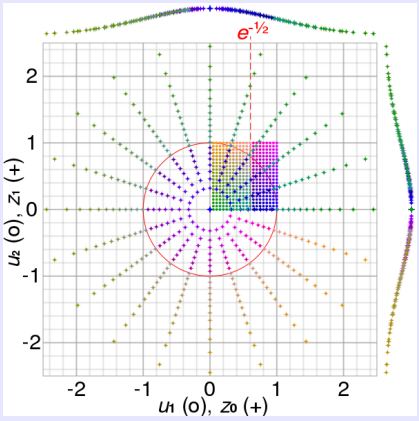
- Let U be a uniform $(0, 1)$ random variable. Suppose our target X has cumulative distribution function F. Since $F : \mathbb{R} \to (0, 1)$, and F is strictly monotonic hence invertible, $Y = F^{-1}(U)$ means Y also has cumulative distribution F. This method is known as **inverse transform sampling**.
- Suppose we are able to reliably sample from some distribution g. We can then scale g by a constant c until $cg \geq f$ everywhere (we create an *envelope*). Then we can sample Y from cg, and then draw a number U from $Uniform(0, 1)$. If $U \leq f(Y)/cg(Y)$ then we keep Y as our drawn sample; otherwise we discard Y and try again. This method of rejecting samples that fall out of the distribution is known as **rejection sampling**.

Box-Muller Transform

The **Box-Muller transform** is a method for sampling from a normal distribution. If U_1 and U_2 are independent uniform random variables on $(0, 1)$, then

$$Z_0 = R\cos(\theta) = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$
$$Z_1 = R\sin(\theta) = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

where $R^2 = -2\ln U_1$ and $\theta = 2\pi U_2$. We omit a deep dive into the derivation for brevity, but there are great summaries online.



There is an alternate form of the Box-Muller transform in polar form which avoids the use of sin and cos and uses rejection sampling. This version samples two uniform random variables u and v to find a radius $s = R^2 = u^2 + v^2$. We reject s if it falls outside of the unit circle, and keep it if it does not. We can then

use this sampled value of s in a similar way to the above.

All of these continuous-domain examples have natural analogs in the discrete case.

10.1 Variance Reduction

Suppose we want to estimate a value $\theta = \mathbb{E}[g(X_1, ..., X_n)]$. Such a value is sometimes intractably difficult to compute. Once again simulations come to the rescue. We can generate $(X_1^{(1)}, ..., X_n^{(1)})$ whose joint distribution is the same as $(X_1, ..., X_n)$. Then letting $Y_1 = g(X_1^{(1)}, ..., X_n^{(1)})$, $Y_2 = g(X_1^{(2)}, ..., X_n^{(2)})$, all the way to $Y_k = g(X_1^{(k)}, ..., X_n^{(k)})$ will give us a collection of random variables Y_i, each of which has the same distribution as $g(X_1, ..., X_n)$. Then taking a simple average can give us $\overline{Y} = \frac{1}{k}\sum Y_i = \theta$.

This general approach, the idea of taking repeated random sampling in order to obtain a result that may otherwise be computationally intractable, is known as a **Monte Carlo method**. The expected difference between $\overline{Y}$ and θ is $\mathbb{E}[(\overline{Y} - \theta)^2] = \text{Var}(\overline{Y})$; so it is reasonable that our goal is to get the variance of $\overline{Y}$ to be as low as possible.

Suppose we only generate two variables, Y_1 and Y_2. Then $\text{Var}(\overline{Y}) = \text{Var}\left(\frac{1}{2}(Y_1 + Y_2)\right) = \frac{1}{2}\text{Var}(Y_1) + \frac{1}{2}\text{Cov}(Y_1, Y_2)$. So if we want to decrease the variance, we might want $\text{Cov}(Y_1, Y_2)$ to be negative, so that Y_1 and Y_2 are negatively correlated versus being fully independent. We won't describe the generation technique here, but these negatively correlated variables are termed **antithetic variates**.

Recall the conditional variance formula $\text{Var}(X) = \mathbb{E}[\text{Var}(X|Y)] + \text{Var}(\mathbb{E}[X|Y])$. Rearranging terms tells us that $\text{Var}(\mathbb{E}[X|Y]) \leq \text{Var}(X)$. Once again, we can try to compute $\mathbb{E}[g(X_1, ..., X_n)$ by simulating $X = (X_1, ..., X_n)$ and computing $Y = g(X)$. From the above inequality this must mean $\text{Var}(\mathbb{E}[Y|Z]) \leq \text{Var}(Y)$. So if we can compute the conditional expectation $\mathbb{E}[Y|Z]$ we will get a lower variance estimator than calculating $\mathbb{E}[Y]$ directly.

Finally, suppose we have a case where we have some function f for which $\mathbb{E}[f(X)] = \mathbb{E}[f(X_1, ..., X_n)] = \mu$ is known. Then for some constant a the random variable

$$W = g(X) + a[f(X) - \mu]$$

is an estimator for $\mathbb{E}[g(x)]$. Using some calculus, we can minimize $\text{Var}(W)$ when

$$a = \frac{-\text{Cov}(f(X), g(X))}{\text{Var}(f(X))}.$$

The quantities here ($\text{Var}(f(X))$ and $\text{Cov}(f(X), g(X))$) are in general not possible to compute, so we can further estimate them with simulated data. This idea of **control variates** is the most broadly applicable and efficient method of variance reduction for Monte Carlo problems.