

18.600: Probability and Random Variables

Ace Chun

December 15, 2025

Contents

1	Combinatorics	4
1.1	Basic Counting Principles	4
1.2	Permutations	4
1.2.1	Cycles	5
1.2.2	Arrangements with permutations	5
1.3	Combinations	6
1.4	Multinomials	7
1.5	Pascal's Triangle	7
1.6	Binomial Theorem	8
2	Probability Basics	9
2.1	Sample Spaces and Events	9
2.1.1	DeMorgan's Laws	9
2.2	Axioms of Probability	10
2.2.1	Inclusion-Exclusion Principle	10
2.3	Equal Likelihood	11
2.3.1	Derangements	11
2.4	Conditional Probabilities	12
2.4.1	Independence	13
2.4.2	Generalized Multiplication Rule	14
2.4.3	Bayes' Theorem	14
2.5	Odds	14

3	Discrete Random Variables	15
3.1	Expectation	16
3.1.1	Functions on RVs	16
3.1.2	Linearity of Expectation	17
3.2	Variance	17
3.2.1	Covariance	18
3.3	Bernoulli RVs	18
3.4	Binomial RVs	18
3.5	Poisson RVs	19
3.5.1	Poisson Point Processes	20
3.6	Geometric RVs	21
3.7	Negative Binomial RVs	22
4	Continuous Random Variables	22
4.1	Definitions	23
4.1.1	Functions on RVs	24
4.2	Uniform Distributions	25
4.3	Normal Distributions	25
4.3.1	DeMoivre-Laplace Limit Theorem	26
4.4	Exponential Distributions	26
4.4.1	Memorylessness	27
4.4.2	Relation to Poisson Processes	27
4.4.3	Sequence of Exponential Events	27
4.5	Gamma Distributions	28
4.6	Cauchy Distributions	29
4.7	Beta Distributions	29
4.8	Moment Generating Functions	30
5	Multiple Random Variables	30
5.1	Joint Distributions	30
5.1.1	Discrete	30
5.1.2	Continuous	31
5.2	Independence	32
5.3	Sums of Independent Distributions	32
5.4	Conditional Distributions	33
5.4.1	Discrete	33
5.4.2	Continuous	34
5.4.3	Conditional Expectation	34

5.5	Covariance and Correlation	35
6	Limiting Behavior	35
6.1	Markov's Inequality	35
6.2	Chebyshev's Inequality	36
6.3	Law of Large Numbers	36
6.3.1	Weak Law	36
6.3.2	Strong Law	37
6.4	Central Limit Theorem	37
6.5	Markov Chains	38
7	Entropy	40
7.1	Noiseless Coding	41
8	Martingales	42
8.0.1	Optional Stopping Theorem	43
8.1	Finance	44
8.1.1	Risk-Neutral Probability	44
8.1.2	Call Function	44
8.1.3	Black-Scholes Model	45

1 Combinatorics

Combinatorics, fundamentally, deals with counting the number of ways arrangements or events can occur within a given environment and set of conditions. This is integral to then computing *probabilities* of specific arrangements out of larger sets of possible outcomes.

1.1 Basic Counting Principles

There are two basic principles for counting outcomes of events or arrangements.

Multiplication

Given two (independent) separate sets A and B , if there are x ways to choose one element from A and y ways to choose one element from B , then there are $x \cdot y$ ways to choose a pair of elements, with one from A and the other from B .

This can also be framed in terms of events. If one event can occur in x ways and a second event following the first can occur in y ways, then there are $x \cdot y$ total possible outcomes (or ways that the events can occur). For some experiment A with x possible outcomes whose occurrences are all independent from each other, running A k times will result in x^k total outcomes.

Addition

Given two separate (disjoint) sets A and B , if there are x ways to choose one element from A and y ways to choose one element from B , then there are $x + y$ ways to choose an element from either A or B .

1.2 Permutations

A permutation describes a particular arrangement of objects that pays attention to *order*. Formally, a permutation is a bijective function σ that takes a set S and maps it back to itself.

$$\sigma : S \rightarrow S$$

In particular, if S is the set of the first n integers ($S = \{1, 2, \dots, n\}$), then σ maps a particular integer j to another integer in the set (which may be j itself — if this is the case, j is called a *fixed point*).

Permutations can be composed. If σ and ρ are both permutations, then their composition is written as

$$\sigma \circ \rho(j) = \sigma(\rho(j))$$

1.2.1 Cycles

Permutations may be decomposed into cycles, which is some subset R of the whole domain S in which σ maps each element of R back to elements of R — that is, σ on R is entirely self-contained.

For a length- k cycle, applying σ to it k times to some member of the cycle j will produce j again.

For example, suppose we define a permutation σ that maps the first 7 positive integers back to itself. If σ has the cycles

$$(2, 3, 5), (1, 7), (4, 6)$$

then

$$\sigma(2) = 3, \sigma(3) = 5, \sigma(5) = 2$$

$$\sigma(1) = 7, \sigma(7) = 1$$

$$\sigma(4) = 6, \sigma(6) = 4$$

σ is called an *involution* if $\sigma(\sigma(x)) = x$ for all x — that is, all cycles of σ are at most length 2. A fixed point is a cycle of length 1; if a permutation has no cycles of length 1, then it is said to be *fixed point free*.

1.2.2 Arrangements with permutations

In general, for a set of n elements, there are $n!$ permutations that can be produced. To see why, we can envision counting the number of permutations as lining the n elements up in a row.

There are n choices for elements to be placed in the first spot. Once the first spot is chosen, there are $n - 1$ choices for the second spot, $n - 2$ choices for the third spot, and so on. By the multiplication principle, there are

$$n \cdot (n - 1) \cdot (n - 2) \cdots 1 = n!$$

total ways to line up the n elements — there are $n!$ possible permutations.

Now, suppose we want to permute only a subset of size $k < n$ elements from our set. Following the same logic, there are

$$\underbrace{n \cdot (n-1) \cdots (n-k+1)}_{k \text{ terms}} = \frac{n!}{(n-k)!}$$

ways to create such an arrangement. Another, “subtractive” way to think about this is to consider that there are $n!$ total ways to permute the set of n elements. There are $(n-k)$ elements that we effectively do not care about with regards to their ordering, and yet the differences every single possible ordering of these $(n-k)$ elements (of which there are $(n-k)!$) is accounted for in the total number of permutations. Hence, we need to divide out the total number of permutations by the number of arrangements we choose not to differentiate between — this is $\frac{n!}{(n-k)!}$.

1.3 Combinations

Now, suppose we simply want to choose k elements from a set of n total elements, without caring about their ordering. In this case, we no longer care about differentiating between orderings of both the k chosen elements and the $(n-k)$ non-chosen elements, so we need to divide out the total number of permutations by both quantities.

The number of ways, then, to choose k elements is

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

Note the symmetry here:

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

Choosing k elements from a set without caring about order is equivalent to choosing $n-k$ elements to *exclude*.

$\binom{n}{k}$ is called a *binomial coefficient*.

1.4 Multinomials

We can generalize the binomial. Instead of choosing two partitions out of a set of n elements, suppose that we want to make r different partitions, with the i th partition consisting of n_i elements such that

$$n_1 + n_2 + \cdots + n_r = n$$

Following the same logic, there are

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

ways to make these divisions. This is notated with the multinomial coefficient:

$$\binom{n}{n_1, n_2, \cdots, n_r}$$

1.5 Pascal's Triangle

Pascal's triangle is a triangular array with rows for $n = 0, 1, 2, \cdots$, where the k th element of the n th row satisfies the recurrence relation

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

that is, each element of the triangle is found by summing the two elements directly/diagonally above it to the left and right.

$$\begin{array}{ccccccc} & & & & 1 & & \\ & & & & & & \\ & & & 1 & & 1 & \\ & & & & & & \\ & & 1 & & 2 & & 1 \\ & & & & & & \\ & 1 & & 3 & & 3 & & 1 \\ & & & & & & \\ & & & & \vdots & & \end{array}$$

Pascal's triangle can be formulated in terms of the binomial coefficients.

$$\begin{array}{ccccccc}
& & & \binom{0}{0} & & & \\
& & & & & & \\
& & \binom{1}{0} & & \binom{1}{1} & & \\
& & & & & & \\
& \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & \\
& & & & & & \\
\binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} \\
& & & & & & \\
& & & \vdots & & &
\end{array}$$

Observe that summing up all of the entries to the n th row of the triangle yields 2^n :

$$\sum_{i=0}^n \binom{n}{i} = 2^n$$

1.6 Binomial Theorem

For the binomial expansion of $(x + y)^n$, we can find the coefficients in front of each term by using the binomial coefficients. More specifically,

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

Intuitively, we can think about this as *laying out* all n binomials $(x + y)$, from which we will choose k to “take” x from and the remaining $(n - k)$ to “take” y from. This is known as the Binomial theorem.

We can use this to prove the fact that the n th row of Pascal’s triangle adds up to 2^n . The sum of the entries of the n th row of Pascal’s triangle can be formulated as

$$\sum_{k=0}^n \binom{n}{k} = \sum_{k=0}^n \binom{n}{k} 1^k 1^{n-k} = (1 + 1)^n = 2^n$$

We can generalize this to trinomial (or n -nomial) expansions by using multinomials.

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{n_1, \dots, n_r} \binom{n}{n_1, \dots, n_r} x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}$$

for all partitions n_1, \dots, n_r such that

$$n_1 + n_2 + \cdots + n_r = n$$

2 Probability Basics

2.1 Sample Spaces and Events

Suppose that we run some experiment X . We define the *sample space* S of X to be the set of all possible outcomes of this experiment. For example, the sample space of tossing a coin n times is the set

$$\{H, T\}^n$$

Any subset $E \subset S$ is called an *event* — an event is a set containing possible outcomes of an experiment.

Two subsets (or, events) A and B can be operated on accordingly:

1. $A \cup B$, or the union of A and B , refers to the set of elements that are contained in either A or B (or both).
2. $A \cap B$, or the intersection of A and B , refers to the set of elements that are contained in both A and B .
3. $A^C = \bar{A}$, or the complement of A , refers to the set of elements in S that are not in A .
4. $A \setminus B$, or A minus B , refers to the set of points in A that are not in B — we can also phrase this as $A \cap B^C$.

2.1.1 DeMorgan's Laws

DeMorgan's Laws are logical axioms that can be applied to sets.

$$(A \cup B)^C \iff A^C \cap B^C$$

$$(A \cap B)^C \iff A^C \cup B^C$$

More generally, for n events E_i ,

$$\left(\bigcup_{i=1}^n E_i \right)^C \iff \bigcap_{i=1}^n (E_i)^C$$

$$\left(\bigcap_{i=1}^n E_i \right)^C \iff \bigcup_{i=1}^n (E_i)^C$$

2.2 Axioms of Probability

We can define the probability of some event occurring (or, some subset of the sample space being the outcome) in terms of its relative frequency in comparison to the rest of the sample space. The probability of some event A occurring is denoted by the probability measure: $P(A)$.

We outline three axioms that should be satisfied by this probability measure.

1. $P(A) \in [0, 1], \forall A \subseteq S$
2. $P(S) = 1$
3. $P(\bigcup_{i=1}^{\infty} E_i)^C = \sum_{i=1}^{\infty} P(E_i)$ for a sequence of mutually exclusive events E_1, E_2, \dots .

A few properties fall out of these axioms. We take $P(AB)$ to mean $P(A \cap B)$. In particular,

1. $P(A^C) = 1 - P(A)$
2. If $A \subseteq B$, then $P(A) \leq P(B)$
3. $P(A \cup B) = P(A) + P(B) - P(AB)$.
4. $P(AB) \leq P(A)$

2.2.1 Inclusion-Exclusion Principle

Suppose we would like to compute the probability of the union of n events, E_1, E_2, \dots, E_n . Since we are given no other information about the relationships between each event (i.e. we don't know if they are mutually exclusive), computing this can be difficult. Instead, we can reframe the probability of the union of the events occurring in terms of the probability of the intersection of some of the events.

$$\begin{aligned}
 P\left(\bigcup_{i=1}^n E_i\right) &= \sum_i P(E_i) - \sum_{i_1 < i_2} P(E_{i_1} E_{i_2}) + \dots \\
 &\quad + (-1)^{r+1} \sum_{i_1 < \dots < i_r} P(E_{i_1} E_{i_2} \dots E_{i_r}) + \dots \\
 &\quad + (-1)^{n+1} P(E_1 E_2 \dots E_n)
 \end{aligned}$$

In other words, we first compute the sum of the probabilities of all individual events, and then we subtract to compensate for any extraneous “overlap” that we’ve double counted. At this point, we will have overcompensated and removed all of the triple-intersections, so we need to add those back — this continues until we hit the intersection of all n events.

To convince ourselves of why this is the case, we can consider a venn diagram of the n events. Now, consider a region of the venn diagram that is contained in exactly $m > 0$ of the events. There are $\binom{m}{1}$ single-intersections that cover this region, $\binom{m}{2}$ double-intersections, $\binom{m}{3}$ triple-intersections, and so on. Using the principle, this region would be counted

$$\begin{aligned} \binom{m}{1} - \binom{m}{2} + \binom{m}{3} + \cdots \pm \binom{m}{m} &= \binom{m}{0} - \sum_{i=0}^m \binom{m}{i} \\ &= 1 - \sum_{i=0}^m \binom{m}{i} (-1)^i \binom{m}{i} \\ &= 1 - (1 - 1)^m = 1 \end{aligned}$$

so, by the principle, each region of the venn diagram would be counted exactly once, which allows us to sum over all such regions to find the total intersection.

For $n = 2$, this works out to

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

2.3 Equal Likelihood

If a sample space has n elements with equal probabilities/likelihoods, then the probability of an event containing one of the elements must be $\frac{1}{n}$. For a general subset $A \subseteq S$, once again, with equal probabilities,

$$P(A) = \frac{|A|}{|S|}$$

2.3.1 Derangements

Suppose we shuffle an array of n elements. An arrangement that occurs such that none of the n elements end up in the position they originally started

is known as a *derangement*. We can count the probability of a derangement occurring (we call this event A) in the following manner.

First, consider the event E_i : the i th element ends up in its original spot. Since we are counting derangements, then, is the complementary probability that each E_i occurs: that is,

$$P(A) = 1 - P(E_1 \cup E_2 \cup \dots \cup E_n)$$

Now, we just need to count the union of each of the E_i s.

Consider the event that r elements end up back in their original spots: $P(E_{i_1} E_{i_2} \dots E_{i_r})$. In this event, there are $n - r$ “degrees of freedom” — that is, once we have placed the r elements in their original spots, there are $n - r$ elements left over for which we can decide an arrangement. Therefore, there are $(n - r)!$ permutations for which this event occurs, so

$$P(E_{i_1} E_{i_2} \dots E_{i_r}) = \frac{(n - r)!}{n!}$$

There are $\binom{n}{r}$ such groups, meaning that there are

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

such terms in the inclusion-exclusion sum. Therefore,

$$P(E_1 \cup E_2 \cup \dots \cup E_n) = 1 - \frac{1}{2!} + \frac{1}{3!} - \dots \pm \frac{1}{n!}$$

Recalling our complementary counting, then,

$$P(A) = 1 - \left(1 - \frac{1}{2!} + \frac{1}{3!} - \dots \pm \frac{1}{n!}\right) = \frac{1}{2!} - \frac{1}{3!} + \dots \pm \frac{1}{n!} \approx \frac{1}{e}$$

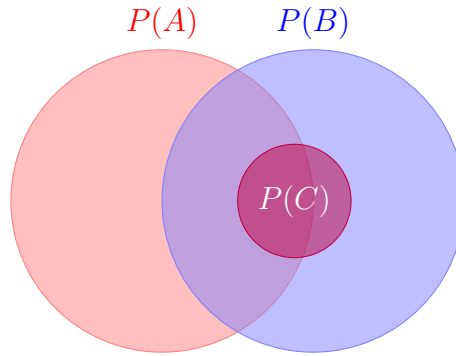
2.4 Conditional Probabilities

Suppose that I am conducting a series of experiments or trials, and I want to count the probability of some event A happening in a larger sample space S . However, I learn some *new* information about my experimental setup; namely, some other event, B , has occurred for certain, which effectively *shrinks* my sample space to the space of outcomes in which B happens. Knowing

that B happened, then, changes my knowledge of the probability of A ; we call this new probability $P(A|B)$, or the probability of A conditioned on B .

$$P(A|B) = \frac{P(AB)}{P(B)}$$

The above equation can be interpreted as taking the likelihood that A and B both happen *overall*, and then normalizing this by the probability that B happens. We're effectively limiting our possible sample space given what we have learned.



Note that

$$P(A) = P(AB) + P(AB^C) = P(A|B)P(B) + P(A|B^C)P(B^C)$$

2.4.1 Independence

Two events A and B are independent if knowing that one of them happened does not change what I know about the other:

$$P(A) = P(A|B)$$

A set of events E_1, E_2, \dots, E_n are *pairwise independent* if, for any pair of distinct indices (i, j) with $i \neq j$,

$$P(E_i) = P(E_i|E_j), \quad P(E_j) = P(E_j|E_i), \quad P(E_i E_j) = P(E_i)P(E_j)$$

E_1, E_2, \dots, E_n are *mutually independent* if

$$P(E_1 E_2 \dots E_n) = P(E_1)P(E_2) \dots P(E_n)$$

Note that mutual independence is a stronger statement than pairwise independence. Mutual independence implies pairwise independence, but not the other way around.

2.4.2 Generalized Multiplication Rule

In general, for events E_1, E_2, \dots, E_n , we cannot simply apply the multiplication rule without knowing that each of the E_i 's are independent from the rest of the events. However, in general, it is the case that

$$P(E_1 E_2 \cdots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2) \cdots P(E_n|E_1 \cdots E_{n-1})$$

This framing is useful when we think about an overarching experiment that is divided into multiple stages, each with a different tree of outcomes.

2.4.3 Bayes' Theorem

From the definition of conditional probability, we can derive the following equivalent statement:

$$P(AB) = P(A|B) \cdot P(B)$$

However,

$$P(AB) = P(BA) = P(B|A) \cdot P(A)$$

Substituting this back into the initial definition for $P(A|B)$, we see

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The above statement is *Bayes' Theorem*. Intuitively, Bayes' theorem tells you how to *update* your estimate of the probability of A happening when you learn new information about your experiment (i.e., that B happened). $P(A)$ is called a prior probability, and $P(A|B)$ is your posterior estimate.

2.5 Odds

The odds that some event A happens is a comparison of the probability of favorable outcomes to the probability of non-favorable outcomes. More formally, this can be written as the ratio

$$\frac{P(A)}{P(A^C)}$$

For A conditioned on B , we write the conditional odds of A given B to be

$$\frac{P(A|B)}{P(A^C|B)}$$

We can describe the ratio between these two odds:

$$\frac{P(A|B)/P(A^C|B)}{P(A)/P(A^C)} = \frac{P(B|A)}{P(B|A^C)}$$

In effect, this lets us do a similar “estimate update” on odds, rather than probabilities.

3 Discrete Random Variables

In general, a *random variable* (RV) X is a real-valued function on the state space.

$$X : S \rightarrow \mathbb{R}$$

We can interpret X as some sort of quantity whose value depends on the outcome of some experiment with sample space S .

For example, we can conceive of a random variable X on the outcome of n coin tosses that counts the number of times the coin lands on heads. For a given number of heads k ,

$$P\{X = k\} = \frac{1}{2^n} \binom{n}{k}$$

An *indicator random variable* functions like a boolean — it takes the value 1 if some event E happens, and 0 if not. They indicate whether or not some event occurred.

$$1_E = \begin{cases} 1 & \text{if } E \text{ occurs} \\ 0 & \text{if } E \text{ does not occur} \end{cases}$$

If E_1, E_2, \dots, E_k are events, then

$$\sum_{i=1}^k 1_{E_i}$$

counts the number of events that occurred.

Indicator random variables are helpful because they often allow us to decompose a larger problem or event into more manageable pieces.

A RV X is called a *discrete random variable* if it takes one out of a countable set of values. For each a in this countable set of values, we can write

$$p_X(a) := P\{X = a\}$$

We call $p_X(a)$ the *probability mass function*. The *cumulative distribution function* is defined as follows:

$$F(a) := P\{X \leq a\} = \sum_{x \leq a} p_X(x)$$

3.1 Expectation

The expectation of X is defined as

$$E[X] = \sum_{\{x | p_X(x) > 0\}} x \cdot p_X(x)$$

We can interpret this as a weighted average of all possible values that could be taken by X , each value weighted by the probability that X takes that value.

Alternatively, if the state space S itself is countable, we can reframe this sum in terms of the elements of the state space itself.

$$E[X] = \sum_{s \in S} P(\{s\})X(s)$$

This last definition requires a technical caveat; if S is countable but infinite, then the expectation may differ depending on how the elements of s are enumerated. However, we only have to watch out for this if each $P\{s\}X(s)$ forms a *conditionally convergent* series; if the series they form is unconditionally convergent, this definition still works. We must therefore have

$$\sum_{s \in S} P(\{s\})|X(s)| < \infty$$

3.1.1 Functions on RVs

We can compose random variables and functions on real numbers. For X , g , where X is a RV and $g : \mathbb{R} \rightarrow \mathbb{R}$, $g(X)$ is also a random variable, computed as the following:

$$E[g(X)] = \sum_{\{x | p_X(x) > 0\}} g(x) \cdot p_X(x)$$

or

$$E[X] = \sum_{s \in S} P(\{s\})g(X(s))$$

Note that the expectation of an indicator random variable is just the probability that the event it describes happens:

$$E[1_E] = P(E)$$

3.1.2 Linearity of Expectation

Expectation is a linear map — that is, it follows the principle of superposition on its arguments. For two random variables X and Y ,

$$E[aX + bY] = aE[X] + bE[Y]$$

In addition, $E[a] = a$ for any given constant a .

3.2 Variance

Variance quantifies how much a random variable deviates from its mean, or expectation. Let X be a random variable with mean $\mu = E[X]$. The variance of X is defined as

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

We can also express variance as

$$\text{Var}[X] = E[X^2] - E[X]^2$$

Often, this second formula will be much more convenient to compute.

We have the following properties:

$$\begin{aligned}\text{Var}[X + b] &= \text{Var}[X] \\ \text{Var}[aX] &= a^2 \text{Var}[X]\end{aligned}$$

In addition, we can define the standard deviation σ , which provides a measure of spread *in the same units as X* .

$$\text{SD}[X] = \sigma = \sqrt{\text{Var}[X]}$$

3.2.1 Covariance

The covariance between two RVs X and Y is a measure how X varies with Y . It is defined as

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y]$$

Note that $\text{Cov}[X, X] = \text{Var}[X]$.

We define the *correlation* between X and Y as

$$\frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

The correlation always lies between -1 and 1 . If the correlation has a magnitude close to 1 , this means that X is very highly correlated with Y , and vice-versa; a lower magnitude for correlation implies that X doesn't really have much to do with Y . The sign of the correlation indicates the direction of the correlation (i.e. if X is larger, Y tends to be larger if they are positively correlated, and the opposite if they are negatively correlated).

3.3 Bernoulli RVs

A Bernoulli trial describes a particular experiment as one of two outcomes: success, or failure. A Bernoulli random variable describes the probability that a Bernoulli trial will be successful (with probability p) or will fail (with probability $q = 1 - p$). The probability mass function of a Bernoulli random variable X takes the following form:

$$\begin{aligned} P(X = 1) &= p \\ P(X = 0) &= 1 - p \end{aligned}$$

3.4 Binomial RVs

Binomial RVs express the outcomes of repeated (independent) Bernoulli trials. In particular, binomial random variables are parameterized by (n, p) , where n is the number of trials taking place, and p is the probability that each trial will result in a successful outcome, and they count the number of successful outcomes that will occur within the n trials.

If X is a binomial random variable with parameters (n, p) , then

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

The expectation of X is np :

$$E[X] = np$$

Intuitively, this makes sense, as we can interpret p as the fraction of trials that will be successful in a given set. We can also calculate the variance of X :

$$\text{Var}[X] = npq$$

In addition, the expectation of raising X to the k th power can be expressed as the following:

$$E[X^k] = npE[(Y + 1)^{k-1}]$$

where Y is another binomial random variable with parameters $(n - 1, p)$.

3.5 Poisson RVs

If X is a Poisson RV with parameter λ , its probability mass function takes the form

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

for some $k \in \mathbb{N}$.

A Poisson RV can be thought of as the limit of a Binomial RV as n grows large in comparison to k and $\lambda = np$, the mean of the distribution, is fixed. In other words, Poisson RVs describe the outcome of a very large number of total trials, each with a small enough probability such that the *expected number* of positive outcomes is relatively “moderate” (i.e., the number of accidents that will happen over 10 years with a vehicle with a low accident rate).

We can observe what happens in the limit $n \gg k$ in the following deriva-

tion:

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

Notice that, if we take the sum of each of these $p_X(k)$ s for all k , we see

$$e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1$$

If X is a Poisson RV with parameter λ , then

$$\begin{aligned}
E[X] &= \lambda \\
\text{Var}[X] &= \lambda
\end{aligned}$$

Something important to note is that the formulation of Poisson RVs relies on the assumption that each individual event/trial is independent, which may or may not be the case.

3.5.1 Poisson Point Processes

A *Poisson process* describes independent events that occur over some time interval t . For example, suppose that we are interested in the number of plane crashes that occur in t years, or the number of p -coins that end up as heads in the first t flips. We encode this information in a random variable parameterized by time, $N(t)$, which describes the number of events that occur during the first t units of time (relative to wherever your “start point” is).

Poisson processes must satisfy the following axioms:

1. $N(0) = 0$
2. Independence: the number of events that occur within disjoint time intervals are independent. That is, for disjoint time intervals $[t_1, s_1]$ and $[t_2, s_2]$, $N(s_1) - N(t_1)$ is independent from $N(s_2) - N(t_2)$.

3. Homogeneity: the distribution of the number of events that occur in some interval depends only on its length, and not its location. We don't expect the number of events occurring in the interval $[t, t + 5]$ to differ from $N(5)$. $E[N(h)] = \lambda h$.
4. Non-concurrence: the probability that at least two events occur within the first h time steps is much less than the probability that a single event occurs within that interval when h is small. More specifically:
 - (a) $P(N(h) = 1) = \lambda h + o(h)$
 - (b) $P(N(h) \geq 2) = o(h)$

A process that obeys the above axioms is called a Poisson Point Process with rate λ .

The non-concurrence axioms can be used to derive the probability of no event occurring within the first t time steps:

$$P(N(t) = 0) = e^{-\lambda t}$$

Let the random variable T_1 indicate the time of the first event. Equivalently, then,

$$P(T_1 \geq t) = e^{-\lambda t}$$

T_1 is called an *exponential random variable* with rate λ .

We can denote subsequent intervals: for example, T_2 is the time between the first and second event, T_3 is the time between the third and second event, etc: T_k is the time between the k th and $(k - 1)$ th event. We know that each T_1, T_2, \dots are independent, based on our axioms, so each T_i is an exponential random variable with rate λ .

We can use this information and the binomial theorem to derive the probability that k events happen within an interval of length t :

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

3.6 Geometric RVs

Consider a series of Bernoulli trials, each with a probability p of success. Let X be a random variable that describes how long it takes for a successful trial

to occur in a sequence of trials — i.e., if $X = k$, the first success occurred on the k th trial. X is a *geometric random variable* with parameter p .

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

as we must have had $k - 1$ failures before finding a success on the k th trial. The expected value of X is

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

and its variance is

$$\text{Var}[X] = \frac{1 - p}{p^2} = \frac{q}{p^2}$$

3.7 Negative Binomial RVs

For a sequence of Bernoulli trials, let X be the random variable that indicates that the r th success occurs on the X th trial. X is a negative binomial RV with parameters (r, p) .

To figure out the probability that X is a particular k , we consider that there must have been $r - 1$ successes in the $k - 1$ preceding trials. This lends us the structure of a Binomial RV. Therefore,

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

The expectation of X is

$$E[X] = \frac{r}{p}$$

which, intuitively, corresponds to taking a geometric variable r times.

$$\text{Var}[X] = \frac{rq}{p^2}$$

4 Continuous Random Variables

Where discrete random variables could map to a set of finite or countably infinite values, continuous random variables can take on an uncountable set of values.

4.1 Definitions

X is a *continuous* random variable if there exists a function f_X (called a *probability density function*, or *pdf*) on \mathbb{R} such that $P(X \in B) = \int_B f_X(x)dx$ for $B \subseteq \mathbb{R}$. The pdf is the continuous analogue to the probability mass function of a discrete random variable.

By the axioms of probability, f_X must satisfy

$$P(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f_X(x)dx = 1$$

In addition, f_X should be nonnegative everywhere, as it does not make sense for probabilities to be negative/subtractive. If B is a contiguous interval $[a, b]$,

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

In a continuous distribution, the probability that X is any singular *point* is 0. This can be shown by setting the relevant interval to $[a, a]$:

$$P(X = a) = \int_a^a f_X(x)dx = 0$$

This is in contrast to discrete random variables, where it is possible that the probability that the random variable takes on a specific value is non-zero. This does not hold in the continuous case because f_X describes probabilities on *infinitesimally small* interval (i.e. $[x, x + \Delta x]$).

$$P(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(x)dx \approx f_X(x) \cdot \Delta x$$

However, we can define the *cumulative distribution function* (cdf) of X , exactly analogous to the discrete case:

$$F_X(a) = P(X \leq a) = \int_{-\infty}^a f_X(x)dx$$

Note that

$$\frac{d}{da} F_X(a) = f_X(a)$$

The expected value of X is defined to be

$$E[X] = \mu_X = \int_{-\infty}^{\infty} x f(x)dx$$

Again, this is analogous to the discrete case, in which we took a summation — here, we've simply replaced the sum with an integral and the probability mass function $p(x)$ with $f(x)dx$. Similarly, the variance is defined as

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

In addition,

$$E[X] = \int_0^\infty P(X > y)dy = \int_0^\infty \int_y^\infty f_X(x)dx dy$$

The linearity of expectation is maintained.

$$E[aX + b] = aE[X] + b$$

and for variance,

$$\text{Var}[aX + b] = a^2\text{Var}[X]$$

4.1.1 Functions on RVs

Let $Y = g(X)$ be a random variable and a function on X , for which a pdf f_X and cdf F_X are known. To determine the relevant functions for Y , we first note that

$$F_Y(a) = P(Y \leq a) = P(g(X) \leq a) = P(X \leq g^{-1}(a))$$

Using what we know about X , we know that this quantity is

$$P(X \leq g^{-1}(a)) = F_X(g^{-1}(a)) = F_Y(a)$$

We can take the derivative of both sides to find f_Y :

$$\frac{d}{da}F_Y(a) = \frac{d}{da}F_X(g^{-1}(a)) = f_X(g^{-1}(a)) \cdot \left(\frac{d}{da}g^{-1}(a)\right)$$

We can find the expectation of Y by computing the integral

$$E[Y] = \int_{-\infty}^{\infty} g(X)f_X(x)dx$$

Finding the distribution functions for a scaled/shifted random variable is fairly easy, using the chain rule. Suppose we have some random variable X

with a pdf $f_X(x)$. The distribution of $Y = aX + b$ for constants a, b , can be found:

$$F_Y(y) = P(Y \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Taking the derivative,

$$f_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right)$$

4.2 Uniform Distributions

A random variable is uniformly distributed over the interval $[\alpha, \beta]$ if it is described by the pdf

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

The corresponding cdf takes the form

$$F(x) = \begin{cases} 0 & x < \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \alpha \leq x \leq \beta \\ 1 & x > \beta \end{cases}$$

Its expected value is

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_{\alpha}^{\beta} x \cdot \frac{1}{\beta-\alpha} dx = \frac{\beta + \alpha}{2}$$

Its variance is

$$\text{Var}[X] = \frac{(\beta - \alpha)^2}{12}$$

4.3 Normal Distributions

A random variable is *normally distributed* with parameters μ and σ^2 if its pdf is described as

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

Its mean is μ , and its variance is σ^2 .

If $X \sim \text{Normal}(\mu, \sigma^2)$, then $Y = aX + b \sim \text{Normal}(a\mu + b, a^2\sigma^2)$.

The integral of $f(x)$ has no closed form, though it integrates to 1 over its domain (this can be shown by converting the integral to polar). Instead, we refer to the cdf of a normal distribution with $\mu = 0$ and $\sigma^2 = 1$ as $\Phi(x)$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

such that $\Phi(a)$ refers to the probability that a sample from $X \sim \text{Normal}(0, 1)$ falls below a . In general, then, for $Y \sim (\mu, \sigma^2)$,

$$P(Y < a) = P(\sigma X + \mu < a) = P\left(X < \frac{a - \mu}{\sigma}\right) = \Phi\left(\frac{a - \mu}{\sigma}\right)$$

The quantity $(a - \mu)/\sigma$ is referred to as the z -score of a with respect to the distribution Y .

4.3.1 DeMoivre-Laplace Limit Theorem

The DeMoivre-Laplace limit theorem states that the distribution of a binomial random variable parameterized by (n, p) will approach a normally distributed random variable with mean np and variance $npq = np(1 - q)$.

4.4 Exponential Distributions

An exponential random variable with parameter $\lambda > 0$ is described by the pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

with

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

The expectation of an exponential random variable is $\frac{1}{\lambda}$, and its variance is $\frac{1}{\lambda^2}$.

Intuitively, the exponential distribution can be conceptualized as the distribution describing the amount of time *until* some event occurs.

4.4.1 Memorylessness

A random variable X is memoryless if

$$P(X > s + t | X > t) = P(X > s)$$

that is, given that some event has already occurred, the time until an identical event occurs does not depend on whether or not the event has already occurred — each future event is independent of what has already happened.

$$P(X > s + t) = P(X > s)P(X > t)$$

Exponential random variables are memoryless; if X is an exponential RV with parameter λ ,

$$P(X > s + t) = 1 - F_X(s + t) = e^{-\lambda(s+t)} = e^{-\lambda s} e^{-\lambda t} = P(X > s)P(X > t)$$

4.4.2 Relation to Poisson Processes

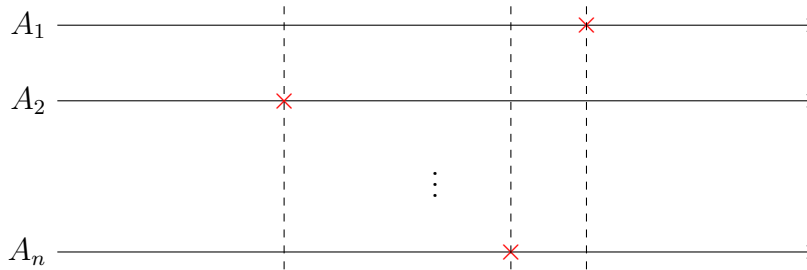
Previously, we had thought about Poisson processes in terms of the expectation of the number of events that would occur in an *given interval of length* T . However, exponential RVs give us a different framing: we think about the process in terms of the *points in time at which events occur*. If events occur at times according to an exponential RV, the events occur according to a Poisson process.

4.4.3 Sequence of Exponential Events

Let $A_1, A_2, A_3, \dots, A_n$ be independent and identically distributed exponential random variables with parameter λ . The distribution of $A = \min_i \{A_i\}$ is exponential with parameter $n\lambda$, which can be seen by observing that

$$P(A \geq a) = P(A_1 \geq a)P(A_2 \geq a) \cdots P(A_n \geq a) = (e^{-\lambda a})^n = 1 - F_A(a)$$

Suppose we would like to calculate the expected time until the i th event out of $A_1 \cdots A_n$ occurs. We have the following picture:



The time until the first event is the minimum of all n events, so it is exponential with rate $n\lambda$. The time from the first event until the second event is the minimum of the subsequent $n - 1$ events, so it is exponential with rate $(n - 1)\lambda$ by memorylessness. Recalling that the expectation of an exponential random variable is $\frac{1}{\lambda}$, then, we find that the time until the i th of the n variables “goes off” is

$$\frac{1}{n\lambda} + \frac{1}{(n-1)\lambda} + \cdots + \frac{1}{(n-i+1)\lambda}$$

and the expected time until all n events happens is

$$\frac{1}{\lambda} \sum_{i=1}^n \frac{1}{i}$$

4.5 Gamma Distributions

A random variable has a gamma distribution with parameters (α, λ) for $\alpha > 0$, $\lambda > 0$ if its pdf is given by

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x} (\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} e^{-y} y^{\alpha-1} dy$$

is the *gamma function*, the analytic continuation of the factorial to the real numbers. For $n \in \mathbb{N}$,

$$\Gamma(n+1) = n!$$

The gamma distribution is, in some sense, the continuous analogue of the negative binomial random variable; it describes the amount of time one has to wait until $\alpha = n$ events have occurred, where each event can be thought of as an exponential random variable with rate parameter λ . The sum of n independent exponential random variables with the same parameter λ is a gamma distribution with parameters (n, λ) .

When $\lambda = \frac{1}{2}$ and $\alpha = \frac{n}{2}$ for a positive integer n , we get the χ^2 distribution.

The expectation of a gamma distribution is $\frac{\alpha}{\lambda}$, and its variance is $\frac{\alpha}{\lambda^2}$.

4.6 Cauchy Distributions

A Cauchy random variable with parameter θ is distributed according to

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

and its cumulative distribution function is

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan(x)$$

A Cauchy RV can be thought of as describing the intersection point of a beam from a flashlight, having been spun around on its center, located at $(0, 1)$.

The sum of two Cauchy random variables is also Cauchy, and the inverse of a Cauchy random variable is Cauchy as well.

Cauchy RVs do not have a well defined expectation or variance.

4.7 Beta Distributions

A random variable has a beta distribution with parameters (a, b) if its density is described by

$$f(x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1} (1-x)^{b-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $B(a, b)$ is a normalization constant given by

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Beta distributions describe updating beliefs about probabilities p for Bernoulli trials. For example, suppose that we have a p -coin (that is, a coin that flips to heads with probability p), but we are not given what p is, and we would like to describe the probability that p is some particular value in $[0, 1]$ based on the outcomes of coin flips (i.e. new evidence). We can say that, if $a - 1$ is the number of heads we have seen so far, and $b - 1$ is the number of tails, then the conditional probability distribution for p based on what we have seen is described by $\text{Beta}(a, b)$. For each coin flip outcome, we can increment either a or b and update our belief about the distribution of p .

The expectation of a Beta distribution is

$$\frac{a}{a+b}$$

4.8 Moment Generating Functions

The moment generating function (MGF) of a RV X is defined to be

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

The n th derivative of $M_X(t)$ with respect to t gives us

$$\frac{d^n}{dt^n} E[e^{tX}] = E[X^n e^{tX}]$$

Evaluating at $t = 0$, we see that

$$M_X^{(n)}(0) = E[X^n]$$

which is the n th moment of X — M_X effectively encodes information about the moments of X into its derivatives, and more specifically, its Taylor series centered around 0.

If X and Y are independent, the mgf of $Z = X + Y$ can be expressed in terms of M_X and M_Y :

$$M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$$

Additionally,

$$M_{aX}(t) = M_X(at)$$

$$M_{X+b}(t) = e^{bt} M_X(t)$$

A related function, called the *characteristic function*, is defined as

$$\phi_X(t) = E[e^{itX}]$$

Where M_X may not necessarily be well defined, ϕ_X is always guaranteed to be defined. Note that $\phi_X(t)$ is the Fourier transform of $f_X(t)$; it encodes some notion of periodicity patterns within the distribution function f_X .

5 Multiple Random Variables

5.1 Joint Distributions

5.1.1 Discrete

Suppose X and Y (not necessarily independent) take values in $\{1, 2, \dots, n\}$, we can define a joint probability mass function

$$A_{i,j} = P(X = i, Y = j)$$

We can view the $A_{i,j}$'s as the entries of an $n \times n$ matrix.

Suppose I just want to know $P(X = i)$ at large, without caring about Y . We can sum over every possible value Y can take on:

$$P(X = i) = \sum_{j=1}^n A_{i,j}$$

And similarly,

$$P(Y = j) = \sum_{i=1}^n A_{i,j}$$

The above are called the *marginal* distributions of X and Y .

5.1.2 Continuous

Given continuous random variables X and Y , we can define their joint cumulative distribution function as

$$F(a, b) = P(X \leq a, Y \leq b)$$

From this, we may also define marginal cumulative distribution functions:

$$F_X(a) = P(X \leq a) = \lim_{b \rightarrow \infty} F(a, b)$$

$$F_Y(b) = P(Y \leq b) = \lim_{a \rightarrow \infty} F(a, b)$$

We may also develop a joint probability density function f , which would need to satisfy

$$P((X, Y) \in A) = \int_A f(x, y) dx dy$$

for $A \subseteq \mathbb{R}^2$. It turns that

$$f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y)$$

fulfills this requirement (and is also the obvious/intuitive choice, given the relationship between the cdf and pdf in the single-variable case). We can also find their marginal pdfs:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

In the above operations, we essentially integrate out the variable we do not want, similar to how we sum over the entire range in the discrete case.

5.2 Independence

Two random variables X and Y are independent if, for any two measurable sets A and B ,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Knowing anything about X doesn't tell me about Y , and vice versa.

If X and Y are discrete, we can say that they are independent if

$$\forall x \in S_X, y \in S_Y, P(X = x, Y = y) = P(X = x)P(Y = y)$$

where S_X and S_Y are the sample spaces of X and Y respectively.

If X and Y are continuous, they are independent if

$$f(x, y) = f_X(x)f_Y(y)$$

5.3 Sums of Independent Distributions

In many cases, we will be interested in random variables expressed as sums of independent distributions. Suppose $Z = X + Y$, for independent random variables X, Y . Then,

$$\begin{aligned} F_Z(a) &= P(Z \leq a) = P(X + Y \leq a) \\ &= \iint_{x+y \leq a} f_X(x)f_Y(y)dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dx dy \\ &= \int_{-\infty}^{\infty} f_Y(y) \left[\int_{-\infty}^{a-y} f_X(x)dx \right] dy \\ &= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy \end{aligned}$$

Then,

$$f_Z(a) = \frac{d}{da}F_Z(a) = \int_{-\infty}^{\infty} f_X(a - y)f_Y(y)dy$$

The above operation is called the *convolution* of f_X and f_Y .

$$(f_X \star f_Y)(a) = \int_{-\infty}^{\infty} f_X(x)f_Y(a-x)dx = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$$

If X and Y are independent and uniformly distributed on $[0, 1]$, then the probability density of $Z = X + Y$ is

$$f_Z(a) = \begin{cases} a & 0 \leq a \leq 1 \\ 2 - a & 1 < a < 2 \\ 0 & \text{otherwise} \end{cases}$$

If X and Y are independent and normally distributed with parameters (μ_X, σ_X^2) and (μ_Y, σ_Y^2) respectively, then $X + Y = Z$ will also be normally distributed, with parameters $(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

If X and Y are Poisson with parameters λ_X, λ_Y , then $X + Z$ will also be Poisson with parameter $\lambda_X + \lambda_Y$.

5.4 Conditional Distributions

5.4.1 Discrete

For any two discrete events X and Y , recall that the conditional probability of X given Y is

$$P(X|Y) = \frac{P(XY)}{P(Y)}$$

We can, in turn, define the conditional probability mass function of X given $Y = y$:

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

If X is independent of Y , then

$$p_{X|Y}(x|y) = P(X = x)$$

as expected. We can also define a cumulative distribution function:

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

5.4.2 Continuous

We can define the conditional probability density of X given $Y = y$ as

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

and similarly,

$$F_{X|Y}(a|y) = \int_{-\infty}^a f_{X|Y}(a, y) dx$$

5.4.3 Conditional Expectation

The value $E[X|Y = y]$ is just the weighted average over the possible values X could take, given that we have restricted the sample space to the event that $Y = y$.

$$E[X|Y = y] = \sum_x xP(X = x|Y = y) = \sum_x xp_{X|Y}(x|y)$$

In a continuous context,

$$E[X|Y = y] = \int_{-\infty}^{\infty} x \frac{f(x, y)}{f_Y(y)} dx$$

From this, we can conceptualize $E[X|Y]$ as a *random variable* itself, depending on the value that Y takes on.

$$E[E[X|Y]] = E[X]$$

Conditional variance works the same.

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

It can be shown that

$$\text{Var}[X] = \text{Var}[E[X|Y]] + E[\text{Var}[X|Y]]$$

5.5 Covariance and Correlation

The covariance of RVs X with mean μ_X and Y with mean μ_Y is defined to be

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

This is also equivalent to the expression

$$E[XY] - E[X]E[Y]$$

Note that $\text{Cov}(X, X) = \text{Var}[X]$. The covariance is, in some sense, a measure of how tied the values of X and Y are to each other. If the magnitude of the covariance is large, that means that as X changes, Y tends to change as well; if the covariance is zero, changes in X have no bearing on the changes in Y .

Covariance is bilinear, so

$$\text{Cov}(aX + bY, Z) = a\text{Cov}(X, Z) + b\text{Cov}(Y, Z)$$

The correlation $\rho(X, Y)$ is defined as follows:

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

ρ effectively normalizes the covariance into a unitless quantity. ρ will always lie between -1 and 1.

6 Limiting Behavior

6.1 Markov's Inequality

Let X be a random variable that takes only non-negative values. For a given constant $a > 0$, Markov's inequality states

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

Intuitively, this indicates that if X is expected to be small, then it is very unlikely that X will stray far from its mean (i.e., take on a large value).

Proof

Let Y be a random variable based on X , such that

$$Y = \begin{cases} a & X \geq a \\ 0 & X < a \end{cases}$$

$X \geq Y$ with probability 1, so $E[X] \geq E[Y]$. By direct computation,

$$E[Y] = aP(X \geq a)$$

So therefore,

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

6.2 Chebyshev's Inequality

Let X be a random variable with a finite mean μ and finite variance σ^2 . Chebyshev's inequality states that, for some $k > 0$,

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

Proof

Consider the random variable $Y = (X - \mu)^2$ and the constant k^2 . By Markov's inequality,

$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X - \mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

Additionally,

$$(X - \mu)^2 \geq k^2 \iff |X - \mu| \geq k$$

6.3 Law of Large Numbers**6.3.1 Weak Law**

Suppose X_i are i.i.d. (independent and identically distributed) random variables with mean μ . The empirical average A_n of the first n trials is defined to be

$$A_n := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The Weak Law of Large Numbers states that, for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|A_n - \mu| > \varepsilon) = 0$$

As we average over more and more identical trials with a *true* mean μ , the likelihood that we will find an empirical mean that differs from μ by any distance ε tends to 0.

6.3.2 Strong Law

Given the empirical mean A_n of n i.i.d. trials of X_i , each with a true mean μ ,

$$\lim_{n \rightarrow \infty} A_n = \mu$$

with probability 1. Note that the limit is *inside* of the probability measure, which is a contrast from the weak law: more specifically, the weak law refers to a convergence *in probability* (i.e. how the sequence of probabilities pertaining to the random variables will converge), whereas the strong law refers to an *almost surely* total convergence (i.e. how the sequence of random variables themselves will converge). The weak law says that A_n is *very likely* to be near μ as $n \rightarrow \infty$, which leaves some room for the possibility that it isn't; the strong law says that A_n *is* μ as $n \rightarrow \infty$.

6.4 Central Limit Theorem

Let X_1, X_2, \dots be a sequence of i.i.d. random variables with mean μ and variance σ^2 . We define a sum of the X_i 's:

$$S_n := \sum_{i=1}^n X_i$$

such that $E[S_n] = n\mu$ and $\text{Var}[S_n] = n\sigma^2$. We can furthermore define

$$B_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

so B_n measures the difference between S_n and its expectation, in standard deviations. The Central Limit Theorem states

$$\lim_{n \rightarrow \infty} P(a \leq B_n \leq b) = \Phi(b) - \Phi(a)$$

where Φ is the cdf of the normal distribution. That is, as more samples over X_i are taken, the difference between the sum of the trials S_n and its expectation approaches a normal distribution.

6.5 Markov Chains

Let X_0, X_1, \dots be a sequence of random variables, taking values from a discrete finite set $\{0, 1, \dots, M\}$. We can say that each of the $M + 1$ values represents a state, and the value of the RV X_n represents the particular state a system takes on at time step n .

We can model transitions between states as probabilities, such that P_{ij} represents the probability that the system will transition to state j in the next time step, given that it is currently in state i .

$$P(X_{n+1} = j \mid X_n = i) = P_{ij}$$

such that

$$P(X_{n+1} = j) = P_{ij} \cdot P(X_n = i)$$

More specifically, the sequence X_0, X_1, \dots is a *first-order Markov chain* if its state transitions follow the probabilities P_{ij} . The chain is first order because the probabilistic transitions only depend on the current state taken by the system — if we were to consider the history of the chain (the last k states taken by the system), we would call such a progression a k th order Markov chain.

We can place the probabilities into a convenient matrix representation:

$$P = \begin{bmatrix} P_{00} & P_{01} & \cdots & P_{0M} \\ P_{10} & P_{11} & \cdots & P_{1M} \\ \vdots & \vdots & \ddots & \vdots \\ P_{M0} & P_{M1} & \cdots & P_{MM} \end{bmatrix}$$

P is called a (right) Markov matrix. The i th row of P represents the probability distribution of possible *next states* the system could take, given that the system is currently in state i . Therefore, each row must add up to 1:

$$P_{00} + P_{01} + \cdots + P_{0M} = 1$$

We may define an initial *state vector* $\mathbf{p}_0 = [p_0 \ p_1 \ \cdots \ p_M]$, with each p_i indicating the probability that the system is in state i at time 0. From there,

we note that the probability that the system is in state i in the next time step is the sum:

$$p_0 P_{0i} + p_1 P_{1i} + \cdots + p_M P_{Mi}$$

which is the dot product between \mathbf{p}_0 and the i th column of P . Therefore,

$$\mathbf{p}_1 = \mathbf{p}_0 P$$

Further,

$$\mathbf{p}_2 = \mathbf{p}_1 P = \mathbf{p}_0 P^2$$

More generally,

$$\mathbf{p}_n = \mathbf{p}_0 P^n$$

It turns out that the eigenvalues of any Markov matrix have magnitudes less than or equal to 1. If P has a (left) eigenvector π corresponding to $\lambda = 1$, π called the *stationary distribution* of P , such that any further state transition from π results in the same distribution.

$$\pi = \pi P$$

To see why, consider the diagonalization of P : $P = X \Lambda X^{-1}$, and

$$P^n = X \Lambda^n X^{-1} = X \begin{bmatrix} \lambda_1^n & 0 & \cdots & 0 \\ 0 & \lambda_2^n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_M^n \end{bmatrix} X^{-1}$$

As $n \rightarrow \infty$, if $\lambda_i < 1$, then $\lambda_i^n \rightarrow 0$. The magnitude of λ must be 1 for its corresponding eigenvector state to “survive.” If $\lambda = 1$, then its corresponding distribution is the steady state of the system; $\lambda = -1$ corresponds to long-term oscillating behavior.

A Markov chain is called *ergodic* if there exists some power k such that its corresponding probability matrix P raised to the k th power has all non-zero entries. Intuitively, this corresponds to the idea that all states in our environment are reachable *at some point*. Furthermore, if a Markov chain is ergodic, then it has a unique stationary distribution; there is only one eigenvector π that corresponds to $\lambda = 1$ for P . Considering stationary distributions is useful for reasoning about the long-term behavior of any particular system, modeled as a Markov chain.

7 Entropy

Colloquially, entropy has come to refer to the amount of chaos, or disorder in a given system. If an environment is more disordered, it is said to have a higher entropy. We can also express entropy in terms of events and “surprises” — for an event with a lower probability of occurring, we would expect to be more surprised to discover that it happened. Entropy is the quantified measure of this surprise.

Suppose that X is a discrete RV that takes on some value x_i with probability p_i . We say that the surprise of finding out that X takes on some value x_i is $-\log P(X = x_i) = -\log p_i$. The *expectation* of the surprise of X is

$$H(X) = - \sum_i p_i \log p_i$$

$H(X)$ is the *entropy* of the random variable X . In information theory, $H(X)$ indicates the average amount of *information* we receive when we learn the value of X .

For two random variables X and Y ,

$$H(X, Y) = - \sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

where we see the ordered pair (X, Y) as a random variable in itself. If X and Y are independent, their entropies add:

$$H(X, Y) = H(X) + H(Y)$$

More generally,

$$H(X, Y) = H(X) + H_X(Y)$$

where $H_X(Y)$ is the conditional entropy of Y given X :

$$\begin{aligned} H_{X=x_j}(Y) &= - \sum_i p(y_i | x_j) \log p(y_i | x_j) \\ H_X(Y) &= \sum_j H_{X=x_j}(Y) p_X(x_j) \end{aligned}$$

In addition,

$$H_Y(X) \leq H(X)$$

and they are equal iff X and Y are independent. This is because we cannot “forget” anything in this model — the amount of information unknown to us can only decrease. Once we definitively learn something about X (through Y), there is less to be surprised about regarding X . Furthermore, for any function f ,

$$H(f(X)) \leq H(X)$$

with equality if f is injective (i.e. each value that $f(X)$ can take on is uniquely determined by one value of X). This is true for a similar reason; the number of different values of $f(X)$ is either equal to or less than the number of different values of X ; the number of bits of separate information can either stay the same or decrease after applying f , but it can never increase.

7.1 Noiseless Coding

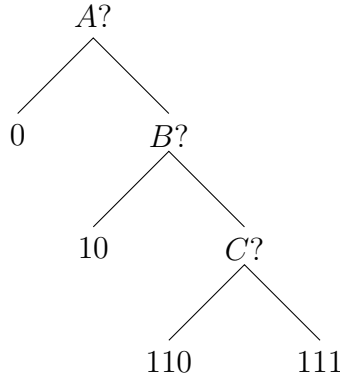
Entropy is a well-used concept in coding theory, which concerns the encoding of objects into bits for digital representation and transmission. In particular, we are interested in minimizing the number of bits required to uniquely determine a message with a given encoding.

Encodings are required to be *prefix-free*, meaning that no encoded character may be the prefix for another. This is because information is transmitted as a single continuous bitstring (e.g. 100011110) that is read left to right, with no separation symbols.

Suppose each character in our message is a random variable X that can take on four values with probabilities:

$$\begin{aligned} P(X = A) &= \frac{1}{2} \\ P(X = B) &= \frac{1}{4} \\ P(X = C) &= \frac{1}{8} \\ P(X = D) &= \frac{1}{8} \end{aligned}$$

The most effective coding scheme is to treat each bit like a “yes/no” question, roughly halving the probability space with each additional bit. For X , the resulting tree looks like



such that the mapping becomes

$$\begin{aligned} A &\mapsto 0 \\ B &\mapsto 10 \\ C &\mapsto 110 \\ D &\mapsto 111 \end{aligned}$$

This encoding takes

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75 = H(X)$$

expected bits, which is the entropy of X . In fact, the noiseless coding theorem states that the expected number of bits you need to achieve a minimal encoding is at most $H(X) + 1$, and is exactly $H(X)$ if the probability space of outcomes of X can be exactly divided in half with every bit.

8 Martingales

Let X_0, X_1, X_2, \dots be a sequence of random variables. We can think about this sequence as learning information in stages, such that X_i represents what we learn at the i th stage/time step. Furthermore, we denote all the information that is known up to the n th stage (that is, the outcomes of X_0, X_1, \dots, X_n) as \mathcal{F}_n . For any random variable Z , we say that $E[Z|\mathcal{F}_n]$ is the conditional expectation of Z , given all of the information known/contained in \mathcal{F}_n .

The sequence X_0, X_1, \dots is a *martingale* if, for all n ,

$$E[|X_n|] < \infty \text{ and } E[X_{n+1}|\mathcal{F}_n] = X_n$$

that is, each X_i has a finite expectation, and the expectation of the value of the next stage *conditioned on* information known up to this point is the current value.

Conditional expectations are martingales, by the law of iterated expectations. For any random variables Y_i , the sequence

$$E[X], E[X|Y_1], E[X|Y_1, Y_2], E[X|Y_1, Y_2, Y_3] \dots$$

is a martingale.

8.0.1 Optional Stopping Theorem

Suppose T were a non-negative, integer-valued random variable. We can think of T as the time at which some *decision* is made, for example, selling a stock. Formally, T is a *stopping time* if the event that $T = n$ depends only on the values X_i for $i \leq n$ — a decision is made knowing only the values up until the n th time step.

For example, T can represent the smallest time at which $|X_T| = 50$, but T cannot be the value within the first 100 time steps at which X_T is largest (as, for some n , we cannot predict that $T = n$ without knowing whether X_n will ever take on larger values).

Doob's Optional Stopping Theorem states that, if the sequence X_0, X_1, \dots is a *bounded* martingale, and T is a stopping time, then

$$E[X_T] = X_0$$

Suppose $X_0 = x$ for $x \in [a, b]$. Suppose that the probability that the sequence of X_i 's reaches a before b is p_a , and let T be the first time at which X_i reaches a or b . Then,

$$E[X_T] = a \cdot p_a + b \cdot (1 - p_a) = x$$

and

$$p_a = \frac{b - x}{b - a}$$

8.1 Finance

8.1.1 Risk-Neutral Probability

Risk neutral probability is a probability measure that can be deduced just by looking at prices on the market. It is best described as a ratio of market prices. For any event A , the risk-neutral probability $P_{RN}(A)$ is

$$P_{RN}(A) = \frac{\text{Price of contract paying \$1 if } A \text{ occurs}}{\text{Price of contract paying \$1 no matter what}}$$

So if the price of a contract paying a dollar if A occurs is \$0.75, then we say that $P_{RN}(A) = 0.75$.

The fundamental theorem of asset pricing states that the interest-discounted price of an asset is a martingale with respect to risk-neutral probability. If interest is calculated according to a constant r , then the price of a contract that pays some amount X at time T is

$$X_0 = E_{RN}[X_T]e^{-rT}$$

where $E_{RN}[X_T]$ is the risk-neutral expectation. If A and B are disjoint events, and the contract pays \$2 if A occurs and \$3 if B occurs, then the price of the overall contract is

$$X_0 = (2P_{RN}(A) + 3P_{RN}(B))e^{-rT}$$

8.1.2 Call Function

For a non-negative random variable X ,

$$E[X] = \int_0^\infty (1 - F_X(x))dx$$

More generally, we can start the integral bound at some positive constant K . The resulting value is called the *call function* of X .

$$C_X(K) = \int_K^\infty (1 - F_X(x))dx$$

Taking derivatives of C , we see that

$$C'_X(x) = 1 - F_X(x), \quad C''_X(x) = -f_X(x)$$

We can express C_X as an expectation.

$$C(K) = E[\max(X - K, 0)]$$

A European call option gives the holder the right to buy a stock for K dollars at time T , but the holder is not obligated to buy the stock if it will not benefit them. If X is the stock price at time T , then the value of buying the stock for K dollars at time T is $\max(X - K, 0)$. If K exceeds X , then the holder will be spending more money than X is worth, so the stock will have no value to them.

The risk neutral expectation of a European call option, then, is

$$E[\max(X - K, 0)] = C(K)$$

so that the price of the contract is

$$C(K)e^{-rT}$$

8.1.3 Black-Scholes Model

The Black-Scholes model assumes that the logarithm of an asset price X at time T is a normal random variable $N = \mathcal{N}(\mu, T\sigma^2)$ — X is a *lognormal* random variable.

$$X = e^N$$

We note that

$$E[X] = E[e^N] = e^{\mu + T\sigma^2/2}$$

If X_0 is the current price,

$$X_0 = E_{RN}[X]e^{-rT} = e^{\mu + (\sigma^2/2 - r)T}$$

so that

$$\mu = \log X_0 + (r - \sigma^2/2)T$$

The big result of the Black-Scholes model was discovering that μ can be determined by just knowing X_0 , σ^2 , and r — it is not a degree of freedom.

In general, if g is any function, then the price of a contract that pays $g(X)$ at time T is

$$E[g(e^N)]e^{-rT}$$

for normal N .