Contents

1 Axioms of Probability

A random experiment is a process leading to outcomes in a sample space Ω, with the actual outcome depending on influences that we can't predict.

1.1 Sample Space and Outcomes

The sample space is the set of all possible outcomes of a random experiment. For example, the sample space of tossing a coin is

$$\Omega = \{H, T\}$$

The sample space will depend on the specific observation/outcomes we are interested at in the random experiment. More detailed observations lead to larger sample spaces.

Examples

1. For the measurement of number of phone calls arriving at a call center in a fixed time period, $\Omega = \mathbb{Z}^+$
2. For the proportion of people who approve of the PM, $\Omega = [0, 1] \cap \mathbb{Q}$
3. For an observation of whether a species is extinct in 100 years time, $\Omega = \{T, F\}$

Simulations of random experiments consists of performing the experiments on a computer instead of in real life. Its benefits are

1. Try multiple or all possibilities before building
2. To perform a large number of repetitions really quickly
3. Study the behavior of complicated random experiements

Events is a set of possible outcome, that it is a subset of Ω. Define that an event occurs if the observed outcome of the experiment ω is in the event set.

The sample space Ω is the certain event. The impossible event is the empty set.

Events are sets so we can do set operations on them.

- $A \cup B$ is the event that either A or B occurs
- $A \cap B$ is the event that both A and B occurs
- A^c is the even that A does not occur
- $A \subset B$ means that A is a subset (or equal) of B. we put a line through the symbol to represent not equal

- $\#A$ is the number of elements in A
- $A \setminus B$ is the event A but not in B

Two events are mutually exclusive or disjoint if their intersection is the empty set

$$A_i \cap A_j = \emptyset$$

This is also for a sequence of events.

A sequence of events is exhaustive if their union is the whole sample space. So, any outcome is in at least one event.

$$A_1 \cup A_2 \cup \cdots = \Omega$$

The distributive laws

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

De Morgan's laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

The intuitive definition of a probability function is the relative frequencies of the outcome when repeated a lot of times. But this is troublesome to replicate in real life for events they only happen once. The Bayesian interpretation of probability includes the personal bias that reflects the odds you will bet on.

1.2 Theorems of the Probability Function

The Axiomatic probability function P (a mapping from events to real numbers) has 3 axioms that reflect the intuitive probability function features

- $P(\Omega) = 1$, let the probability of the certain be 1
- $P(A) \geq 0$, let the probability always be non-negative
- For a sequence of disjoint events $\{A_1, A_2, \dots\}$, let the probability of the union of those events be the sum of their individual probabilities

  $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

  This is the countable additivity axiom. The finite additivity is this axiom with a specific number of disjoint events.

Putting a dot on the union means that the events are disjoint.

From the axioms, we can deduce the following properties of the probability function (never use these in a proof and always use the fundamental 3 axioms)

- $P(\emptyset) = 0$ because

$$\begin{aligned} P(\Omega \cup \emptyset \cup \emptyset \cup \dots) &= P(\Omega) + \sum P(\emptyset) \\ &= P(\Omega) \\ 1 + \sum P(\emptyset) &= 1 \\ \sum P(\emptyset) &= 0 \\ P(\emptyset) &= 0 \end{aligned}$$

- Finite additivity can be proven by induction, or we can do

$$\begin{aligned} P\left(\bigcup_i^n A_i\right) &= P\left(\bigcup_i^n A_i \cup \emptyset \cup \emptyset \dots\right) \\ &= \sum_i^n P(A_i) + \sum P(\emptyset) \\ &= \sum_i^n P(A_i) \end{aligned}$$

- For any event

$$A \subset \Omega, P(A^c) = 1 - P(A)$$

  because $\Omega = A^c \cup A$ and use finite additivity

- If $A \subset B$, then $P(A) \leq P(B)$, because $B = A \cup (B \setminus A)$ and use finite additivity
- For all

$$A \in \Omega, P(A) \leq 1,$$

  because $A \subset \Omega$ so by the last theorem, $P(A) \leq P(\Omega)$.

- The addition theorem

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

because

$$\begin{aligned} A \cup B &= A \cup (B \setminus A) \\ \implies P(A \cup B) &= P(A) + P(B \setminus A) \\ B &= (B \setminus A^c) \cup (B \setminus A) \\ \implies P(B) &= P(B \setminus A^c) + P(B \setminus A) \\ \implies P(B) &= P(A \cap B) + P(B \setminus A) \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

- Continuity theorem, if

$$A_i \subset A_2 \subset A_3 \dots$$

and $A = \cup_i^\infty A_i$, then

$$\lim_{n\to\infty} P(A_i) = P(A)$$

And the decreasing continuity has

$$A_1 \supset A_2 \supset A_3 \dots, A = \bigcap_i^\infty A_i$$

and the same limit continuity holds.

To show this, define $D_i = A_i \setminus A_{i-1}$. Notice that $A_n = \cup_i^n D_i$, and that they are all disjoint. So

$$\begin{aligned} \lim_{n\to\infty} P(A_n) &= \lim_{n\to\infty} P(\bigcup_i^n D_i) \\ &= \lim_{n\to\infty} \sum_i^n P(D_i) \\ &= \sum_i^\infty P(D_i) \end{aligned}$$

Realize that $\cup_i^n A_i = \cup_i^n D_i$, so taking the limit

$$\begin{aligned} P(A) &= P(\bigcup_i^\infty A_i) \\ &= P(\bigcup_i^\infty D_i) \\ &= \sum_i^\infty P(D_i) \\ &= \lim_{n\to\infty} P(A_n) \end{aligned}$$

We can use the proof for the first part to prove the decreasing continuity. Notice that

$$A_1^c \subset A_2^c \subset \dots$$

because $B \subset C \implies C^c \subset B^c$, and

$$A^c = (\bigcap_i^\infty A_i)^c = \bigcup_i^\infty A_i^c$$

from the countable De Morgan's law.

Hence by the first part, we have

$$\begin{aligned}
\lim_{n\to\infty} P(A_i^c) &= P(A^c) \\
1 - \lim_{n\to\infty} P(A_i^c) &= 1 - P(A^c) \\
\lim_{n\to\infty} 1 - P(A_i^c) &= 1 - P(A^c) \\
\lim_{n\to\infty} P(A_i) &= P(A)
\end{aligned}$$

We are sure that the limit is a finite real number because the RHS is limited between $[0, 1]$, which also means that the LHS series must converge because it is monotonically increasing in $[0, 1]$ (but why).

The probability function is a set function from the class of events to real numbers within $[0, 1]$.

For a discrete outcome space (Ω is finite or countably infinite)[1], we have that

$$P(A) = P(\bigcup_{\omega\in A} \{\omega\}) = \sum_{\omega\in A} P(\{w\})$$

due to countable additivity.

In general, an event is possible to have zero probability but is not the empty set (say that P of a normal random variable takes on a specific value is zero, because we cannot construct $\mathbb{R}$ as a countable union of partitions), and an event can have probability of 1 but is not the whole sample space.

[1] A set is discrete if it is countable, which means that it is finite or countably infinite. A set is finite if the number of elements in it is a natural number. A set is countably infinite if there is a bijection from it to the natural numbers

1.3 Countable sets

The set of all finite sequences of zeros and ones is countable, because it keeps partitioning the space of $[0, 1]$ where 0 means to take the left part and 1 means the right part, hence we can bfs from the binary search tree to generate all of them.

The set of all infinite sequences of zeros and ones is uncountable, because each can be mapped to a real value (taking the limit of the finite partition above), but $[0, 1]$ is not countable, so neither is the sequences.

1.4 Classical Probability

The classical probability function on a finite sample space is a valid probability function. Say $\#(\Omega) = n$, then

$$\omega \in \Omega, P(\{\omega\}) = \frac{1}{n}$$

and

$$A \subset \Omega, P(A) = \frac{\#(A)}{n}$$

This indicates that each outcome is equally likely.

Practically, the sample space for n independently tossed coins is the set of ordered n-tuples, forming a classical probability function.

Birthday Problem The sample space of the birthdays of n people is an ordered n tuple. Under the classical probability function, each outcome has a probability of

$$P(\{d_1, d_2, \ldots, d_n\}) = \frac{1}{365^n}$$

Let A be the event that at least 2 people share the same birthday, but we will use A^c which is that no people share the same birthdays. The number of elements in A^c is a permutation

$$\#(A^c) = 365 \times 364 \times \cdots \times (365 - n + 1)$$

And $P(A^c)$ is

$$P(A^c) = 1 \times (1 - \frac{364}{365}) \times (1 - \frac{363}{365}) \times \ldots$$

after using the classical probability function.

To find $P(A)$, we use the theorem where $P(A) = 1 - P(A^c)$.

Bertrand's Paradox Randomly draw two chords in a circle with radius and find the average distance between the two. This problem is ill-defined, for what does random mean?

2 Conditional Probability

If event H occurred, what effect does this information have on the probability of another event A. This is like working with a new sample space H and adjusting the probabilities.

2.1 Definition

Define the conditional probability of A given H as

$$P(A|H) = \frac{P(A \cap H)}{P(H)}, P(H) > 0$$

for it is the relative probability of A and H compared to H.

The multiplication theorem is

$$P(A \cap H) = P(A|H)P(H) = P(H|A)P(A)$$

and allows us to compute event intersections.

For example, roll two fair dices, let A be the events where their numbering differs by at most 1, and H be the events where they add to 7. The conditional probability of A given H is

$$\begin{aligned} \#(H) &= 6 \\ \#(A \cap H) &= 2 \\ P(A|H) &= \frac{\#(A \cap H)}{\#(H)} \\ &= \frac{1}{3} \end{aligned}$$

2.2 Relations

When $P(A|H) > P(A)$, where A is more likely to occur when H occurs, we can derive that it is symmetrical and

$$P(A|H) > P(A) \equiv P(H|A) > P(H)$$

Define a positive relationship between A and H when

$$P(A|H) > P(A)$$

and a negative relationship between them when

$$P(A|H) < P(A)$$

When $P(A|H) = P(A)$, where knowing H does not impact the chance of A, we say that they are independent.

A positive relationship implies that knowing one event increases the probability of the second event occuring; a negative relationship means that knowing one event decreases the probability of the second event; independent events means that knowing one event does not impact the chance of the other event.

2.3 Independent events

Define that two events A and B are independent when

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

This is a special version of the multiplication theorem

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

The empty event is independent to all other events, and the sample space is also independent to all other events.

Two events are dependent when they are not independent.

If the physical processes do not influence each other, we assume that they are independent. But probabilities on the same random physical process may be dependent.

Independence is different to mutually exclusive. Disjoint sets are about sets and does not depend on the probability function, while independent must be with regard to a probability function.

Unless one or both events have zero probability, two disjoint sets cannot be independent for knowing one changes the probability of outcomes for the other.

$$0 = P(A \cap B) = P(\emptyset) < P(A)P(B)$$

If A and B are independent, we have that

- A^c and B are independent
- So, A and B^c are independent
- A^c and B^c are independent

To extend the idea of independence to more than two events. The definition that mutual independence over countable many events is pairwise independence does not work, because a triplet of events may not be independent.

Therefore, define mutual independence of the set $A_1, A_2, \ldots, A_n$ when for any finite subcollection $\{j_1, j_2, \ldots, j_n\}$ where $1 \leq j_i \leq n$ are indices of the set,

$$P(A_{j_1} \cap A_{j_2} \cap \ldots) = P(A_{j_1})P(A_{j_2}) \times \ldots$$

This is to say that a sequence of events is mutually independent if and only if any finite subcollection is independent within. The amount of equations to check/verify will grow exponentially.

Given a mutually independent set of events, we can create other mutually independent sets by picking sets of non-overlapping events and doing set operations within those sets. So

- $\{A_1^c, A_2^c\}$
- $\{A_1 \cup A_2, A_3\}$

are all mutually independent.

In independent systems: systems in parallel each with probability of p to fail has a working rate of $(1 - p^n)$ by De Morgan's law; systems in series has a working rate of p^n by independence. Parallel systems have higher working rates then an equivalently sized series system.

2.4 Uniform Sampling

We say a uniform sample from a finite set S with n elements is when

$$\omega \in S \implies P(\{\omega\}) = \frac{1}{\#S}$$

We represent k uniform samples in S of size n with replacements by the sample space

$$\Omega = \{(x_1, x_2, \ldots, x_k) \colon x_i \in S\}$$

where each outcome is equally likely by $P(\{\omega\}) = \left(\frac{1}{n}\right)^k$. When sampling with replacements, we can sample more elements than items in the set.

To have a k uniform sample in S without replacements, the sample space is

$$\Omega = \{(x_1, x_2, \ldots, x_k) \colon x_i \in S, x_i \text{ are all distinct}\}$$

with $P(\{\omega\}) = (n-k)!/n!$. This is because there is $n!/(n-k)!$ total elements (by permutations) in the set, and with a uniform sample, each outcome is equally likely.

We can exploit symmetry when computing the probability of the ith sample when sampling without replacement is a specific element by realizing that every element has an equal chance to be the ith sample.

2.5 Law of Total Probability

Define a partition of a set S to be a countable sequence that is mutually disjoint and exhaustive

$$\{A_1, A_2, \dots\} \iff \bigcup_i^n \{A_i\} = \Omega, A_i \cap A_j = \emptyset \iff i \neq j$$

Consider a countable partition of the set where each item in the partition has non-zero probability, $P(A_i) > 0$, we have that

$$\begin{aligned} P(H) &= P(H \cap \Omega) \\ &= P(\bigcup_i^n H \cap A_i) \\ &= \sum_i^n P(H \cap A_i) \\ &= \sum_i^n P(H|A_i)P(A_i) \end{aligned}$$

This is the law of total probability, given a countable partition of the set A_i, we have that

$$P(H) = \sum_i^n P(H|A_i)P(A_i)$$

this computes the probability of the effect H given the probabilities of the causes A_i.

For the simple partition $\{A, A^c\}$

$$P(H) = P(H|A)P(A) + P(H|A^c)P(A^c)$$

2.6 Baye's theorem

Consider the events where H^+ is when the person is HIV positive, T^+ is when the person is tested positive. Note that $P(H^+)$ is the HIV positive probability for the whole population, and $P(T^+)$ is the probability that a random person is tested positive.

We note that

$$P(T^+) = P(T^+|H^+)P(H^+) + P(T^+|H^{c+})P(H^{c+})$$

by the law of total probability.

Consider the opposite

$$P(H^+|T^+) = \frac{P(H^+ \cap T^+)}{P(T^+)} = \frac{P(T^+|H^+)P(H^+)}{P(T^+)}$$

which is Baye's theorem.

In general for rare diseases, a test must be very accurate and specific for it to have any effect on the posterior probability on whether the person carries the disease.

Define Bayes' formula by: let A_i be a partition of Ω, then for any event H

$$P(A_i|H) = \frac{P(H|A_i)P(A_i)}{P(H)} = \frac{P(H|A_i)P(A_i)}{\sum_j P(H|A_j)P(A_j)}$$

an interpretation is that given the effect, what is the probability of the cause.

3 Random Variables

Define a random variable X as a function $X\colon \Omega \to \mathbb{R}$. The random variable is essentially a mapping from the outcomes to real numbers.

We denote random variables by capital letters, and their values with lower case letters.

Define the set of possible values (state space) of X by $S_X \subset \mathbb{R}$. Namely, $S_X = ran(X) = X[\Omega]$.

Let the indicator random variable be defined by: let Ω be a non-empty set and $A \subset \Omega$ be an event, the indicator of A is the random variable

$$1_A\colon \Omega \to \mathbb{R}, 1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

and $S_{1_A} = \{0, 1\}$.

The random variable doesn't have to be bijective.

3.1 Events of random variables

Denote the set $\{X = x\}$ by $\{\omega\colon X(\omega) = x\}$ for the random variable X and value $x \in S_X$, and likewise with other binary operations ($\{X \leq x\} = \{\omega\colon X(\omega) \leq x\}, \{x < X \leq y\} = \{\omega\colon x < X(\omega) \leq y\}$). The probability that X is x is defined by

$$P(X = x) = P(\{\omega\colon X(\omega) = x\})$$

A countable set is finite or has a bijection to the natural numbers, an uncountable set does not.

3.2 Discrete random variables

Define a discrete random variable to be the random variables where S_X is countable. That is, X can only take countable many values. This is guaranteed if Ω is countable.

For a discrete random variable X, let the probability mass function (pmf) $p_X : S_x \to [0, 1]$ be

$$p_X(x) = P(X = x)$$

It is the masses of probability assigned to each value of X. We can use anything in place of that x, and omit the X on $p_X(x)$ if X is in context. The pmf can also be defined on $\mathbb{R}$ and set to zero for all values outside S_X.

We use the pmf to determine the probability distribution (distribution function) of the random variable X.

Thus for the indicator random variable

$$p_{1_A}(1) = P(A) \quad p_{1_A}(0) = P(A^c)$$

Properties of the pmf are

- $p_X(x) \geq 0$ for all $x \in S_X$, because $p_X(x)$ is a probability
- $\sum_{x \in S_X} p_X(x) = 1$, because

$$P(X \in S_X) = 1 = P(X = x_1 \cup X = x_2 \dots) = \sum_{x \in S_x} p_X(x)$$

any function defined on a countable subset of $\mathbb{R}$ satisfies both properties will be a pmf for some random variable.

In general due to the additive axiom on disjoint unions

$$P(X \in B) = \sum_{x \in B \cap S_X} p_X(x)$$

and specifically when $B = \{S_X \leq x\}$

$$P(X \leq x) = \sum_{y \leq x} p_X(y)$$

3.3 Cumulative distribution function

Define the (cumulative) distribution function F_X of any random variable X be a function $F_X \colon \mathbb{R} \to [0,1]$ where

$$F_X(x) = P(X \leq x)$$

For example, suppose that $A \subset \Omega$

$$F_{1_A}(x) = \begin{cases} 0 & x < 0 \\ P(A^c) & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

because the $P(1_A = 0) = P(A^c)$ and $P(1_A = 1) = P(A)$.

Properties of the cdf

- $0 \leq F_X(x) \leq 1$. Because the cdf is a probability
- $P(a < X \leq B) = F_X(b) - F_X(a)$ when $a < b$. Because $\{X \leq a\} \cup \{a < X \leq b\} = \{X \leq b\}$
- $F_X(x)$ is non-decreasing because of the previous property
- $\lim_{n \to -\infty} F_X(x) = 0$, $\lim_{n \to \infty} F_X(x) = 1$
- $F_X(x)$ is right continuous as in

$$\lim_{h \to 0^+} F_X(x+h) = F_X(x)$$

- $P(X = x)$ is the jump of the cdf at x from the left, namely

$$P(X = x) = F_X(x) - \lim_{h \to 0^+} F_X(x-h)$$

Any function that satisfies property 3,4,5 must be a distribution function of some random variable.

Proof. To see 4, consider $A_n = \{X \leq n\}$, and $A_1 \supset A_2 \ldots$, and $B = \cap A_n = \emptyset$. Hence by continuity

$$\begin{aligned} \lim_{n \to -\infty} F_X(n) &= \lim_{n \to \infty} P(A_n) \\ &= P(B) = 0 \end{aligned}$$

and similarly with the positive infinity case.

To see 5, notice that

$$F_X(x+h) - F_X(x) = P(x < X \leq x+h)$$

and let $A_n = \{x < X \leq x + 1/n\}$, and A_n is getting smaller and smaller, with $\cap A_n = \emptyset$. Therefore by continuity

$$\lim_{h \to 0^+} P(x < X \leq x+h) = \lim_{n \to \infty} P(A_n) = P(\cap A_n) = P(\emptyset) = 0$$

Hence

$$\lim_{h \to 0+} F_x(x+h) - F_X(x) = \lim_{h \to 0+} P(x < X \leq x+h) = 0$$

and by rearranging, we have right continuity.

To see 6, consider $A_n = \{X \leq x - 1/n\}$, and that $\cup A_n = \{X < x\}$. Because $A_1 \subset A_2 \ldots$, by the continuity theorem, we have

$$\begin{aligned} P(\cup A_n) = P(\{X < x\}) &= F_X(x) - P(X = x) \\ &= \lim_{n \to \infty} P(A_n) = \lim_{h \to 0^+} F_x(x-h) \end{aligned}$$

and the theorem is given by rearranging the three components. □

If the cdf of a random variable is continuous, then its cdf's left and right limits must equal for all points, leading to

$$P(X = x) = 0$$

for all x. This intuitively means a lack of jumps in the distribution function.

We can define a distribution function by including positive and negative infinities at infinities. All axioms except 4 will hold.

We can have an uncountable sample space but have a valid probability function on events. These probabilities must be a real number so some events may have probability zero at infinity (probability of not seeing a six in infinite many dice rolls).

For a sample space Ω and two random variables on that sample space X and Y. Then by function properties, $X + Y$, $X - Y$, $\min(X, Y)$, etc are all random variables. The composition X/Y is a random variable if $0 \notin S_Y$ (Y cannot be zero).

3.4 Continuous random variables

We can approximate a discrete cdf with small pmf values by a continuous function. This gets more accurate as the jump approaches zero.

We say that the random variable X has a continuous distribution if its distribution function F_X (satisfying cdf properties) is continuous.

If X is a continuous random variable

- S_X is uncountable because there are no jumps
- $P(X = x) = 0$ for all $x \in \mathbb{R}$. The probability mass function is just all zero.
- We only assign probabilities to intervals from the cdf F_X
- $P(a < X < b) = P(a \leq X \leq b)$. the endpoints of intervals don't matter

We will restrict to a subclass of distributions which are absolutely continuous, where the interval probabilities can be computed by integration.

For a continuous df F_X on a continuous random variable X with density, the probability density functions $f \colon \mathbb{R} \to [0, \infty)$ are defined by

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

The pdf doesn't have to be continuous and is not unique for a specific continuous random variable with density.

If pdf exists for almost all values of x, then it is the derivative of the cdf

$$\frac{dF}{dx} = f(x)$$

Properties of the pdf

- for all $a < b$, by the fundamental theorem of calculus

  $$\int_{a}^{b} f_X(t)\, dt = F_X(b) - F_X(a) = P(a < X \leq b)$$

  where probability of intervals are represented by the area under the graph of f_X.
- that the area under the pdf is 1

  $$\int_{-\infty}^{\infty} f_X(t)\, dt = F_X(\infty) - F_X(-\infty) = 1$$

  for $F_X(\infty) = 1$ and $F_X(-\infty) = 0$.
- $f_X(x) \geq 0$ because F_x is non-decreasing

Any integrable non-negative function that integrates to 1 on entire real number domain is a valid pdf for some continuous random variable with density.

Not all continuous random variables with continuous cdf has a valid density function. But for this subject, we assume all continuous random variables has valid probability density functions.

Discrete	Continuous
pmf	pdf
prob masses	prob densities
$p\_x(x)$	$P(X = x) = 0$
$\sum p\_X(x) = 1$	$\int_{-infty}^{\infty} f_X(t)\,dt = 1$
$P(I) = \sum_{x \in I} p_X(x)$	$P([a, b]) = \int_a^b f(t)\,dt$
$0 \leq p_X(x) \leq 1$	$f_X(x) \geq 0$

We don't need the density function to be lesser than 1 because it is not a probability. It can be interpreted as a probability density around a point x or as the relative probability of observing value near x.

However, the distribution function works in describing both continuous and discrete random variables. It describes the distribution of any random variable X.

We'll focus on continuous random variables with density. There can be random variables that have both discrete and continuous sections. So the categories of random variables are

- Discrete random variable
- Continuous random variable
- Continuous random variable with density
- Neither discrete or continuous

We often only focus on the distribution and density/mass functions of random variables without noting the sample space.

3.5 Expected values

Expected value summarizes the distribution of a random variable.

Expectation value is the expected average result of a large amount of random trials. If the expected value of winnings for a game is negative, then don't play the game in the long run.

3.5.1 Discrete

In the discrete case, with a random variable X, with set S_X and pmf $p_X(x)$, define the expected value/mean to be

$$E[X] = \sum_{x \in S_X} x p_X(x)$$

Provided that either

$$\sum_{x \in S_X \cap [0,\infty)} x p_X(x) \qquad \sum_{x \in S_X \cap (-\infty,0]} x p_X(x)$$

are finite (so if both are infinite, the expected value is not defined for it doesn't converge). The reason for this condition is that we can't compute $\infty - \infty$.

The expected value will always be defined when S_X is finite.

The expected value does not have to be in S_X.

The law of large numbers states that the sample mean approaches the expected value at large sample sizes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{\infty} X_i - E[X] = 0$$

for independent random variable realizations X_i of the random variable X.

We also have the notation $E[X] = \mu$. We can interpret the mean by the center of mass on the pmf.

3.5.2 Continuous

Let X be a continuous random variable with density f_X, the expected value of X is

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

provided that not both the positive $\int_0^{\infty} x f_X(x)\, dx$ and negative $\int_{-\infty}^{0} f_X(x)\, dx$ are infinite (at least one is finite).

Some Riemann sum approximation about the expected value.

3.5.3 Accounting Trick

The accounting trick or the Lebsegue integral defines the expected value for a random variable X by summing over the sample space instead of the range of the random variable. When Ω is countable and that the random variable is always non-negative, we have

$$E[X] = \sum_{\omega\in\Omega} X(\omega) P(\{\omega\})$$

This is simply a probability weighted sum of all outcomes.

We can also use symmetry on the sample space to find the expected value.

Additionally, we can partition a counting random variables into various different possible outcomes and define indicator random variables over it. The sum of the indicator random variables will be the counting variable, hence the sum of the expected values of the indicators will also be the expected value of the counting variable.

3.5.4 Expectation of functions

For a discrete random variable X with S_X and a pmf $p_X(x)$, under the real-valued function $g\colon S_X \to \mathbb{R}$, we have

$$E[g(X)] = \sum_{x \in S_X} g(x) p_X(x)$$

provided that the sum converges.

Similarly, for a continuous random variable with density $f_X(x)$, we have

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

We will only have nice functions where the expected values exist.

Proof. To prove the discrete case, let $Y = g(X)$, and

$$\begin{aligned} E[Y] &= \sum_{y \in S_y} y P(Y = y) \\ &= \sum_{y \in S_Y} y \sum_{x \in S_X, g(x) = y} P(X = x) \\ &= \sum_{y \in S_Y} \sum_{x \in S_X, g(x) = y} g(x) P(X = x) \\ &= \sum_{x \in S_X} g(x) P(X = x) \end{aligned}$$

□

Notice that in general, the expectation of a random variable under a function is different to the function on the expectation, where

$$E[g(X)] \neq g(E[X])$$

(in general means that it does not hold for all possible X and $g(x)$). This equality does hold when $g(X)$ is a linear function.

3.5.5 Linearity of expected values

The sums of expected values for X and Y is

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

as long as we don't have $\infty - \infty$.

To prove the discrete case where S_X and S_Y are finite. Consider $Z = X + Y$

$$\begin{aligned}
E[Z] &= \sum_{z \in S_Z} zP(Z = z) \\
&= \sum_{z \in S_Z} \sum_{x \in S_X} zP(X = x, Z = z) \\
&= \sum_{x \in S_X} \sum_{z \in S_Z} zP(X = x, Z = z) \\
&= \sum_{x \in S_X} \sum_{z \in S_Z} (x + (z - x))P(X = x, Y = z - x) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} (x + y)P(X = x, Y = y) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} xP(X = x, Y = y) + \sum_{y \in S_Y} \sum_{x \in S_X} yP(X = x, Y = y) \\
&= \sum_{x \in S_X} xP(X = x) + \sum_{y \in S_y} yP(Y = y) \\
&= E[X] + E[Y]
\end{aligned}$$

By substitution of $y = z - x$ and the finite additivity on partitioning $\{Z = z\}$ into sets of $\{Z = z, X = x\}$.

There is also

$$E[cx] = cE[X]$$

which is a special case of the expectation of functions, where $g(X) = cX$.

Combined, we have that for random variable X and real numbers a, b:

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

The proof uses the linearity property of summation and integration.

If for two random variables, $0 \leq X \leq Y$. Then

$$0 \leq E[X] \leq E[Y]$$

and the pure inequalities hold when $0 < X < Y$.

Proof. Let $t > 0$, then $\{X > t\} \subset \{Y > t\}$ because for all $\omega \in \Omega, X(\omega) \leq Y(\omega)$.

Then by the subset inequalities

$$P(X > t) \leq P(Y > t)$$

and by tail probabilities

$$E[X] = \int_0^\infty P(X > t)\, dt \leq \int_0^\infty P(Y > t)\, dt = E[Y]$$

□

3.6 Variances

The variance measures the spread of the distribution around the center/mean.

Define the variance of a random variable X by

$$V[X] = E[(X - E[X])^2]$$

such that it measures the expected squared distance from the mean.

The variance measures the consistency of the outcome, smaller variance means the bulk of the distribution is closer to the mean, vice versa.

The variance of a constant random variable is zero.

Denote the variance as $\sigma^2 = V[X]$. The square root of the variance is the standard deviation σ with the same units as the mean.

Alternatively, another measure of spread would be

$$E[|X - \mu|]$$

This is harder to use due to the absolute value.

Variance properties

- When the variance exists due to all expected values existing, $V[X] \geq 0$ due to $(X - \mu)^2 \geq 0$.
- $V[X] = 0 \iff P(X = \mu) = 1$. Variance is zero when the variable is a constant
- $V[X]$ is not defined when $E[X]$ is not finite. It may be infinite when $E[X]$ is finite

The linearity property of variance is

$$V[aX + b] = a^2 V[X]$$

Proof. In general assuming the expected values exist

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

□

This implies that adding a constant does not change the spread of a random variable. But scaling it does.

For a random variable X, if $E[X]$ is finite then we have

$$V[X] = E[X^2] - E[X]^2$$

Proof. Let $\mu = E[X]$, then

$$\begin{aligned} V[X] &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + E[\mu^2] \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - E[X]^2 \end{aligned}$$

□

The standardized random variable for a random variable X is computed by

$$Z_X = \frac{X - E[X]}{\sqrt{V[X]}}$$

assuming the variance is non-zero. We can compute to see that its mean is zero, and its variance and sd is one

$$E[Z_X] = 0 \quad V[Z_X] = \sigma_{Z_X} = 1$$

3.7 Higher moments

Define the kth moment of a random variable X by

$$\mu_k = E[X^k]$$

Define the kth central moment of a random variable X by

$$\mu_k = E[(X - \mu)^k]$$

The expected value is the first moment, the variance is the second central moment.

The tail probability method is an expression to compute the kth moment. For a continuous (or discrete) random variable X with where $P(X \geq 0) = 1$ (implying that it is never negative), if the kth moment exists, we have

$$E[X^k] = \int_0^\infty kx^{k-1} P(X > x)\, dx$$

Particularly, we have the base case where $k = 1$ to compute the expected value as

$$E[X] = \int_0^\infty P(X > x)\, dx$$

Proof. We can prove the continuous case using integration by parts

$$\begin{aligned}
E[X^k] &= \int_0^\infty x^k f(x)\, dx \\
&= x^k F(x)\big|_0^\infty - \int_0^\infty kx^{k-1} F(x)\, dx \\
&= \infty^k - k \int_0^\infty x^{k-1} F(x)\, dx \\
&= \int_0^\infty kx^{k-1}\, dx - k \int_0^\infty x^{k-1} F(x)\, dx \\
&= k \int_0^\infty x^{k-1} (1 - F(x))\, dx \\
&= k \int_0^\infty x^{k-1} P(X > x)\, dx
\end{aligned}$$

where we assume that $\infty = M$ where we take the limit as $M \to \infty$.

Alternatively, we can prove the base case sing the indicator function and the expected value of the indicator.

$$\begin{aligned} X &= \int_0^X 1\,dx \\ &= \int_0^\infty 1_{\{x<X\}}\,dx \\ E[X] &= E[\int_0^\infty 1_{\{x<X\}}\,dx] \\ &= \int_0^\infty E[1_{\{x<X\}}]\,dx \\ &= \int_0^\infty P(x<X)\,dx \end{aligned}$$

□

The equivalent discrete case method for $X \geq 0$ is

$$E[X] = \sum_{n=0}^{\infty} P(X > n)$$

Proof. With the fact that $P(X > t) = P(X > n) - P(n < X \leq t) = P(X > n)$

$$\begin{aligned} E[X] &= \int_0^\infty P(X>t)\,dt \\ &= \sum_{n=0}^{\infty} \int_n^{n+1} P(X>t)\,dt \\ &= \sum_{n=0}^{\infty} \int_n^{n+1} P(X>n)\,dt \\ &= \sum_{n=0}^{\infty} P(X>n) \end{aligned}$$

□

There is an alternative negative version of the tail probability method, where for a random variable X, and $P(X \leq 0) = 1$

$$E[X^k] = -k \int_{-\infty}^{0} x^{k-1} F(x)\,dx$$

4 Discrete Distributions

There are some distributions which are common enough that we've named their pdf, pmf, and df.

4.1 Bernoulli Distribution

Consider a random variable with two possible values. These options are success or failures. Such random variables must be an indicator function on some set of outcomes. The Bernoulli distribution is a distribution on an indicator random variable with a pmf of

$$p(1) = p \quad p(0) = (1-p)$$

for some $p \in [0, 1]$.

A random variable X with such a pmf is called a Bernoulli random variable of parameter p. We denote that by $X \sim B(1, p)$.

A trial that generates Bernoulli random variables are called Bernoulli trials.

The moments for the Bernoulli random variables are

$$E[X] = np \quad V[X] = p(1-p)$$

Many other discrete distributions focus on a sequence of independent Bernoulli trials.

4.2 Binomial distribution

For a sequence of $n > 0$ independent bernoulli trials with success probability $p \in [0, 1]$. The sample space is the ordered set of n successes or failures. The probability of getting a specific sequenced outcome is

$$P(\{\omega\}) = p^{\text{successes}}(1-p)^{\text{failures}}$$

The number of successes in the trials has a binomial distribution is a random variable X. The probability mass function is

$$\begin{aligned} p_X(k) &= \text{number of sequences with } k \text{ successes} \times \text{P(event)} \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

Where the choose function returns the number of ways to order k successes in n elements

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

We can say that X has a binomial distribution with parameter p and n. Namely, $X \sim B(n, p)$

If a discrete random variable pmf has the probability distribution of a binomial distribution with parameters n and p, it is binomially distributed with parameters n and p.

The binomial theorem states that for all $n > 0$ and any numbers a and b:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

We can use the theorem to check that our pmf is a valid pmf

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

and of course the pmf is positive.

The moments of a binomial distributed random variable are

$$\begin{aligned} E[X] &= np \\ E[X(X-1)] &= n(n-1)p^2 \\ V[X] &= np(1-p) \end{aligned}$$

Proof. For the expected value, notice that $X \sim B(n, p)$, $J \sim B(n-1, p)$

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\ &= \sum_{k=1}^{n} n \frac{(n-1)!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\ &= np \end{aligned}$$

The general moments relationship for the $m > 0$th moment is

$$\begin{aligned}
E[X^m] &= \sum_{k=0}^{n} k^m \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k^{m-1} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} k^{m-1} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{j=0}^{n} (j+1)^{m-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} \\
&= npE[(Y+1)^{m-1}]
\end{aligned}$$

where $Y \sim B(n-1, p)$.

Therefore, the variance calculation is

$$\begin{aligned}
E[X^2] &= npE[Y+1] = np((n-1)p+1) \\
V[X] &= np((n-1)p+1) - (np)^2 = np(1-p)
\end{aligned}$$

□

A binomial distributed random variable can be decomposed into the sum of n independent Bernoulli random variables each with parameter p. $X \sim B(n,p)$, $Y \sim B(1,p)$

$$X = \underbrace{Y + Y + Y + \ldots}_{n}$$

we can use the linearity of expectations to find the mean and variance of binomials in this way.

Therefore, the assumptions for a variable to be binomial distributed is: a discrete set of outcomes, same probability of successes, and independent.

To find the shape of the binomial distribution, consider the ratio of successive binomial probabilities

$$r(x) = \frac{p_X(x)}{p_X(x-1)} = \frac{\frac{n+1}{x} - 1}{\frac{1}{p} - 1}$$

Notice that this ratio is decreasing and it is a recursive formula for the binomial probabilities.

To find when the pmf is increasing or decreasing:

- If $x < p(n+1)$, then $r(x) > 1$ and the pmf is increasing
- If $x > p(n+1)$, then $r(x) < 1$ and the pmf is decreasing
- Therefore the pmf increases until and including floor $p(n+1)$ and decreasing after when $p(n+1)$ is not an integer.

If $p(n+1)$ is an integer, $r(p(n+1)) = 1$ and thus the binomial probabilities $x = p(n+1)$ and $x = p(n+1) - 1$ are equal. The pmf will increase up to $p(n+1) - 1$ and decreasing after $p(n+1)$.

The binomial distribution models sampling with replacements.

4.3 Geometric Distribution

For a sequence of independent bernoulli trials of probability $p \in (0, 1]$, let X be the discrete random variable for the number of failures until the first success. Then

$$S_X = \{0, 1, 2, \dots\}$$

If we are counting the number of trials til the first success, we use $Y = X + 1$.

The pmf of X is

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

We say that X has geometric distribution $X \sim G(p)$ with parameter p if it has a pmf like above.

We can check that this is a valid pmf using the geometric series formula.

The memory-less property of the geometric distribution on the random variable X is summarized by

$$P(X = x | X \geq s) = P(X = x - s)$$

which is saying that if we already had s failures, the number of total failures needed for a success has the same distribution as the geometric distribution on the extra number of failures, as if the s failures never happened.

The moments of the geometric random variable are

$$E[X] = \frac{1-p}{p} \quad V[X] = \frac{1-p}{p^2}$$

we can use the discrete tail probabilities formula and infinite geometric sum to find the values.

Proof. Using the discrete tail probability.

$$\begin{aligned} E[X] &= \sum_{n=0}^{\infty} P(X > n) \\ &= \sum_{n=0}^{\infty} (1-p)^{n+1} \\ &= \sum_{n=1}^{\infty} (1-p)^{n} \\ &= \frac{1-p}{1-(1-p)} = \frac{1-p}{p} \end{aligned}$$

□

The number of failures after the first success until the second success is also a geometric distribution. Therefore, the number of failures until the second success is just the sum of the two geometric random variables.

4.4 Negative Binomial

Negative binomial measures the number of failures until the rth success.

The probability of seeing r successes at the $r+k$th trial is seeing $r-1$ successes before the $r+k$th trial. Each event would have identical probability and there are $\binom{r+k-1}{r-1}$ choices for those successes, the pmf is

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

and $S_X = \mathbb{Z}^+$. A random variable with the pmf and parameters $r \geq 1$ and $p \in (0, 1]$ has a negative binomial distribution where $X \sim NB(r, p)$.

For the alternative notation where negative binomial measures the number of trials to the rth success. The random variable would be $Y = X + r$.

We can generalize this by defining the generalized binomial coefficient

$$\binom{x}{k} = \frac{x(x-1)\dots(x-k+1)}{k!} \quad \forall x \in \mathbb{R}, k \in \mathbb{N}$$

and write the pmf of the negative binomial as

$$\begin{aligned} \binom{-r}{k} &= (-1)^k \binom{r+k-1}{r-1} \\ p(k) &= \binom{-r}{k} p^r (p-1)^k \end{aligned}$$

where $r > 0$, under the generalized binomial coefficient.

This is a negative binomial because of its similar pmf as the binomial, but the n is negative.

The negative binomial of parameter r is equivalent to the sum of r geometric variable with the same p. This is because we can treat the number of failures to get another success after a success as another independent geometric variable. The sum of r of them would be the total number of failures to get r successes.

$$X = \underbrace{Y + Y + \dots}_{r}$$

where $X \sim NB(r, p)$ and $Y \sim G(p)$.

The moments of the negative binomial are

$$E[X] = \frac{r(1-p)}{p} \quad V[X] = \frac{r(1-p)}{p^2}$$

which can be derived by the decomposition into geometric variables and the linearity of expected values.

Proof. We only consider the integer case where $r \in \mathbb{N}$.

Due to the geometric variable decomposition

$$E[X] = E[\sum_{i=1}^{r} Y] = \sum_{i=1}^{r} E[Y] = \frac{r(1-p)}{p}$$

Similarly, for variances because the geometric variables are independent. □

The ratio of successive probabilities are

$$r(k) = \left(\frac{r-1}{k} + 1\right)(1-p)$$

which decreases with k. This is also a good formula to compute negative binomial probabilities.

The shape of the distribution is

- If $k < (r-1)(1-p)/p$ then the pmf is increasing
- If $k > (r-1)(1-p)/p$ then the pmf is decreasing
- If the cutoff is an integer, there will be two maximum values.

4.5 Hypergeometric distribution

A distribution that models sampling without replacement.

For n objects where the number of defective items is d, the number of defective items X selected in a sample of m without replacement (uniformally random sequentially or all at once) has the probability distribution

$$p(k) = \frac{\binom{d}{k}\binom{n-d}{m-k}}{\binom{n}{m}}$$

and $S_X = \{0, 1, \ldots, m\}$.

The actual domain for the random variable is

$$\{\max(0, m - (n - d)), \ldots, \min(d, m)\}$$

but this is accounted for in the binomial coefficients.

The moments are

$$E[X] = \frac{md}{n} \quad V[X] = \frac{md(n-d)}{n^2}(1 - \frac{m-1}{n-1})$$

Proof. To prove the expected value formula.

Let $X = \sum_{i=1}^{m} 1_{A_i}$ where A_i is the event that the ith selection is defective. Notice that $P(1_{A_i}) = d/n$ for symmetry.

Then by linearity

$$E[X] = E[\Sigma] = \sum_{i=1}^{m} E[1_{A_i}] = \frac{md}{n}$$

□

Notice that the first part of the variance formula $p(1-p)$ is equivalent to the binomial and assumes that the indicators are independent. The correction term makes the variance of the hypergeometric variable smaller than the binomial.

For a large enough n compared to m, there is enough elements such that sampling without replacement is close enough to sampling with replacement. Hence the hypergeoemtric distribution is approximated by the binomial distribution with parameters m and $p = d/n$ at large values of n and small values of m.

4.6 Poisson Distribution

A poisson variable is a binomial variable where the bernoulli trials are happening continuously and independently. It is a limited sequences of Bernoulli trials.

Assume each bernoulli trial takes $1/n$ time and has a probability of success λ/n. Within a unit time the number of successes is

$$p(k) = \binom{n}{k}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k}$$

by letting $n \to \infty$ (shrinking the time length and success probability of each trial), we have

$$p(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

where $X \sim Pn(\lambda)$ is a poisson random variable if it has this pmf with parameter $\lambda > 0$. $S_X = \mathbb{Z}^+$.

To show that the pmf is valid, notice that $p(k) \geq 0$ and notice the taylor series

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

so that $\sum p(k) = 1$.

The moments for the poisson distribution has the mean be

$$\begin{aligned}
E[X] &= \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\
&= \sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}e^{-\lambda} \\
&= \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} \\
&= \lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{(k)!}e^{-\lambda} \\
&= \lambda
\end{aligned}$$

and the variance is

$$\begin{aligned} E[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!}e^{-\lambda} \\ &= \lambda^2 \end{aligned}$$

where

$$\begin{aligned} V[X] &= E[X(X-1)] + E[X] - E[X]^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda \end{aligned}$$

We see that both the mean and the variance are equal to the rate.

4.6.1 Binomial Approximation

The poisson distribution is an approximation to the binomial distribution when n is large and p is small, namely

$$\lim_{n\to\infty} B(n,p) = Pn(np)$$

hence we can model a binomial distribution by a poisson distribution directly with these conditions, usually when $p \leq 0.05$.

4.6.2 Poisson model

The poisson model can be used to model calls in a call center in a time span, radioactive decay.

The assumptions for a poisson model is a binomial distribution where p is very small, but np is normal sized. By letting $\lambda = np$, we have an approximation to the binomial.

In general, a poisson model on the random variable X requires the condition that

- $X = X_1 + X_2 + \cdots + X_n$ where X_i has the same bernoulli distribution
- The p of each bernoulli distribution is very small, but np is normal sized
- X_i is locally independent (that knowing X_i would only affect a few X_j).

we can approximate X by $X \sim Pn(np)$.

4.6.3 Poisson process

A poisson process is the process that led to the derivation of the poisson distribution, namely, it assumes that the count in two different periods are independent, and events happen continuously with a rate λ.

4.7 Discrete Uniform Distribution

We say that X has a discrete uniform distribution if all its outcome in an integer range $[a, b]$ has equal probability, its pmf is

$$p(k) = \frac{1}{b - a + 1}$$

where $X \sim U(a, b)$ for integers a, b and $a < b$. If a random variable has this pmf, it is a discrete uniform random variable.

Consider a standard discrete uniform distribution $X \sim U(0, n)$, its moments are

$$\begin{aligned}
E[X] &= \sum_{k=0}^{n} k \frac{1}{n+1} \\
&= \frac{1}{n+1} \frac{n(n+1)}{2} \\
&= \frac{n}{2} \\
E[X^2] &= \sum_{k=0}^{n} k^2 \frac{1}{n+1} \\
&= \frac{1}{n+1} \sum_{k=0}^{n} k^2 \\
&= \frac{1}{n+1} \frac{n(n+1)(2n+1)}{6} \\
&= \frac{n(2n+1)}{6} \\
V[X] &= E[X^2] - E[X]^2 \\
&= \frac{n(n+2)}{12}
\end{aligned}$$

4.7.1 Bernoulli relation

Given a sequence of independent bernoulli trials with parameter p, given that there is exactly one success in the first n trials. Let X be the index of the trial with the success in the first n trial, then

$$X \sim U(0, n)$$

4.8 Relations between distributions

For $Y \sim B(n,p)$ and $X_i \sim B(1,p)$, then

$$Y = X_1 + X_2 + \cdots + X_n$$

For $X_i \sim G(p)$ and $Y \sim NB(r,p)$, then

$$Y = X_1 + X_2 + \cdots + X_r$$

For $X \sim B(m, d/n)$ and $Y \sim Hg(m,n,d)$, then

$$n \gg m \implies Y \approx X$$

For $Y \sim NB(r,p)$ and $X \sim B(n,p)$, then

$$P(Y > n) = P(X < r)$$
$$P(Y \leq n) = P(X \geq r)$$

For $Y \sim Po(np)$ and $X_i \sim B(1,p)$ where X_i are only locally dependent and n is large

$$X_1 + X_2 + \cdots + X_n \approx Y$$

For $X \sim B(n,p)$, $Y \sim U(1,n)$

$$\text{success index} | Y = 1 = X$$

5 Continuous special random variables

5.1 Continuous uniform random variable

Suppose $a < b$ are numbers, a continuous random variable X with uniform pdf within $[a,b]$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

and we say that $X \sim R(a,b)$. For a continuous uniform distribution, the probability density across $[a,b]$ are constant.

The uniform distribution implies that the relative probability of the variable being anywhere within (a,b) is equal. This is useful to model uniform experiments.

The cdf of a uniform random variable $X \sim R(a, b)$ is

$$F(x) = P(X \leq x) = \begin{cases} \frac{x-a}{b-a} & x \in [a, b] \\ 1 & x > b \\ 0 & x < a \end{cases}$$

The moments of the random variable (and the standard uniform $U \sim R(0, 1)$) is

$$\begin{aligned} E[U] &= \frac{1}{2} \\ E[X] &= \int_a^b x \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a}(b^2/2 - a^2/2) \\ &= \frac{a+b}{2} \\ V[U] &= \frac{1}{12} \\ V[X] &= \frac{(b-a)^2}{12} \end{aligned}$$

we can also derive this by shifting and scaling the standard uniform of $R(0, 1)$. Namely, if $U \sim R(0, 1)$, and $X = (b - a)U + a$, then

$$X \sim R(a, b)$$

and vice versa.

5.2 Exponential distribution

The exponential random variable is the continuous variant of the geometric variable. It models the waiting time until a success for a continuously independent series of bernoulli trials.

Suppose that each trial takes $1/n$ time with $p = \alpha/n$ for n trials in unit time. In time t, there are nt trials, with the probability of no success up to time t

$$P = (1 - \frac{\alpha}{n})^{nt}$$

by letting the number of trials in unit time go to infinity, we have

$$\lim_{n \to \infty} P = e^{-\alpha t}$$

Let T be the waiting time for the first success, it is to say that the chance of T being longer than a time t is having no success in t

$$P(T > t) = e^{-\alpha t}$$

The cdf of T is therefore

$$F(t) = 1 - P(T > t) = 1 - e^{-\alpha t} \quad t \geq 0$$

and notice that $S_T = [0, \infty)$, with $F(t) = 0$ for $t < 0$.

The pdf is calculated by differentiating the cdf

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

and we can check that this is a valid cdf/pdf function.

The random variable with this cdf and pdf is called an exponential random variable with parameter α. We say that the random variable T has an exponential distribution and $T \sim \exp(\alpha)$.

The moments of the exponential distribution $X \sim \exp(\alpha)$ is

$$\begin{aligned} E[X] &= \int_0^\infty P(T > t)\, dt \\ &= \int_0^\infty e^{-\alpha t}\, dt \\ &= \frac{1}{\alpha} \\ E[X^2] &= \int_0^\infty 2t(1 - F(t))\, dt \\ &= \frac{2}{\alpha} \int_0^\infty t\alpha e^{-\alpha t}\, dt \\ &= \frac{2}{\alpha} E[X] \\ &= \frac{2}{\alpha^2} \\ V[X] &= E[X^2] - E[X]^2 \\ &= \frac{1}{\alpha^2} \end{aligned}$$

The lack of memory property of the exponential distribution is that the probability distribution of the additional waiting time given an already amount of waiting is also exponentially distributed with the same rate. That is for the exponential rv $T \sim \exp(\alpha)$ for all $t \geq 0$

$$(T-t)_{T>t} \sim T$$

Proof. Let $Y = (T-t)_{T>t}$ and $T \sim \exp(\alpha)$, then for non negative y,

$$\begin{aligned} P(Y=y) &= P(T-t=y|T>t) \\ &= P(T=y+t|T \geq t) \\ &= \frac{P(T=y+t)}{P(T \geq t)} \\ &= \frac{\alpha e^{-\alpha(y+t)}}{e^{-\alpha t}} \\ &= \alpha e^{-\alpha y} \end{aligned}$$

showing that $Y \sim \exp(\alpha)$. □

For a random variable T with a probability p of being zero and an exponential distribution of parameter α otherwise with probability $1-p$, the pdf is

$$f(t) = \begin{cases} p & t = 0 \\ (1-p)\alpha e^{-\alpha t} & t > 0 \end{cases}$$

5.3 Gamma distribution

A gamma random variable is a continuous variant of the negative binomial random variable. It models the waiting time until the rth occurrence of a success, assuming a continuous series of independent bernoulli trials (a poisson process).

To derive the gamma distribution, notice that $T > t$ implies at most $r-1$ successes in the first nt trials for a continuous bernouli trial with $p = \frac{\alpha}{n}$ with n trials in unit time. For the number of successes follows a binomial distribution

$$P(T>t) = \sum_{k=0}^{r-1} \binom{nt}{k} (\frac{a}{n})^k (1-\frac{a}{n})^{nr-k}$$

by taking $n \to \infty$, we have

$$P(T>t) = \sum_{k=0}^{r-1} \frac{\alpha^k}{k!} t^k e^{-\alpha t}$$

and the cdf is

$$F(t) = \begin{cases} 1 - \sum_{k=0}^{r-1} \frac{\alpha^k}{k!} t^k e^{-\alpha t} & t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We can derive the pdf by differentiation (the process is tedious and requires canceling sums), leading to the gramma distribution's pdf

$$\begin{aligned} f(z) &= \frac{\alpha^r z^{r-1}}{\Gamma(r)} e^{-\alpha z} \\ &= \frac{\alpha^r z^{r-1}}{(r-1)!} e^{-\alpha z} \end{aligned}$$

when $z \geq 0$, and $f(z) = 0$ for all other z. A random variable with such a pdf is called a gamma random variable with parameter $r > 0$ and rate $\alpha > 0$, then $Z \sim \gamma(r, \alpha)$.

The gamma distribution with $r = 1$ is simply an exponential distribution

$$\gamma(1, \alpha) = \exp(\alpha)$$

When r is an integer, we have that $Y \sim \gamma(r, \alpha)$

$$Y = \underbrace{X + X + \dots}_{r}$$

where $X \sim \exp(\alpha)$. This implies that the gamma distribution is the sum of r independent exponential distributions.

5.3.1 Gamma function

To show that this is a valid pdf and extend r outside of integers, we introduce the gamma/erlang function

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, dx$$

which has the properties

- $\Gamma(1) = 1$
- $\Gamma(r) = (r-1)\Gamma(r-1)$ by integration of parts

implying that $\Gamma(r) = (r-1)!$ for positive integer values r. Then we can show that

$$\int_0^\infty f(z) \, dz = 1$$

by integration by parts.

5.3.2 Moments of the gamma function

The kth moment of $X \sim \gamma(r, \alpha)$ is

$$\begin{aligned} E[X^k] &= \int_0^\infty \frac{\alpha^r z^{r-1}}{\Gamma(r)} e^{-\alpha z}\, dx \\ &= \frac{1}{\alpha^k \Gamma(r)} \int_0^\infty e^{-u} u^{r+k-1}\, du \quad u = \alpha x \\ &= \frac{\Gamma(r+k)}{\Gamma(r)\alpha^k} \\ E[X] &= \frac{r}{\alpha} \\ V[X] &= E[X^2] - E[X]^2 \\ &= \frac{r}{\alpha^2} \end{aligned}$$

notice that k can be any value where the integral converges. If k is a positive integer, we can use the factorial representation of $\Gamma(r)$.

We can also use the independent sum of exponential decomposition and use the linearity of mean and variance.

5.4 Beta distribution

Represents continuous quantitative process where the observed value is in a bounded interval. The proportion of people who voted for a party is beta distributed. (We often model proportions using beta distributions)

A random variable X follows a beta distribution with parameter $a > 0$ and $b > 0$ if it has a pdf

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

where

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

and we say that $X \sim \text{Beta}(a, b)$.

When $a = b = 1$, we have a standard uniform distribution. When $a = b$, we get a symmetric distribution. The beta distribution generalizes the uniform distribution.

5.4.1 Beta moments

The kth moment of a beta distribution X

$$\begin{aligned} E[X^k] &= \frac{\mathrm{B}(a+k,b)}{\mathrm{B}(a,b)} \\ &= \frac{\Gamma(a+k)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+k)} \end{aligned}$$

which is derived from noticing that the integral of the kth moment is another beta function.

And other moments are

$$\begin{aligned} E[X] &= \frac{a}{a+b} \\ V[X] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

5.5 Pareto distribution

The exponential distribution has a thin tail where the tail probability decays exponentially.

Many processes will have a tail distribution that decays as a reciprocal of a power, this is the power law and we use a Pareto distribution.

A random variable with pdf

$$f(x) = \frac{\gamma a^\gamma}{x^{\gamma+1}}$$

for $a \leq x < \infty$. We say the variable has a Pareto distribution $X \sim \text{Pareto}(a, \gamma)$ with parameter $a > 0$ and $\gamma > 0$. The denominator is the decay power term and the numerator is a scalar to maintain the pdf.

The tail probability is

$$P(X > x) = \frac{a^\gamma}{x^\gamma}$$

The pareto distribution models processes with outcomes above a with a tail decay rate of γ. Say the income distribution of population above a level a is pareto distributed.

5.5.1 Moments of pareto distribution

Notice that for the kth moment to converge and exist, $\gamma > k$.

$$\begin{aligned} E[X] &= \int_a^\infty x \frac{\gamma a^\gamma}{x^{\gamma+1}}\, dx \\ &= \gamma a^\gamma \int_a^\infty x^{-\gamma}\, dx \\ &= \gamma a^\gamma \frac{a^{\gamma-1}}{1-\gamma} \\ &= \frac{\gamma a}{\gamma - 1} \quad \gamma > 1 \\ E[X^2] &= \frac{a^2\gamma}{\gamma - 2} \quad \gamma > 2 \\ V[X] &= \frac{\gamma a^2}{(\gamma-1)^2(\gamma-2)} \quad \gamma > 2 \end{aligned}$$

Because of the heavy tail property of pareto distributions, it is not always that it has a finite mean and variance.

5.6 Normal distribution

The normal distribution is the limit of the sum of independently, identically distributed random variables (the central limit theorem).

The domain of the normal distribution is the entire real number line. The pdf of a normal distribution is

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

and X is a normal distribution with parameter μ and $\sigma > 0$, and $X \sim N(\mu, \sigma^2)$.

When $Z \sim N(0,1)$, then it is a standard normal with pdf

$$f(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

It is clear that $f(z) \geq 0$, but to check that it is a pdf $\int_{-\infty}^{\infty} f(z)\, dz = 1$ is a bit harder, and requires the double integral polar transformation.

5.6.1 Standardization of Normal distribution

If $Z = \frac{X-\mu}{\sigma}$, with $\sigma > 0$ and $X \sim N(\mu, \sigma^2)$ being a normal distribution, then the df for Z is

$$\begin{aligned} F_Z(z) &= P(Z \leq z) \\ &= P(\frac{X-\mu}{\sigma} \leq z) \\ &= F_X(\mu + \sigma z) \end{aligned}$$

the pdf is the derivative of the df, so

$$f_Z(z) = \sigma f_X(\mu + \sigma z) = \varphi(z)$$

by substitution into the normal pdf formula. So $Z \sim N(0,1)$ and is a standard normal variable.

The other direction also works, where if Z is a standard normal, then

$$X = \mu + \sigma Z, X \sim N(\mu, \sigma^2)$$

which uses the same idea. Also notice that $E[X] = \mu + \sigma^2 E[Z]$.

This means that we only need the standard normal distribution to compute any normal distribution. The distribution function for the standard normal is

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx$$

there is no explicit formula, this is usually computed as a z table numerically. Notice that $\varphi(z)$ is symmetrical and

$$\Phi(-z) = 1 - \Phi(z)$$

5.6.2 Moments of a standard normal

For $Z \sim N(0,1)$, its moments are

$$\begin{aligned} E[Z^n] &= \int_{-\infty}^{\infty} x^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int x^{n-1} (-e^{-\frac{1}{2}x^2})'\, dx \\ &= \frac{1}{\sqrt{2\pi}} \left([x^{n-1}(-e^{-\frac{1}{2}x^2})]_{-\infty}^{\infty} - (n-1) \int x^{n-2}(-e^{-\frac{1}{2}x^2})\, dx \right) \\ &= (n-1) \frac{1}{\sqrt{2\pi}} \int x^{n-2} e^{-\frac{1}{2}x^2}\, dx \\ &= (n-1) E[Z^{n-2}] \end{aligned}$$

by integrating by parts and noticing the last integral is the $n-2$th moment of Z.

Therefore, because $E[Z^0] = 1$ (φ is a pdf) and $E[Z^1] = 0$ (because φ is even), we see that odd moments of Z are all zero by induction

$$E[Z^{2k+1}] = 0$$

and the even moments are

$$E[Z^{2k}] = (2k-1)(2k-3)\dots(1) = (2k)!/(2^k k!)$$

The variance and mean of Z is therefore

$$E[Z] = 0, \quad V[Z] = 1$$

and the mean and variance of $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$ is

$$E[X] = \mu, \quad V[X] = \sigma^2$$

by the linear properties of mean and variances.

Notice that in general, two distributions with the same mean and variance are not necessarily equal. But two normals with the same parameter are equal in distribution.

5.7 Normal Approximation

We might can't find probabilities of the binomial distribution when n is large because individual probabilities are small and the choose function is large.

In general for $X \sim B(n, p)$

- If p is small and np is normal sized, we use a Poisson approximation
- If p is close to 1, we can switch success with failure and then p is small
- If p is away from 0 and 1, we use a normal approximation.

The rule of thumb for the normal approximation is $np > 5$ and $n(1-p) > 5$. Then $X \approx N(np, np(1-p))$. Otherwise we use a poisson approximation.

We can also approximate poisson and gamma distributions by normal distributions.

- If $X \sim Pn(\lambda)$ and λ is large, then $X \approx N(\lambda, \lambda)$.
- If $X \sim \gamma(r, \alpha)$ and r is big, then $X \approx N(\frac{r}{\alpha}, \frac{r}{\alpha^2})$

This is the result of the central limit theorem, where a sum of identically independently distributed random variables is normal.

To see the normal approximation of poisson distribution, we use the convolution formula. For instance X_1 and X_2 are independent poisson distributed, then $X_1 + X_2$ has the distribution

$$\begin{aligned} P(X_1 + X_2 = k) &= \sum_{i=0}^{i=k} P(X_1 = i, X_2 = k - i) \\ &= \sum_{i=0}^{k} \frac{e^{-\lambda_1}\lambda_1^i}{i!} \frac{e^{-\lambda_2}\lambda_2^{k-i}}{(k-i)!} \\ &= e^{-\lambda_1-\lambda_2} \sum_{i=0}^{k} \frac{\lambda_1^i \lambda_2^{k-i}}{i!(k-i)!} \\ &= \frac{e^{-\lambda_1-\lambda_2}}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda_1^i \lambda_2^{k-i} \\ &= \frac{e^{-\lambda_1-\lambda_2}}{k!} (\lambda_1 + \lambda_2)^k \end{aligned}$$

using the binomial coefficient and binomial formula. We notice that the pdf of the sum is a poission distribution with $X_1 + X_2 \sim Po(\lambda_1 + \lambda_2)$. This implies that the sum of n independent poisson random variables is another poisson random variable with the sum of the rates. For we can approximate poisson distributions with normal distributions if r is large, as $n \to \infty$, $r \to \infty$, and thus the sum of many independent poisson RV is normally distributed.

5.8 Weibull distribution

To model exponential decaying tail behavior, we use the Weibull distribution. The Weibull distribution has an exponential tail, its pdf is

$$f(x) = \frac{\gamma x^{\gamma-1}}{\beta^\gamma} e^{-(x/\beta)^\gamma}$$

where $\beta > 0$, $\gamma > 0$ and $x \geq 0$. We also say the random variable has a weibull distribution $X \sim \text{Weibull}(\beta, \gamma)$.

If $\gamma > 2$, its tail decays faster than normal. If $\gamma < 1$, its tail decays slower than exponential.

The tail behavior of a Weibull distribution is

$$P(X > x) = e^{-(x/\beta)^\gamma}$$

when $x \geq 0$. Its df is

$$F(x) = \begin{cases} 1 - e^{-(x/\beta)^\gamma} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

By directly integrating against the pdf using the definition, and performing u-sub and gamma functions, the expectation and variance of the Weibull distribution is

$$\begin{aligned} E[X] &= \int_0^\infty x f(x)\, dx \\ &= \beta\Gamma(\frac{\gamma+1}{\gamma}) \\ V[X] &= E[X^2] - E[X]^2 \\ &= \beta^2 \left(\Gamma(\frac{\gamma+2}{\gamma}) - \Gamma(\frac{\gamma+1}{\gamma})^2 \right) \end{aligned}$$

neither are nice functions to evaluate by hand.

5.9 Transformation of Distribution

5.9.1 Distribution Scaling

The positive scaling of an exponential distribution is another exponential distribution, for $X \sim \exp(\alpha)$, $d > 0$ and $Y = dX$

$$Y \sim \exp(\frac{\alpha}{d})$$

The scaling and translation of a uniform distribution or a normal distribution is also a uniform/normal distribution.

5.9.2 General Transformations

To compute the distribution of a transformation φ of a random variable X which is $\varphi(X)$, we either use the pdf, df, or the tail distribution. The tail distribution and the df is a lot easier to use than the pdf.

In general, if $Y = \varphi(X)$, then by the cdf

$$F(Y \leq y) = F(\varphi(X) \leq y)$$

and by the tail probability

$$F(Y > y) = F(\varphi(X) > y)$$

and solve the inequality.

For example on the exponential scaling, say $X \sim \exp(\alpha)$

$$P(dX > y) = P(X > \frac{y}{d}) = e^{-\frac{dy}{\alpha}}$$

which is the tail distribution of the exponential random variable $Y \sim \exp(\frac{d}{\alpha})$

Also, for independent exponential random variable $X_1 \sim \exp(d_1)$ $X_2 \sim \exp(d_2)$, the transformation $\min(X_1, X_2)$ has the tail probability

$$\begin{aligned} P(\min(X_1, X_2) > y) &= P(X_1 > y, X_2 > y) \\ &= P(X_1 > y)P(X_2 > y) \\ &= e^{-(d_1+d_2)y} \end{aligned}$$

which is the tail probability of $\exp(d_1 + d_2)$.

For linear functions, say $Y = aX + b$ for $a > 0$, we have the cdf and pdf of Y be

$$F_Y(y) = F_X(\frac{y-b}{a}) \quad f(y) = \frac{1}{a} f_x(\frac{y-b}{a})$$

and in general when X is a continuous random variable with pdf $f_X(x)$, when $a \neq 0$, the pdf of Y is

$$f(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

5.9.3 Generalized Inverses

The supremum is the lowest upper bound in a set A. For instance, the $\sup A$ is the smallest value such that it is larger than or equal to all elements in A, and $\sup\{x \in A \colon p(x)\}$ for predicate $p(x)$ is the greatest x such that $p(x)$ is true (or the smallest value x where $p(x)$ is false for all values above x). The supremum doesn't need to be in the set A.

Define the general inverse function $g^{-1}(y)$ of an increasing function $g(x)$ to be

$$g^{-1}(y) = \sup\{x \colon g(x) \leq y\}$$

for any y. This allows any increasing (and not just strictly increasing) function to have an inverse.

Notice that for the generalized inverse

$$g(x) \leq y \iff g^{-1}(y) \geq x$$

and also

$$g(g^{-1}(y)) = y$$

and in general

$$g^{-1}(g(x)) \neq x$$

Proof. For the first statement, on the forward direction, we have if x is such that $g(x) \leq y$ then by definition of the supremum, $g^{-1}(y) \geq x$. On the reverse direction, we use negation. If $g(x) > y$, for g is continuous, there exists an $x_0 < x$ where $g(x_0) > y$. For g is also increasing, $\forall u \geq x_0, g(u) \geq g(x_0) > y$. Therefore, $\sup\{x \colon g(x) \leq y\} \leq x_0$ else x_0 is a smaller upperbound. But $g^{-1}(y) \leq x_0 < x$, and $g^{-1}(y) < x$.

For the second statement, we see that

$$x > g^{-1}(y) \implies g(x) > y$$

then $g(x) \geq y$, and when

$$x < g^{-1}(y) \implies g(x) \leq y$$

so because $g(x)$ is continuous and by the left and right limits

$$\begin{aligned} g(g^{-1}(y)) &= \lim_{x \to g^{-1}(y)} g(x) \\ &= \lim_{x \to g^{-1}(y)^+} g(x) \geq y \\ &= \lim_{x \to g^{-1}(y)^-} g(x) \leq y \\ g(g^{-1}(y)) &= y \end{aligned}$$

For the third part, consider the cdf $F(x)$ of a standard uniform random variable. When $x < 0, F(x) = 0$. So

$$\begin{aligned} F(x) &= 0 \\ F^{-1}(F(x)) &= F^{-1}(0) \\ &= \sup x \colon F(x) \leq 0 \\ &= 0 \\ x &\neq 0 \end{aligned}$$

So $F^{-1}(F(x)) \neq x$ for this case when $x < 0$. □

5.10 Monotonic Functions

Suppose $Y = g(X)$ where g is an increasing function on S_X, then for $y \in g(S_X)$

$$\begin{aligned} F_Y(y) &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

For example, suppose X is a random variable, and $Y = F_X(X)$ is a transformation using the increasing cdf function of X. therefore the distribution function for $y \in [0, 1]$

$$F_Y(y) = \begin{cases} F_X(F_X^{-1}(y)) = y & y \in [0, 1] \\ 0 & y < 0 \\ 1 & y > 1 \end{cases}$$

and we can tell that this is the df for a continuous uniform RV, $Y \sim R(0, 1)$.

5.11 Pseudo random numbers

We can use the fact that $F_X(X) = R$ to generate random variables X from any distribution F_X given a standard uniform random sample. Let $R \sim R(0, 1)$, and suppose $Y = F_X^{-1}(R)$, then

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X^{-1}(R) \leq y) \\ &= P(F_X^{-1}(R) < y) \\ &= P(R < F_X(y)) \\ &= P(R \leq F_X(y)) \\ &= F_X(y) \end{aligned}$$

notice that R is continuous and the negation $g^{-1}(y) < x \implies g(x) > y$.

Because equal cdf implies equal distributions we have $Y = X = F_X^{-1}(R)$, and we can sample values in X by transforming the uniform sample by F_X^{-1}.

5.12 Cauchy Distribution

Suppose that there is a light pole of height 1, and the angle of the light beam towards the pole is a uniform random variable $\Theta = R(-\pi/2, \pi/2)$. Let X be the ground distance of the light beam, notice that $X = \tan \Theta$. The cdf of X is

$$\begin{aligned} P(X \leq x) &= P(\tan \Theta \leq x) \\ &= P(\Theta \leq \arctan x) \\ &= \frac{\arctan x + \pi/2}{\pi} \end{aligned}$$

The pdf of X is

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

which is named the standard Cauchy distribution. The standard cauchy distribution has a heavy tail and does not have a mean nor variance.

More generally, if X has pdf

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

for all x and parameter m, a, we say that X has a cauchy distribution

$$X \sim C(m, a)$$

The standard cauchy is $X \sim C(0, 1)$.

5.13 Lognormal distribution

Suppose $X \sim N(\mu, \sigma^2)$, let $Y = e^X$.

The possible values of Y are positive, so for all $y > 0$ under the standard normal is

$$\begin{aligned} F(y) &= P(e^X \leq y) \\ &= P(X \leq \log y) \\ &= P(Z \leq \frac{\log y - \mu}{\sigma}) \\ &= \int_{-\infty}^{(\log y - \mu)/\sigma} \frac{1}{2\pi} e^{-x^2/2} \, dx \end{aligned}$$

then the pdf is computed by differentiation

$$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\log y - \mu}{\sigma})^2}$$

for all $x \geq 0$, and we say that $X \sim LN(\mu, \sigma^2)$ is a lognormal distribution with parameter μ, σ^2.

The lognormal distribution has a single mode.

The moments of the lognormal random variable $Y \sim LN(\mu, \sigma^2)$ are

$$\begin{aligned} E[Y^r] &= e^{r\mu + \frac{1}{2}r^2\sigma^2} \\ E[Y] &= e^{\mu + \frac{1}{2}\sigma^2} \\ V[Y] &= e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) \end{aligned}$$

Proof. Consider the standard lognormal $Y \sim LN(0,1)$

$$\begin{aligned} E[Y^r] &= E[e^{rZ}] \\ &= \int_{-\infty}^{\infty} e^{rx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x^2-2rx)/2}\, dx \\ &= e^{r^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-v)^2/2}\, dx \\ &= e^{r^2/2} \end{aligned}$$

Therefore because $Y \sim LN(\mu, \sigma^2)$ has the property

$$Y = e^X = e^{\mu+\sigma Z} = e^{\mu} e^{\sigma Z}$$

so

$$\begin{aligned} E[Y^r] &= E[(e^{\mu} e^{\sigma Z})^r] \\ &= E[e^{r\mu} e^{r\sigma Z}] \\ &= e^{r\mu} e^{r^2\sigma^2/2} \end{aligned}$$

And we can derive the expectation and variance by substituting $r = 1$ and $r = 2$. □

Lognormal distribution models

- Stock prices
- Particle sizes
- Length of words and sentences
- Production data
- Lifetimes of mechanical systems
- Length of stay for patients in hospital

5.14 DF of minimum function

Let $Y = g(X)$ where $g(x) = \min(x, m)$ for some value m. This implies that Y is capped at m.

The cdf of Y is

$$P(Y \leq y) = \begin{cases} F_X(y) & y < m \\ 1 & y \geq m \end{cases}$$

the graph will jump at m if $F_X(m) \neq 1$ to 1 directly.

If there is a jump, the pdf doesn't exist for all S_X for the rv is both continuous and discrete.

The expected value of Y is directly computed (depending on if X is discrete or continuous)

$$\begin{aligned} E[Y] &= \sum_{x<m} xp(x) + m \sum_{x \geq m} p(x) \\ &= \int_{-\infty}^{m} xf(x)\,dx + m \int_{m}^{\infty} f(x)\,dx \end{aligned}$$

by conditioning on whether $x \geq m$.

5.15 Square function

Let $Y = X^2$, then for all $y \geq 0$

$$\begin{aligned} F(y) &= P(Y \leq y) \\ &= P(X^2 \leq y) \\ &= P(X \leq \sqrt{y}) - P(X < \sqrt{y}) \end{aligned}$$

If X is a continuous random variable (so the second strict lesser becomes a non-strict inequality), then

$$f(y) = \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y}))$$

for $y \geq 0$.

By this logic, if $Z \sim N(0,1)$ is a standard normal and $Y = Z^2$, we have

$$Y = Z^2 \sim \gamma(1/2, 1/2)$$

this shows that the Chi squared distribution when $n = 1$ is a gamma distribution of $r = 1/2$ and $\alpha = 1/2$.

Proof. Because Z is continuous, we can use the squared property above,

$$\begin{aligned} f(x) &= \frac{1}{2\sqrt{x}}(\frac{1}{\sqrt{2\pi}}e^{-1/2(\sqrt{y})^2} + \frac{1}{\sqrt{2\pi}}e^{-1/2(-\sqrt{y})^2}) \\ &= \frac{1}{\sqrt{2x\pi}}e^{-y/2} \\ &= \frac{1}{\sqrt{\pi}}(1/2)^{1/2}y^{1/2-1}e^{-y/2} \end{aligned}$$

which is the pdf for the gamma distribution with parameter $r = 1/2, \alpha = 1/2$, for $\Gamma(1/2) = \sqrt{\pi}$.

Notice that we can also use the $y^{r-1}e^{-\alpha y}$ portion to see the gamma distribution, as everything else is a constant and must be the gamma distribution constant as Y is a random variable. □

In general, we can guess the random variable distribution if the changing part (against x) is the same as the changing part of a common distribution, for their constants in front must be the same for them to be both pdfs.

6 Bivariate Random Variables

We may be interested in values of different random variables out of the same random experiment. They may also be related.

A bivariate random variable is a pair of random variables. It is defined as a function that maps Ω to $\mathbb{R}^2$, and can be thought of $(X(\omega), Y(\omega))$ for $\omega \in \Omega$.

The set of possible values of a bivariate random variable (X, Y) is

$$S_{X,Y} = \{(x, y) \colon (X(\omega), Y(\omega)) = (x, y), \omega \in \Omega\} \subseteq \mathbb{R}^2$$

each component X and Y is a univariate random variable (with a marginal distribution). The set of possible values includes all paired outcomes (a subset of $S_X \times S_Y$).

Univariate random variables are points on a real line, and bivariate random variables are points on a real plane.

Define the distribution function on a bivariate random variable (X, Y) by

$$F_{X,Y}(x, y) = P(\{X \leq x\} \cap \{Y \leq y\}) = P(X \leq x, Y \leq y)$$

for $(x, y) \in \mathbb{R}^2$.

A property of df: we can get the probability of the bivariate random variable within a rectangular range of X and Y

$$P(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$$

We can also obtain the univariate distribution function from the joint distribution

$$F_{X,Y}(x, \infty) = F_X(x) \quad F_{X,Y}(\infty, y) = F_Y(y)$$

the converse is not true in general, because the univariate distributions do not have enough information to specific the joint distribution. Except when X and Y are independent, then

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for every $(x,y) \in \mathbb{R}^2$ which can be used to define random variable independence.

6.1 Discrete Bivariate random variables

If $S_{X,Y}$ is countable, we say that (X,Y) is a discrete bivariate random variable. $S_{X,Y}$ is countable if both S_X and S_Y are countable.

Define the joint pmf to be

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

for all $(x,y) \in S_{X,Y}$. It assigns discrete probability masses to every point in the plane.

Its properties are similar to the univariate case

$$\forall (x,y), p(x,y) \geq 0$$
$$\sum_{(x,y)\in S_{X,Y}} p(x,y) = 1$$

Define the marginal pmfs from the joint pmf to be

$$p_X(x) = \sum_{y\in S_Y, (x,y)\in S_{X,Y}} p_{X,Y}(x,y)$$
$$p_Y(y) = \sum_{x\in S_X, (x,y)\in S_{X,Y}} p_{X,Y}(x,y)$$

which represents the univariate pmfs by summing along the other random variable.

We can rewrite the joint df using the joint pmf

$$F_{X,Y}(x,y) = \sum_{u\leq x, v\leq y, (u,v)\in S_{X,Y}} p_{X,Y}(u,v)$$

and we can find the joint pmf from the joint df using the df area formula on a unit square (assuming that X and Y have values in $0, 1, 2, \dots$)

$$p_{X,Y}(x,y) = F(x,y) - F(x-1,y) - F(x,y-1) + F(x-1,y-1)$$

if the random variables have non-integer values, we can label them against the natural numbers for they are countable.

If the joint pmf values are in a table, the marginal pmfs are computed by summing along rows and columns. The joint df is computed by summing values in the square with its corner at the coordinate.

6.2 Continuous bivariate random variable

A bivariate random variable is continuous if its cdf is continuous. It has a joint density function if the distribution function can be written as an integral of a two variable, non-negative function $f : \mathbb{R}^2 \to [0, \infty)$,

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y)\, dydx$$

for every (x, y). Then we call $f_{X,Y}(x, y)$ the joint probability density function.

The joint pdf has properties

- It is non-negative and the integral volume under its domain is 1
- If the pdf exists for almost all (x, y) value, we have

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = f(x, y)$$

  and the order of partial derivatives does not matter

For a bivariate random variable with density, the probability that $a < X \leq b$ and $c < Y \leq d$ is

$$P(a < X \leq b, c < Y \leq d) \int_a^b \int_c^d f(x, y)\, dydx$$

and in general, the integral of pdf within an area is the probability that the bivariate random variable is in that area

$$P((X, Y) \in R) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{(x,y)\in R} f(x, y)\, dydx$$

The intuition of the double integral for $f(x, y) \geq 0$ is

$$V = \int_a^b \int_c^d f(x, y)\, dydx = \lim \sum_i f(x_i, y_i)\Delta i$$

which is the limit as we take smaller partitions of the rectangle $(a, b) \times (c, d)$ into small areas Δ_i. If $f(x, y)$ is a 2D surface, then the double integral gives the volume below the surface in a region.

The double definite integrals can be evaluated sequentially in any order

$$V = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx = \int_c^d \left(\int_a^b f(x, y)\, dx \right) dy$$

we can always switch the integrals in double integrals.

Often, the inner integral has limits as a function of the outer integral variable. When we are integrating a more complicated area, we let one variable to vary over the area, while letting the other depend on the first in the limits of integration, netting a double integral.

The marginal pdfs can be computed from the joint pdf

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

and vice versa when we are computing $f(y)$, this is simply fixing x and integrating over all y. Then $f(x)$ is the density function of X and $f(y)$ is the density function of Y. The reasons are that $f(x, y) \geq 0$ so $f(x) \geq 0$, and $\int_{-\infty}^{\infty} f(x)\, dx = 1$ by the definition of the joint density function.

6.3 Conditional distribution

For a discrete bivariate random variable (X, Y), define the conditional pmf of X given $Y = y$ to be

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

when $p_Y(y) \neq 0$ and for all $x \in S_X$.

The conditional pdf for a continuous bivariate random variable is

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

for all $f_Y(y) \neq 0$ and $x \in S_X$. This is proven by discretizing and approximating the integral.

The conditional pdf/pmf are valid probability pmf pdf functions for they are all positive and integrate to 1.

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)\, dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = 1$$

by the definition of marginal density functions.

We can interpret the conditional pmf to be the probability distribution conditioning on the other variable, and the conditional pdf as the probability density distribution conditioning on the other variable. The conditional variable is written as

$$X_{Y=y}$$

as the conditional distribution of X given $Y = y$.

6.4 Conditional probability

For a continuous bivariate random variable and a y where $f_Y(y) > 0$ and $f_{X|Y}(x|y)$ as the conditional distribution, define

$$P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)\,dx$$

as the conditional probability of $X \in A$ given $Y = y$. This returns an actual probability.

The conditional expectation is defined to be

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

which returns a number.

6.5 Independent of random variables

Two random variables X and Y are independent if

$$P(X \in N, Y \in M) = P(X \in N)P(Y \in M)$$

for every event N, M.

This is equivalent to say

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

for every $(x, y) \in S_{X,Y}$. To generalize independence to higher dimensions, $X_1, X_2, \ldots, X_n$ are independent random variables if and only if

$$F_{\vec{X}}(\vec{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

for every $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. That is, for every instance of the random vector, we can decompose the cdf function.

For discrete random variables, this is equivalent to saying

$$\begin{aligned} p_{X,Y}(x, y) &= p_X(x)p_Y(y) \\ p_X(x) &= p_{X|Y}(x|y) \quad p_Y(y) > 0 \\ p_Y(y) &= p_{Y|X}(y|x) \quad p_X(x) > 0 \end{aligned}$$

for every $(x, y) \in S_{X,Y}$.

For continuous random variables, this is equivalent to saying

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x)f_Y(y) \\ f_X(x) &= f_{X|Y}(x|y) \quad f_Y(y) > 0 \\ f_Y(y) &= f_{Y|X}(y|x) \quad f_X(x) > 0 \end{aligned}$$

for almost every $(x, y) \in S_{X,Y}$.

Intuitively, if the probability distribution of X depends on the value of Y, then they are not independent.

To summarize, independence is

- The cdf is the product of its marginal cdf
- The pmf/pdf is the product of its marginal pmf/pdf
- The conditional distributions are equal to the marginal distributions

6.6 Bivariate normal distribution

If the pdf of (X, Y) is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-p^2}} \exp\left(-\frac{1}{2(1-p^2)}(x^2 - 2pxy + y^2)\right)$$

where $p \in (-1, 1)$, then we say that (X, Y) is the standard bivariate normal distribution with parameter p, and $(X, Y) \sim N_2(p)$.

When $p = 0$, this is the product of two standard normal distribution pdfs, hence this is the joint distribution of two independent standard normal distributions.

When $p > 0$, X and Y are positively correlated; when $p < 0$, X and Y are negatively correlated; and when $p = 0$, X and Y are independent.

When $p = 0$, the pdf is rotationally symmetric. When $p > 0$, the pdf skews towards the $y = x$ line. When $p < 0$, the pdf skews towards the $y = -x$ line.

We can extend this definition to $p = -1$ or $p = 1$. When $p \to 1$, we have $X = Y$. When $p \to -1$, we have $X = -Y$.

The marginal pdf of Y is

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-p^2}} \exp\left(-\frac{1}{2(1-p^2)}(x^2-2pxy+y^2)\right) dx \\
&= \frac{1}{2\pi}\frac{1}{\sqrt{1-p^2}} \int_{-\infty}^{\infty} \exp(-\frac{x^2-2pxy+p^2y^2-p^2y^2+y^2}{2(1-p^2)})\,dx \\
&= \frac{1}{2\pi}\frac{1}{\sqrt{1-p^2}} \int_{-\infty}^{\infty} \exp(-\frac{x^2-2pxy+p^2y^2+y^2(1-p^2)}{2(1-p^2)})\,dx \\
&= \frac{1}{\sqrt{2\pi}}e^{-y^2/2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1-p^2}} \int_{-\infty}^{\infty} \exp(-\frac{(x-py)^2}{2(1-p^2)})\,dx \\
&= \frac{1}{\sqrt{2\pi}}e^{-y^2/2}
\end{aligned}$$

and notice at the second to last step, we have the pdf of $N(py, 1-p^2)$. Hence it is always case for all p that $X, Y \sim N(0,1)$. The p only determines the joint distribution.

Note that if $X, Y \sim N(0,1)$, it does not imply that $(X,Y) \sim N_2(p)$. We need that

$$\forall a, b \in \mathbb{R}, aX + bY \sim N$$

The conditional pdf of X given $Y = y$ is

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-p^2}} \exp(-\frac{x^2-2pxy+y^2}{2(1-p^2)} + y^2/2) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-p^2}} \exp(-\frac{x^2-2pxy+p^2y^2}{2(1-p^2)}) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1-p^2}} \exp(-\frac{(x-py)^2}{2(1-p^2)})
\end{aligned}$$

and $X|Y = y \sim N(py, 1-p^2)$. And $E[X|Y=y] = py$.

This implies that if $p > 0$ and $y > 0$, then $P(X > 1|Y = y) > P(X > 1)$ for they are positively correlated and $Y > 0$ is positive.

6.7 Generalized Bivariate Normal Distribution

If $((X-\mu_X)/\sigma_X, (Y-\mu_Y)/\sigma_Y) \sim N_2(p)$ is a standard bivariate normal distribution, then (X,Y) is a general bivariate normal distribution with parameters $\mu_X, \mu_Y \in \mathbb{R}, \sigma_X, \sigma_Y \geq$

$0, p \in [-1, 1]$

$$(X, Y) \sim N_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, p)$$

and vice versa.

The alternative notation states that a random vector $\begin{bmatrix} X \\ Y \end{bmatrix}$ has a bivariate normal distribution if

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & p\sigma_X\sigma_Y \\ p\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}) = N_2(\mu, \Sigma)$$

with the mean vector μ and covariance matrix Σ.

We can derive the properties of the general bivariate normal distribution from the standard case.

For the bivariate normal distribution (X, Y), because $Z_X = (X - \mu_X)/\sigma_X$ and Z_Y are random variables of the standard distribution, $Z_X \sim Z_Y \sim N(0, 1)$. By the properties of normal distributions

$$X \sim N(\mu_X, \sigma_X^2) \quad Y \sim N(\mu_Y, \sigma_Y^2)$$

For the conditional distribution, notice that

$$Z_X = z_x \iff X = x$$

therefore $(Z_X|Z_Y = z_y) = (Z_X|Y = y) \sim N(pz_y, 1 - p^2)$. And by the linearity property of normal distributions

$$(X|Y = y) = \mu_X + \sigma_X(Z_X|Y = y) \sim N(\mu_X + p\sigma_X z_y, \sigma_X^2(1 - p^2))$$

And $(X|Y = y) \sim N(\mu_X + p\sigma_X \frac{y-\mu_Y}{\sigma_Y}, \sigma_X^2(1 - p^2))$

6.8 Covariance and Correlation

Define the covariance for a general bivariate random variable (X, Y)

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

which we can derive an alternative representation by expanding and using the linearity of expected values

$$Cov(X, Y) = E[XY] - E[X]E[Y]$$

Properties of the covariance

- $Cov(X, X) = V[X]$

- The bilinear property

  $$Cov(aA+bB, cC+dD) = acCov(A,C)+adCov(A,D)+bcCov(B,C)+bdCov(B,D)$$

  which tells us that the covariance of a linear combination of random variables can be "expanded" like they are multiplying. This also applies to linear transformations on variance.

- $Cov(X,Y) = \sigma_{XY}$
- $Cov(X,Y) = Cov(Y,X)$

The correlation coefficient p or $Cor(X,Y)$ is a standardized covariance defined by

$$p_{X,Y} = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{V[X]V[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Properties of the correlation

- $Cor(X,X) = 1$, and $Cor(X,-X) = -1$
- $Cor(X,Y)\sqrt{V[X]V[Y]} = Cov(X,Y)$
- $-1 \leq Cor(X,Y) \leq 1$ always

Proof. To show that $p \in [-1,1]$, consider

$$V[zX+Y] = z^2V[X] + 2zCov(X,Y) + V[Y] \geq 0$$

for any random variable X, Y and number z. This is a function in z, which is a quadratic function that is never negative. This implies that the discriminant is non-positive

$$\begin{aligned}(2Cov(X,Y))^2 - 4V[X]V[Y] &\leq 0\\ Cov(X,Y)^2 &\leq V[X]V[Y]\\ |Cov(X,Y)| &\leq \sqrt{V[X]V[Y]}\\ |Cor(X,Y)| &\leq 1\end{aligned}$$

implying that the correlation is between -1 and 1. □

If the random variables are independent, then the correlation between them is zero. The inverse is not true, we can have two random variables (X, X^2) that are dependent but having zero correlation. The contrapositive is that, if two random variables have non-zero correlation, then they must not be independent.

For any random variable X and Y

$$V[X+Y] = V[X] + V[Y] + 2Cov(X,Y)$$

and notice that in general $V[X+Y] \neq V[X] + V[Y]$. This can be derived from expanding the variance.

If $Cov(X,Y) = 0$, then X and Y are uncorrelated. The only time that $V[X+Y] = V[X] + V[Y]$ is when X and Y are uncorrelated. Also, if X and Y are independent, their covariance is zero for $E[XY] = E[X]E[Y]$. The converse is not true: zero covariance does NOT imply that the random variables are independent.

In general for the random variables $\{X_1, X_2, \dots X_n\}$, we have

$$V[X_1 + X_2 + \dots] = V[X_1] + V[X_2] + \dots + \sum_i \sum_j Cov(X_i, X_j)$$

which can be proven by expanding the LHS (or by induction).

The standard deviation when X and Y are independent random variables

$$Sd[X+Y] = \sqrt{Sd[X] + Sd[Y]}$$

To see the connection between covariance/correlation of random variables and positive/negative relation in events. Consider the indicators 1_A and 1_B, the covariance between them are

$$Cov(1_A, 1_B) = P(A \cap B) - P(A)P(B)$$

In both cases, positive relation is when the covariance is positive, negative is when covariance is negative.

Additionally, independence between X and Y is equivalent to saying that

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$$

for all bounded functions f and g. The special case when $f(x) = g(x) = x$ shows that independence implies zero covariance, but this is not enough in the converse for independence as it doesn't cover all functions f and g.

Proof. In the forward case, consider the class of functions

$$f_z(x) = 1_{x \leq z}$$

applied to X and Y. Then if the expectation relation holds for all functions, it must hold for all f_x and f_y

$$E[f_x(X)f_y(Y)] = E[f_x(X)]E[f_y(Y)]$$
$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$$

which is the cdf definition for independence, showing that X and Y are independent.

In the reverse case, assuming that X, Y are independent

$$\begin{aligned} E[f(X)g(Y)] &= \int\int f(x)g(y)f_{X,Y}(x,y)\,dxdy \\ &= \int f(x)f_X(x)\,dx \int g(y)f_Y(y)\,dy \\ &= E[f(X)]E[g(Y)] \end{aligned}$$

showing that this holds for all bounded functions f and g. □

For the standard and general bivariate normal distribution (X, Y) with parameter p, we see that

$$Cor(X, Y) = p$$

6.9 Polar Transformation and Normal Distribution Scalar

Suppose that $(X, Y) \sim N_2(0)$, then (X, Y) are points in the plane. Consider the distribution of their polar transformation (R, Θ) where $X = R\cos\Theta$ and $Y = R\sin\Theta$.

We have to first find the scalar for the bivariate/univariate normal pdf. First we find the unnormalized area for the bivariate pdf

$$I = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dxdy$$

let $x = r\cos\theta$ and $y = r\sin\theta$, the change of variables is

$$I = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,drd\theta = 2\pi$$

this uses the double integral change of variables formula by the jacobian on the inverse transformation.

Because

$$I = \int_{-\infty}^{\infty} e^{-y^2/2}\int_{-\infty}^{\infty} e^{-x^2/2}\,dxdy = \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2$$

we find that the univariate constant is $1/\sqrt{I} = 1/\sqrt{2\pi}$.

Now we explore about the transformed bivariate random variables. The joint pdf of (R, Θ) is found by normalizing using I and realizing that the inner function fits the description of a pdf (and some rigorous shit about bijective transformations)

$$f(r,\theta) = \int_0^{2\pi} \frac{1}{2\pi}\int_0^{\infty} re^{-r^2/2}\,drd\theta$$

We notice that this is a product of two pdfs, hence R and Θ are independent with the pdfs

$$f(r) = re^{-r^2/2}$$
$$f(\theta) = \frac{1}{2\pi}$$

with $\Theta \in R(0, 2\pi)$ and R is the rayleigh distribution.

6.10 Expectation of functions of bivariate RV

We can directly modify the univariate case of the expectation of a function of a RV into the bivariate case.

If (X, Y) is discrete and $g(x, y)$ is a function of $g\colon S_{X,Y} \to \mathbb{R}$, then

$$E[g(X, Y)] = \sum_{(x,y)\in S_{X,Y}} g(x, y)p(x, y)$$

If (X, Y) are continuous, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,dxdy$$

If X and Y are independent, then

$$E[XY] = E[X]E[Y]$$

the converse is not true, we can have two dependent X and Y having the same property. The general statement for $X_1, X_2, \dots X_n$ independent random variables is

$$E[X_1X_2 \dots X_n] = E[X_1]E[X_2] \dots E[X_n]$$

Proof. Consider when X and Y for discrete random variables. Because they are independent, $p(x, y) = p(x)p(y)$

$$\begin{aligned} E[XY] &= \sum xyp(x, y) \\ &= \sum_x \sum_y xyp(x)p(y) \\ &= \sum_x xp(x) \sum_y yp(y) \\ &= E[X]E[Y] \end{aligned}$$

The continuous case is almost the same with the integral and the factorization of the pdf. □

The expectation of a sum of random variables is the sum of their expectations. For the bivariate random variables (X, Y) with a joint density function

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

and in general for $X_1, X_2, \ldots, X_n$

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

Proof. For the continuous case

$$\begin{aligned} E[X + Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y)\, dxdy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f(x, y)\, dydx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f(x, y)\, dxdy \\ &= E[X] + E[Y] \end{aligned}$$

□

6.11 Convolution Formula

If X and Y are independent discrete random variables

$$P(X + Y = z) = \sum_{x \in S_X} p_X(x) p_Y(z - x)$$

and we let $p_Y(y) = 0$ when $y \notin S_Y$.

If X and Y are independent continuous, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx$$

Proof. For the continuous case,

$$\begin{aligned} P(X + Y \leq z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{x+y \leq z} f(x, y)\, dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} 1_{x+y \leq z} f_X(x) f_Y(y)\, dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y)\, dydx \quad y \leq z - x \\ &= \int_{-\infty}^{\infty} f_X(x) F_Y(z - x)\, dx \end{aligned}$$

Then the pdf of $X+Y$ is

$$\frac{d}{dz}P(X+Y\le z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\,dx$$

□

The convolution formula is symmetrical in the $z-x$ term. We could integrate along y instead and use $z-y$.

Integrals of this form are call convolution integrals.

We can use this to prove that $Pn(\lambda)+Pn(\mu)\sim Pn(\lambda+\mu)$. And $\exp(\lambda)+\exp(\lambda)\sim\gamma(2,\lambda)$.

For example, consider $X\sim\gamma(r,a)$ and $Y\sim\gamma(s,a)$, then

$$\begin{aligned}
f_{X+Y}(z)&=\int_{-\infty}^{\infty}c1_{x\ge0}1_{z-x\ge0}x^{r-1}e^{-ax}(z-x)^{s-1}e^{-a(z-x)}\,dx\\
&=ce^{-az}\int_0^z x^{r-1}(z-x)^{s-1}\,dx\\
&=ce^{-az}\int_0^1(uz)^{r-1}(z-uz)^{s-1}z\,du\quad u=x/z\\
&=ce^{-az}z^{r-1}zz^{s-1}\int_0^1u^{r-1}(1-u)^{s-1}\,du\\
&=cke^{-az}z^{r+s-1}
\end{aligned}$$

where $k=B(r,s)$ is the beta function which is a numbered constant. The non-constant part has the gamma distribution pdf with parameters $r+s$ and a, hence $X+Y\sim\gamma(r+s,a)$.

6.12 Conditional random variable on Events

For $P(A)\ge0$ for an event A, we define the conditional random variable $X|A$ by

$$F_{X|A}(x)=P(x\le X|A)$$

If the random variable is countable

$$p_{X|A}(x)=P(X=x|A)=\frac{p_X(x)1_{x\in A}}{P(A)}$$

to get the pmf of the conditional random variable.

If the random variable is continuous, and if there is a function $g\colon\mathbb{R}\to[0,\infty)$ such that

$$F_{X|A}(x)=\int_{-\infty}^{x}g(x)\,dx$$

then $g(x) = f_{X|A}(x)$ is the pdf of the conditional random variable.

Also notice that if the event is $A = \{X \in I\}$ and the density function $f_X(x)$ exists, we can write the event conditional pdf as

$$f_{X|A}(x) = \frac{f_X(x)1_{x\in I}}{P(X \in I)}$$

because

$$\begin{aligned} P(X \leq y|X \in I) &= \frac{P(X \leq y, X \in I)}{P(X \in I)} \\ &= \int_{-\infty}^{\infty} \frac{f_X(x)1_{x\leq y}1_{x\in I}}{P(X \in I)}\,dx \\ &= \int_{-\infty}^{y} \frac{f_X(x)1_{x\in I}}{P(X \in I)}\,dx \end{aligned}$$

which satisfies the condition for $f_{X|A}(x)$. This along with the discrete case gives an explicit formula for the event conditional pmf/pdf for X.

Define the conditional expectation to be the expectation on this conditional random variable

$$E[X|A] = \begin{cases} \sum_{x\in S_X} x p_{X|A}(x) \\ \int_{-\infty}^{\infty} x f_{X|A}(x)\,dx \end{cases}$$

for the continuous and discrete cases.

6.13 Conditional random variable on Random Variables

Similarly, with the conditional variance

$$V[X|A] = E[(X|A - E[X|A])^2]$$

To motivate the law of total expectation, consider the function

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

which is a function representing the conditional expectation on X when $Y = y$. Applying the function to the random variable Y, we get

$$Z = \eta(Y) = E[X|Y]$$

which is a random variable called the conditional expectation of X given Y, we write it as $E[X|Y]$, which is a RANDOM VARIABLE ON Y, hence $S_Z = S_{E[X|Y]} = S_Y$.

Alternatively, we have that

$$E[X|Y] = \sum_{y \in S_Y} E[X|Y = y] 1_{Y=y}$$

Proof.

$$\begin{aligned} \sum_{y} E[X|Y = y] 1_{Y=y} &= \sum_{y} \eta(y) 1_{Y=y} \\ &= \eta(Y) \end{aligned}$$

With the indicator sum being similar to the property of dirac delta integration, for the summation returns the same values as $\eta(Y)$ for all $Y = y.$, and taken as assumption. □

The conditional expectation for a continuous bivariate random variable (X, Y) with joint density $f_{X,Y}(x, y)$ is

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

with the conditional expectation of $E[X|Y]$ defined similarly

$$E[X|Y] = \eta(Y)$$

The law of total expectation states that for all bivariate random variables (X, Y)

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

Proof. Assuming they are continuous

$$\begin{aligned} E[E[X|Y]] &= E[\eta(Y)] \\ &= \int \eta(y) f_Y(y)\, dy \\ &= \int \int x f_{X|Y}(x|y)\, dx f_Y(y)\, dy \\ &= \int \int x f_{X,Y}(x, y)\, dy dx \\ &= \int x f_X(x)\, dx \\ &= E[X] \end{aligned}$$

□

This is the expected value version of the law of total probability.

Extending this to a function of X, Y, we have

$$E[g(X,Y)] = E[E[g(X,Y)|Y]]$$

which can be proved by creating a new random variable $W = g(X,Y)$, and applying the law of total expectation on W and substitute back the function. Specifically

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

Proof. Consider

$$\begin{aligned}
E[XY|Y] &= \sum_y E[XY|Y=y]1_{Y=y} \\
&= \sum_y E[Xy|Y=y]1_{Y=y} \\
&= \sum_y yE[X|Y=y]1_{Y=y} \\
&= \sum_y YE[X|Y=y]1_{Y=y} \\
&= Y\sum_y E[X|Y=y]1_{Y=y} \\
&= YE[X|Y]
\end{aligned}$$

using the properties of Y being a constant inside $E[XY|Y=y]$, the linear properties of expecations, and the summation property of indicator functions allowing $y = Y$ and taking Y outside the sum for it doesnt depend on Y. □

The conditional variance of the bivariate random variable (X,Y) requires the conditional variance function

$$v(y) = V[X|Y=y]$$

and the conditional variance is

$$V[X|Y] = v(Y)$$

The law of total variance for (X,Y) is

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

Proof. To find the expected value of the conditional variance

$$\begin{aligned} V[X|Y] &= E[X^2|Y] - E[X|Y]^2 \\ E[V[X|Y]] &= E[E[X^2|Y]] - E[E[X|Y]^2] \\ &= E[X^2] - E[\eta(Y)^2] \end{aligned}$$

using the variance formula on conditional random variables and the law of total expectation.

Additionally, to find the variance of the conditional expectation

$$\begin{aligned} V[E[X|Y]] &= V[\eta(Y)] \\ &= E[\eta(Y)^2] - E[\eta(Y)]^2 \\ &= E[\eta(Y)^2] - E[X]^2 \end{aligned}$$

Hence, adding the two values

$$V[E[X|Y]] + E[V[X|Y]] = E[X^2] - E[X]^2 = V[X]$$

netting the law of total variance. □

7 Inequalities and Approximations

Chebyshev's inequality (Bienayme inequality) gives a quantitative idea of standard deviation as a measure of spread.

It states that if X has mean μ and variance $\sigma^2 > 0$, then for all $k > 0$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

for any random variable X. The inequality provides an upperbound to the probability that the random variable is more than k times the standard deviation away from the mean. The one-minus version provides a lower bound to the probability that the random variable is within k times the standard deviation away from the mean.

This is a relatively crude bound.

Markov's inequality states that for $Y \geq 0$, and $a > 0$

$$P(Y \geq a) = \frac{E[Y]}{a}$$

Proof. For any $a > 0$

$$\begin{aligned} E[Y] &\geq E[Y 1_{Y \geq a}] \\ &\geq E[a 1_{Y \geq a}] \\ &= aE[1_{Y \geq a}] \\ &= aP(Y \geq a) \end{aligned}$$

which is markov's inequality. □

We can use markov's inequality to prove chebyshev's inequality.

$$\begin{aligned} P(|X - \mu| \geq k\sigma) &= P((X - \mu)^2 \geq k^2\sigma^2) \\ &\leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} \\ &= \frac{\sigma^2}{k^2\sigma^2} \\ &= \frac{1}{k^2} \end{aligned}$$

with the first line being true for both $|X - \mu|$ and $k\sigma$ are positive, and squaring is an increasing function.

8 Generating functions

For a sequence of real numbers $\{a_k\}$, then the generating function of the sequence is the power series

$$A(z) = \sum_{k=0}^{\infty} a_k z^k$$

This series will converge for $z \in C \subseteq \mathbb{R}$, the domain of A. We are only interested for $C \neq \{0\}$ because zero is always in C.

By the uniqueness of power series, $A(z)$ uniquely defines $\{a_k\}$ and vice versa. We can extract the sequence $\{a_k\}$ from $A(z)$ by

- Expanding $A(z)$ as a power series and extracting the terms
- Differentiation

  $$k! a_k = \frac{d^k}{dz^k} A(z)|_{z=0}$$

  which is proven by the taylor expansion of $A(z)$ at $z = 0$.

8.1 Probability generating function

Define the probability generating function for a non-negative, integer valued random variable X with a pmf $p_X(k)$

$$P_X(z) = \sum_{k=0}^{\infty} p_X(k) z^k = E[z^X]$$

This will always converge when $|z| \leq 1$.

$$\begin{aligned} |P_X(z)| &\leq \sum_{k=0}^{\infty} |p_X(k) z^k| \\ &\leq \sum_{k=0}^{\infty} |p_X(k)| \\ &= \sum_{k=0}^{\infty} p_X(k) \\ &= 1 \end{aligned}$$

hence it is finite when $|z| \leq 1$, which implies that it converges (lmao).

To invert and find the pmf from the pgf, we have

$$p_X(k) = \frac{P_X^{(k)}(0)}{k!}$$

We can also find the probability that X is even by

$$\begin{aligned} P_X(z) + P_X(-z) &= \sum_{k=0} p_X(k)(z^k + (-z)^k) \\ &= 2 \sum_{j=0} p_X(2j) z^{2j} \\ P(X \text{ even}) &= \sum_{k=0} p_X(2k) \\ &= \frac{P_X(1) + P_X(-1)}{2} \\ &= \frac{1 + P_X(-1)}{2} \end{aligned}$$

and the probability that X is odd as one minus the probability that X is even

$$P(X \text{ odd}) = \frac{P_X(1) - P_X(-1)}{2} = \frac{1 - P_X(-1)}{2}$$

The uniqueness theorem states that the pgf determines the distribution of the random variable. If two random variables have the same pgf, they have the same distribution. In general, for a non-negative discrete distribution with integers, its distribution is completely described by its pgf, cdf, and pmf.

The properties of the pgf

- $P_X(1) = 1$, because it is a sum of the pmfs
- $P'_X(1) = E[X]$
- $P''_X(1) = E[X(X-1)]$ and that in general, the kth derivative is this expectation of falling powers
- $V[X] = P''_X(1) + P'_X(1) - P'_X(1)^2$
- $P_X(z) = E[z^X] = E[E[z^X|Y]]$ from law of total expectations
- The convolution theorem. For independent random variables X and Y, then
$$P_{X+Y}(z) = P_X(z) + P_Y(z)$$
from the definition. This can generalize to the sum of n independent random variables.

8.1.1 Common pgfs

For $X \sim B(n,p)$, then the pgf is
$$P_X(z) = (1-p+pz)^n$$
from simple substitution and the binomial theorem, which is defined/converges for all $z \in \mathbb{R}$. By letting $n = 1$, the pgf of the bernoulli random variable is
$$P_X(z) = 1 - p + pz$$

For $X \sim Pn(\lambda)$, the pgf is
$$P_X(z) = e^{-\lambda(1-z)}$$
using the taylor series of $e^{\lambda z}$, which is defined for all $z \in \mathbb{R}$.

For $X \sim Nb(r,p)$, its pgf is
$$\begin{aligned} P_X(z) &= \sum \binom{-r}{x} p^r (p-1)^x z^x \\ &= p^r \sum \binom{-r}{x} ((p-1)z)^x \\ &= p^r (1 + ((p-1)z))^{-r} \\ &= p^r (1 - (1-p)z)^{-r} \end{aligned}$$

using the general binomial theorem, with the domain $|z(1-p)| < 1$. By letting $r = 1$, we get the geometric distribution pgf

$$P_X(z) = \frac{p}{1-(1-p)z}$$

8.2 Moment generating function

The moment generating function allows us to represent continuous random variables. Define the mgf for a random variable X to be

$$M_X(t) = E[e^{tX}]$$

for $t \in T$, where $T = \{t\colon E[e^{tX}] < \infty\}$ is the set of all values of t where the mgf is finite.

The properties of the mgf

- $M_X(0) = 1$, hence $0 \in T$.
- $M_X(t) = \sum_k E[X^k]\frac{t^k}{k!}$, so the higher order moments of X are uniquely determined by the mgf through differentiation

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

  . Alternatively, We can also find all $\{\mu_k\}$ directly by writing out $M_X(t)$ as a power series in t then matching the terms
- The linear property of the mgf

  $$M_{aX+b}(t) = E[e^{(aX+b)t}] = e^{bt}E[e^{atX}] = e^{bt}M_X(at)$$

- $M_X(t) = E[E[e^{tX}|Y]]$
- The convolution theorem also applies for mgf, for independent X and Y

  $$M_{X+Y}(t) = M_X(t)M_Y(t)$$

- Mgf doesn't exist for some random variables, that is, the set $T = \{0\}$ for its mgf is defined only at zero. This happens for some heavy tail random variables (cauchy, pareto).
- The central moments generating function is are

  $$N_X(t) = E[e^{(X-\mu)t}] = e^{-\mu t}M_X(t)$$

  and we can get the central moments by differentiating $N_X(t)$

- The uniqueness theorem. If the mgf exists for a neighbourhood around 0, then it uniquely determines the distribution of X which is the distribution function $F_X(x)$.
- There exists RV with all finite moments but infinite mgf everywhere except zero

If X is a discrete random variable with non-negative integers, then the conversions between the pgf and mgf are

$$M_X(t) = P_X(e^t) \quad P_X(z) = M_X(\log z)$$

8.2.1 Common mgfs

For $X \sim \exp(\alpha)$, then

$$M_X(t) = \int_0^\infty e^{xt} a e^{-ax}\, dx = \frac{a}{a-t}$$

only when $t < a$. So $T = \{t < a\}$. The moments are

$$M_X(t) = \frac{1}{1-1/a} = \sum (t/a)^k \implies \mu_k = \frac{k!}{a^k}$$

If $X \sim Pn(\lambda)$, we can just use the substitution $z = e^t$

$$M_X(t) = P_X(e^t) = e^{\lambda(e^t - 1)}$$

which is defined for all $T = (-\infty, \infty)$.

If $X \sim \gamma(r, \alpha)$, the mgf is

$$M_X(t) = \left(\frac{a}{a-t}\right)^r$$

with domain $T = (-\infty, a)$.

If $Z \sim N(0,1)$, the mgf is

$$M_Z(t) = e^{t^2/2}$$

by completing the square. The domain is $T = \mathbb{R}$.

We ignore the cumulant generating function.

8.3 Laplace Transform

The laplace transform for a random variable is

$$L_X(t) = M_X(-t)$$

which exists for all $t > 0$ when $P(X \geq 0) = 1$, for that $P(X \geq c) = 1$. This has similar properties as the mgf but is invertible to the pdf.

To recover the distribution from the mgf, we either

- Recognize the form of the mgf through the uniqueness theorem
- If X is integer valued, we can get $P_X(z)$ from the mgf and recover the pmf through differentiation
- If X is non-discrete. If RV is non-negative, use the inverse laplace transform, otherwise, invert the characteristic function

The characteristic function is

$$ch_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx$$

which is the fourier transform of the pdf. This gf is defined for all $T = \mathbb{R}$ because it is bounded by the modulus.

8.4 Limiting Distribution

Define the convergence in distribution by: a sequence of random variables Y_n converges to Y in distribution, and $Y_n \to Y$ if the two cdfs converges pointwise

$$\lim_{n\to\infty} F_{Y_n}(x) = F_Y(x)$$

for all x where $F_Y(x)$ is continuous (or left continuous).

The equivalent mgf version states that $Y_n \to Y$ if and only if their moment generating functions converges pointwise

$$\lim_{n\to\infty} M_{Y_n}(t) = M_Y(t)$$

for all $t \in (-\delta, \delta)$.

8.5 Law of large numbers

For independent and identically distributed random variables X_i with mean μ, let $S_n = \sum_{i=1}^{n} X_i$, then by the law of large numbers

$$\frac{S_n}{n} \to \mu$$

The RHS is interpreted as the deterministic random variable that is always equal to μ.

Proof. Assume that the mgf for X_i exists for a neighbourhood around 0.

Then by taylor expansion

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} E[X^k]\frac{t^k}{k!} \\ &\approx 1 + E[X]t + E[X^2]t^2/2 + o(1/n^2) \\ M_X(t/n) &\approx 1 + E[X]t/n + o(1/n) \\ &= 1 + \mu t/n \end{aligned}$$

by ignoring lower order terms $o(1/n)$ for we are taking $n \to \infty$.

By the convolution theorem

$$\begin{aligned} S_n/n &= \sum X_i/n \\ M_{S_n/n}(t) &= \prod M_{X_i/n}(t) \\ &= (M_X(t/n))^n \\ &= (1 + \mu t/n)^n \\ &\approx e^{\mu t} \\ &= M_\mu(t) \end{aligned}$$

when $n \to \infty$ and for all $t \in (-\delta, \delta)$ by assumption.

Hence by the mgf definition of distribution convergence, $S_n/n \to \mu$. □

The hypothesis of the law of large numbers can be weakened. The theorem implies that the random variable $\bar{X} = S_n/n$ is an unbiased estimator for μ

$$P(|\frac{S_n}{n} - \mu| < \epsilon) = 1$$

as $n \to \infty$ for any $\epsilon > 0$.

8.6 Central limit theorem

Vaguely speaking, the central limit theorem states that the sum of IIDRV converges to a normal distribution.

Let X_i be IIDRV with $E[X_i] = \mu$ and $V[X_i] = \sigma^2$. Then let $S_n = \sum_{i=1}^n X_i$, the central limit theorem states that

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{V[S_n]}} = \frac{S_n - n\mu}{\sigma\sqrt{n}} \to N(0,1) \quad n \to \infty$$

alternatively for large n

$$S_n \approx N(n\mu, n\sigma^2) \quad S_n/n \approx N(\mu, \sigma^2/n)$$

Proof. Assume that the mgf for X_i exists for a small $(-\delta, \delta)$.

Then

$$\begin{aligned} M_{Z_n}(t) &= (M_{X-\mu}(\frac{t}{\sigma\sqrt{n}}))^n \\ &= (M'_{X-\mu}(0) + \frac{t}{\sigma\sqrt{n}} M'_{X-\mu}(0) + (\frac{t}{\sigma\sqrt{n}})^2 M''_{X-\mu}(0)/2 + o(1/n)) \\ &\approx (1 + \frac{\sigma^2}{2}\frac{t^2}{\sigma^2 n})^n \\ &\to e^{t^2/2} \end{aligned}$$

by ignoring higher orders of $o(1/n)$ for $n \to \infty$. This is the mgf for $N(0,1)$, and because both have domains $(-\delta, \delta)$, by the definition of distribution convergence, $Z_n \to N(0,1)$ $\square$

We can use the CLT to show that if X_i is a series of independent exponential distributions with the same parameter α, then

$$\sum_i^n X_i \sim \gamma(n, \alpha) \approx N(n/\alpha, n/\alpha^2)$$

Additionally, if X_i is a series of independent bernoulli random variables with the same parameter p, then

$$\sum_i^n X_i \sim B(n, p) \approx N(np, np(1-p))$$

but these are vague statements because both sides have n.

In practice, try to transform the sequence of random variables to the standardized form then directly apply the CLT instead of transforming the limit distributions, because we haven't done transformations of random variable limiting distributions.

9 Stochastic processes

A stochastic process is a sequence of random variables $\{X_t\}$ for $t \in T$ indexed by time. Time in this context may be continuous or discrete. Let each X_t be random variables defined on the same sample space, and take values in a set S, called the state space of the stochastic process. The set T is the index set usually representing time, but can be a spatial variable like displacement.

The state space S may be discrete or continuous, depending on whether X_t is discrete or continuous. The process are called discrete-space or continuous-state stochastic processes.

The index set T can also be discrete or continuous, we call these processes discrete-time and continuous-time stochastic processes. We can denote discrete time processes by $\{X_n\}$ and continuous time processes by $\{X(t)\}$.

The random variables in the process are just random variables, they are not defined based on previous random variables in the process (we should think of them independently), although their distributions might be correlated/dependent on the result of previous random variables.

A sequence of bernoulli trials is an example of a discrete-space, discrete-time random stochastic process. Let X_n be the number of successes after n trials, then it represents a such a random process with the state space $\mathbb{N}$ and index set $\mathbb{N}$. Note that we assume that $\{X_n\}$ are defined on the same sample space of independent bernoulli trials.

A discrete-state, continuous-time random process that is an analogue of the sequence of bernoulli trials is the Poisson process. The underlying sample space has points occur randomly in continuous time. The poisson process is defined by $\{N(t)\}$ which are random variables on the number of points occuring in the interval $[0, t]$. Its state space is the natural numbers $\mathbb{N}$ and its index set are the positive reals $\mathbb{R}_+$.

Stochastic processes where random variables are independent are really boring. The processes where there are dependence in random variables we'll analyze are: poisson processes, discrete time markov chains.

9.1 Poisson process

Some postulates about the poisson process

- Within a small enough time, there will be either one event or no event. We cannot have two events happening at the same time. The probability that there is at least one event in this $(t, t+h]$ should be proportional to the interval length

$$P(N(t+h) - N(t) \geq 1) = \lambda h + o(h)$$
$$P(N(t+h) - N(t) \geq 2) = o(h)$$

- The number of events in disjoint time intervals are independent to each other, namely for (s_1, t_1) and (s_2, t_2)

$$N(t_1) - N(s_1), N(t_2) - N(s_2)$$

  are independent.

The poisson process can also be generated by taking the limit of independent continuous bernoulli trials.

Properties of the poisson process

- The number of events in $[0, t]$ and the random variables are poisson distributed $N(t) \sim Pn(t\lambda)$.
- The waiting time until the first event is exponentially distributed with parameter λ. This also applies to the waiting time until the next event due to the independent postulate
- The waiting time until the rth event is a gamma distribution with parameters r and λ

A poisson process does not have evenly spaced out events for the waiting time is an exponential distribution.

Furthermore

- In general, the number of occurrences in an interval of size t has a poisson distribution of parameter $t\lambda$
$$N(a+t) - N(a) \sim Pn(t\lambda)$$
- For non-overlapping intervals in a poisson process, the number of occurrences in one is independent to the number of occurrences in another
- For overlapping intervals, the number of occurrences are no longer independent and will be correlated

The properties of the poisson process stems from the continuous memoryless property of exponential waiting times.

Poisson processes can model: car passes, patient arrivals in ER, number of calls in call centers.

9.2 Discrete time markov chain

To motivate the properties of markov chains, consider discrete random variables X_i and integers x_i from their sample space, the joint pmf is

$$P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n)$$

which can be rephrased as

$$\begin{aligned} P(X_0 = x_0, \dots, X_n = x_n) &= P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\quad \times P(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\quad \times P(X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}, \dots) \\ &\quad \times \cdots \times P(X_0 = x_0) \end{aligned}$$

using successive conditional probabilities on intersection of events.

The simplest case of the joint pmf is if the sequence is independent, then the RHS is simply

$$P(X_n = x_n) \dots P(X_0 = x_0)$$

for there are no conditional aspects. The next simplest case is a markov chain. For a markov chain, the RHS would be

$$P(X_n = x_n | X_{n-1} = x_{n-1}) P(X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}) \dots P(X_0 = x_0)$$

due to the markov property where the conditional probabilities only depend on previous random variable value.

9.2.1 Definition

A discrete time markov chain is a stochastic process with discrete index time $\mathbb{N}$ and a countable state space S. To simplify the notation, map S to $\mathbb{N}$. For any $i \in S$, let $P(X_n = i)$ to be the probability that the process has state i at time n. The process is a markov chain if for all $n \geq 0$, all $j, i_n, \dots i_0 \in S$

$$P(X_{n+1} = j | X_n = i_n, \dots X_0 = i_0) = P(X_{n+1} = j | X_n = i_n)$$

which is the markov property. It states that the next state only depends on the current state.

9.2.2 Transition matrix

Define the transition probability from state i to j at time n in a markov chain as

$$P(X_{n+1} = j | X_n = i)$$

which are the one-step transition probabilities of the markov chain. If this does not depend on the time n, then the markov chain is said to be time homogeneous markov chain. More precisely, a markov chain is time homogeneous if for all $n \geq 0$ and $i, j \in S$

$$P(X_{n+1} = j | X_n = i) = P(X_1 = j | X_0 = i)$$

Under a time homogeneous markov chain, we can represent all transition probabilities in a one-step transition matrix where

$$P_{ij} = P(X_{n+1} = j | X_n = i) \quad P = [P_{ij}]_{i,j \in S}$$

because S is countable, we assume a natural numbers indexing for the state space i, j. Properties of the transition matrix are

- All elements are non-negative, they are probabilities
- Each row of the transition matrix sums up to one because the rows forms a conditional pmf given the current state, which must sum up to one

A square matrix with the above 2 properties is called a stochastic matrix.

In general for m steps, the m-step transition probabilites on a time homogeneous markov chain also forms a transition matrix (a stochastic matrix too) with

$$P_{ij}^{(m)} = P(X_{n+m} = j | X_n = i)$$

Notice that when $m = 1$, this is the one-step transition matrix

$$P_{ij}^{(1)} = P_{ij}$$
$$P^{(1)} = P$$

for two matrices are equal when all their elements are equal.

To compute the m-step transition matrix, we have the theorem

$$P^{(m)} = P^m$$

where P is the one-time transition matrix.

Proof. We induct on m. We've already proven the result when $m = 1$.

Given that $P^{(m-1)} = P^{m-1}$ with $m > 1$ so that $n + m - 1 > n$, we have that

$$\begin{aligned}
P_{ij}^{(m)} &= P(X_{n+m} = j | X_n = i) \\
&= \sum_k P(X_{n+m} = j | X_{n+m-1} = k, X_n = i) P(X_{n+m-1} = k, X_n = i) \\
&= \sum_k P(X_{n+m} = j | X_{n+m-1} = k) P(X_{n+m-1} = k | X_n = i) \\
&= \sum_k P_{jk}^{(1)} P_{ki}^{(m-1)} \\
&= (PP^{m-1})_{ij} = P_{ij}^m \\
P^{(m)} &= P^m
\end{aligned}$$

by the law of total conditional probability, the markov property, time homogenuity, and induction case. □

9.2.3 Transition diagram

A transition diagram is another way to represent a transition matrix. The nodes are states of the markov chain, while an arrow from a state to another (including itself) represents the transitional probability from that state to another. We ignore the arrows representing a transitional probability of zero.

9.2.4 Long run behavior of time homogeneous markov chains

The initial state distribution is the pmf of X_0

$$p_{X_0}(x) = P(X_0 = x)$$

The law of total probabilities gives the m-step distribution pmf X_m given an initial distribution

$$p_{X_m}(y) = P(X_m = y) = \sum_x P(X_m = y | X_0 = x) P(X_0 = x)$$

Let π_n be a row vector denoting the pmf of X_n, therefore

$$\begin{aligned} \pi_0 &= (p_{X_0}(x_1), p_{X_0}(x_2), \dots) \\ \pi_m &= \pi_0 P^m \end{aligned}$$

which can be proven by induction on m using the one step updating of π_n.

If the n step transition matrix has a limit

$$P^n \to Q \quad n \to \infty$$

where all the rows of Q are the same, then no matter what the initial distribution π_0 is, the limiting distribution is

$$\begin{aligned} \lim_{n\to\infty} \pi_n &= \pi \\ \lim_{n\to\infty} P_{ij}^n &= \pi_j \end{aligned}$$

where π is the rows of Q. This can be interpreted as: no matter what initial state you started with, in a long time, your state in the future has the same limiting distribution π.

Assuming that the state space is finite, and that the transition matrix always converges with the same rows, then the conditional probability to end up at state j is independent to the initial state/distribution. The limiting distribution π is also known as the equilibrium distribution, long-run distribution, or the stationary distribution. This property does not always occur.

To find the limiting distribution, notice that

$$\lim_{m\to\infty} P_{ij}^{(m)} = \lim_{m\to\infty} \sum_k P_{ik}^{(m-1)} P_{kj}$$
$$\pi_j = \sum_k \pi_k P_{kj}$$

from the law of total probability and the limiting distribution when $m \to \infty$. Written in vector form, this is the equilibrium equation for a markov chian

$$\pi = \pi P \quad \pi^T = P^T \pi^T$$

noting that π is the left-eigenvector of P with eigenvalue 1.

The additional property needed to find the limiting distribution is that the limiting distribution is a pmf, so all entries sum up to 1. Hence when the state space is finite, the system of equations we need to solve to find π is

$$\pi P = \pi$$
$$\sum_j \pi_j = 1$$

which can be done by any linear algebra techniques like row reduction and Gaussian Elimination, but transpose the matrix P beforehand. Due to our assumptions, the matrix equation always has exactly one redundant equation with $N-1$ linearly independent equations, so we can find a unique solution for π by replacing the Nth equation with a normalizing equation.

This implies that the limiting distribution can be interpreted as

- The stationary distribution where applying the one-step transition matrix does nothing
$$\pi = \pi P$$
- The limiting distribution when applying a large number of transition matrices
$$\pi_j = \lim_{n\to\infty} P_{ij}^{(n)}$$
- The ergodic interpretation is that given a set of sample paths with total probability one (or a large set of sample path), the proportion of time that the process spends in state j is π_j.

all three interpretations are equivalent in this subject.