Survey of Probability Concepts

Matthew He

December 9, 2024

This is meant to be a quick survey of important concepts in introductory probability theory. It goes faster and more in-depth than elementary probability class, but not as technical as those pure math courses, which might be good for people just getting into research like me.

Reference textbook:
Introduction to Probability by David F. Anderson, Timo Seppäläinen, and Benedek Valkó (2017)
Mathematical Statistics with Applications by Dennis Wackerly, William Mendenhall, and Richard L. Scheaffer (2007)
MATH447 Stochastic Process notes by Jana Kurrek

1 Experiments with random outcomes

1.1 Ingredients of a Probability Model

Definition 1. *These are ingredients of a probability model.*

- *The **sample space** Ω is the set of all possible outcomes of the experiment. Elements of Ω are called **sample points** and typically denoted by ω.*
- *Subsets of Ω are called **events**. The collection of events in Ω is denoted by $\mathcal{F}$.*
- *The **probability measure** (also called **probability distribution** or simply **probability**) P is a function from $\mathcal{F}$ into the real numbers. Each event A has a probability of P(A), and P satisfies the axioms on the right.*

*The triple $(\Omega, \mathcal{F}, P)$ is called a **probablity space**. Every mathematically precise model of a random experiment or collection of experiments must be of this kind.*

Kolmogorov Axioms (early 1930s)

1. $0 \leq P(A) \leq 1$ for event A.
2. $P(\Omega) = 1$ and $P(\emptyset) = 0$.
3. If A_1, A_2...... is a sequence of pairwise disjoint events $(E_i \cap E_j = \emptyset)$ for $i \neq j$, then $P(E_1 \cup E_2 \cup E_3......) = \sum_{i=1}^{\infty} P(E_i)$ or
$$P(\bigcup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} P(A_i)$$

*Axiom 3 can also be stated in terms of finite union of events as $P(E_1 \cup E_2 \cup E_3......) = \sum_{i=1}^{n} P(E_i)$
*Axiom 3 states that we can calculate probability of an event by summing up probabilities of its disjoint decomposed events.

1.2 Random Sampling

Theorem 1 (Random Sampling)**.** *Let S be a finite sample space with N equally likely events and let E be an event in S. Then*

$$P(E) = \frac{n}{N}$$

This important theorem can reduce the problem of finding probabilities to a counting problem.

Counting Rule 1: Multiplication Rule

Sampling with replacement, order matters. Consider k sets, Set 1 and Set 2 ... Set k. Set 1 has n_1, Set 2 has n_2 ... Set k has n_k distinct objects. Then the number of ways to form a set by choosing one object from each set is $n_1 n_2 ... n_k$.

Counting Rule 2: Factorial Rule

Sampling without replacement, order matters. The number of ways to arrange n distinct objects is $n!$.
$0! = 1$and $1! = 1$

Counting Rule 3: Permutation Rule

Sampling without replacement, order matters. The number of ways to arrange r chosen from n distinct object at a time without replacement, where the order matters, is known as **permutations** of n objects taken r at a time.
It is given by: ${}^nP_r = \dfrac{n!}{(n-r)!}$

Counting Rule 4: Combination Rule

Sampling without replacement, order irrelevant. The number of ways to select r object from n distinct total objects at a time without replacement, where order does not matter, is known as **combination** of n objects taken r at a time. It is given by: $\binom{n}{r} = \dfrac{n!}{(n-r)!r!}$

$\binom{n}{r}$ is also called binomial coefficient.

1.3 *Consequences of the rules of probability*

Decomposing an event

If $A_1, A_2, ...$ are pairwise disjoint events and A is their union, then $P(A) = P(A_1) + P(A_2) + ...$ Calculation of the probability of a complicated event A almost always involves decomposing A into smaller disjoint pieces whose probabilities are easier to find. Both finite and infinite decomposition is possible.

Theorem 2 (Events and complements). *For any event* A ,$P(A)^c = 1 - P(A)$

Theorem 3. $P(\emptyset) = 0$

Theorem 4. $P(A \cup B^C) = P(A) - P(A \cap B^c)$

Proof:
Express A as the union of disjoint events as $A = (A \cap B^C) \cup (A \cap B)$
$P(A) = P(A \cap B^C) + P(A \cap B)$
by Axiom 3,
$\Rightarrow P(A \cup B^C) = P(A) - P(A \cap B^c)$

Theorem 5 (Monotonicity of probability). *If* $A \subset B$ *then* $P(A) \leq P(B)$

Proof:
$B = A \cup (A^C \cap B)$, $P(B) = P(A) + P(A^C \cap B)$ – Axiom 3
As $P(A^C \cap B) \geq 0$ – Axiom 1, $\Rightarrow P(B) \geq P(A) or P(A) \leq P(B)$

Theorem 6 (Inclusion-exclusion formulas).

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

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ -P(B \cap C) + P(A \cap B \cap C) \end{aligned}$$

General Formula:

$$\begin{aligned} P(A_1 \cup ... \cup A_n) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2}) \\ + \sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ - \sum_{1 \le i_1 < i_2 < i_3 \le i_4 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap A_{i_4}) \\ + ... + (-1)^{n+1} P(A_{i_1} \cap ... \cap A_{i_n}) \\ = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < ... < i_k \le n} P(A_{i_1} \cap ... \cap A_{i_k}) \end{aligned}$$

Proof:
$A \cup B = (A \cap B^C) \cup (A \cap B) \cup (A^C \cap B)$
$P(A \cup B) = P(A \cap B^C) + P(A \cap B) + P(A^C \cap B)$
$P(A \cup B) = (P(A) - P(A \cap B)) + P(A \cap B) + (P(B) - P(A \cap B))$
By Theorem 3, therefore $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Proof :
Write E as the union of its simple events (elementary outcomes).
$E = \cup_{i=1}^{n} E_i$
As the simple events are disjoint,
$P(E) = \sum_{n}^{i=1} P(E_i)$ by Axiom 3.
Similarly, $S = \cup_{i=1}^{N} E_i$ and $P(S) = \sum_{N}^{i=1} P(E_i)$ by Axiom 3.
Since all event E_i are equally likely (have the same probability of occurrence)
$\sum_{i=1}^{N} P(E_i) = NP(E_i)$ also $P(S) = 1$ by Axiom 2
Hence, $NP(E_i) = 1$ and $P(E_i) = \frac{1}{N}$
Therefore, $P(E) = \sum_{i=1}^{n} P(E_i) = \sum_{i=1}^{n} \frac{1}{N} = \frac{n}{N}$

Side notes ssss

1.4 *Continuity of the probability measure*

Theorem 7. *Suppose we have an infinite sequence of events $A_1, A_2, \ldots$ that are nested increasing: $A_1 \subset A_2 \subset \cdots \subset A_n \subset \cdots$. Let $A_\infty = \cup_{k=1}^{\infty} A_k$ denote the union. Then*

$$\lim_{n \to \infty} P(A_n) = P(A_\infty).$$

Another way to state this is as follows: if we have increasing events then the probability of the union of all the events(the probability that at least one of them happens) is equal to the limit of the individual probabilities.

Proof. To take advantage of the additivity axiom of probability, we break up the events A_n into disjoint pieces. For $n = 2, 3, 4, \ldots$ let $B_n = A_n \backslash A_{n-1}$.
Now we have the disjoint decomposition of A_n as

$$A_n = A_{n-1} \cup B_n = A_{n-2} \cup B_{n-1} \cup B_n = \cdots = A_1 \cup B_2 \cup \cdots \cup B_n.$$

Taking union of all the events A_n gives us the disjoint decomposition

$$\cup_{n=1}^{\infty} A_n = A_1 \cup B_2 \cup B_2 \cup B_3 \cup \cdots$$

By the additivity of probability, and by expressing the infinite series as the limit of partial sums,

$$P(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} (P(A_1) + P(B_2) + \cdot + P(B_n)) = \lim_{n\to\infty} P(A_n).$$

Q.E.D

Recall from calculus that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous at x if and only if for each sequence of point $x_1, x_2, \ldots$ that converge to x, we have $f(x_n) \to f(x)$ as $x \to \infty$. In a natural way increasing sets A_n converge to their union A_∞, because the difference $A_\infty \backslash A_n$ shrinks away as n $n \to \infty$.

1.5 *Measurability*

Every subset of a discrete sample space Ω is a legitimate event. For example, the sample space of flipping a single coin is $\Omega = \{H, T\}$ and the collection of events is $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$, which is exactly the collection of all subsets of Ω, namely the power set of Ω.

This all seems very straightforward. But there can be good reasons to use smaller collection $\mathcal{F}$ of events. It can be useful for modeling purposes, and solve the technical problems with uncountable sample spaces preventing us from taking the $\mathcal{F}$ as the power set.

To put the theory on a sound footing, we extend the axiomatic framework to impose the following requirements on the collection of event $\mathcal{F}$:

Definition 2 ($\sigma - algebra$)**.** *Any collection $\mathcal{F}$ of sets satisfying the following properties is call $\sigma - algebra$ or $\sigma - field$.*

1. *the empty set $\emptyset$ is a member of $\mathcal{F}$,*
2. *if A is in $\mathcal{F}$, then A^c is also in $\mathcal{F}$,*
3. *if $A_1, A_2, A_3, \cdots$ is a sequence of events in $\mathcal{F}$, then their union $\cup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$.*

The members of a $\sigma - algebra$ are called *measurable* sets. The properties of a $\sigma - algebra$ imply that countably many applications of the usual set operations to events is a safe way to produce new events.

Fortunately all reasonable sets and functions encountered in practice are measurable.

Another aspect of the collection $\mathcal{F}$ of events is that it can represent *information*.

2 *Conditional probability and independence*

2.1 *Condition probability*

Definition 3 (Conditional probability)**.**
Let B be an event in the sample space Ω such that $P(B) > 0$. Then for all events A the ***conditional probability*** *of A given B is defined as*

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

Fact:
Let B be an event in the sample space Ω such that $P(B) > 0$. Then, as a function of the event A, the conditional probability $P(A|B)$ satisfies the Kolmogorov Axioms. Especially, we have $P(\cup_{i=1}^{\infty} B_i|A) = \sum_{i=1}^{\infty} P(B_i|A)$ where, $B_i \cap B_j = \emptyset$ for $i \neq j$

Theorem 8. *Suppose that we have an experiment with* ***finitely many equally likely outcomes*** *and B is not the empty set. Then, for any event A*

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

Theorem 9 (Multiplication rule for n events)**.** *If $A_1, ..., A_n$ are events and all the conditional probabilities below make sense then we have*

$$P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2A_1) \cdots P(A_n|A_1 \cdots A_{n-1})$$

Or simply, we have

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

Hints:

1. If required to find $P(A \cap B)$, look for either $P(A)$ or $P(B)$ and one of the conditional probabilities.
2. In word problems "of those that" implies a conditional probability.
3. Do not confuse "and" with "given that"

Nots: This implies that problems involving the intersection of several events can be simplified to a great extent by conditioning backwards.

Three special cases of connditional probability

1. Let A and B be two disjoint events, then, $A \cap B = \emptyset$ and $P(B|A) = 0$, since $P(A \cap B) = 0$
2. Let A and B be two events, such that $B \subset A$. Then, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)}{P(A)}$
3. Let A and B be two events, such that $A \subset B$. Then, $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A)}{P(A)} = 1$

Calculating probability by decomposition

For example, a general version of the reasoning can be:

$$P(A) = P(AB) + P(AB^c) = P(A|B)P(B) + P(A|B^c)P(B^c). \quad (1)$$

The idea is the decomposition of a complicated event A into disjoint pieces that are easier to deal with. Above we used the pair $\{B, B^c\}$ to split A into two pieces. $\{B, B^c\}$ is an example of a *partition*.

Definition 4 (Partition). *A finite collection of event $\{B_1, \ldots, B_n\}$ is a* ***partition*** *of Ω if the sets B_i are pairwise disjoint and together they make up Ω. That is , $B_iB_j = \emptyset$ whenever $i \neq j$ and $\cup_{i=1}^n B_i = \Omega$*

This equation is true for the same reason as the eq. (1).
Namely, set algebra gives

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^n B_i\right) = \bigcup_{i=1}^n AB_i$$

$$P(A) = P(\bigcup_{i=1}^n AB_i)$$

Theorem 10 (The Law of Total Probability). *Suppose that $B_1, \ldots, B_n$ is a partition of Ω with $P(B_i) > 0$ for $i = 1, \ldots, n$. Then for any event A we have*

$$P(A) = \sum_{i=1}^n P(AB_i) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

More generally, $P(A) = \mathbb{E}[P(A|X)]$.

Bayes' formula for two events.
For events A and B,

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

post = likehood * prior/marginization

Definition 5 (General Version of Bayes' Formula). *Let $B_1, \ldots, B_n$ be a partition of the sample space Ω such that each $P(B_i) > 0$. Then for any event A with $P(A) > 0$, and any $k = 1, \ldots, n$.*

$$P(B_k|A) = \frac{P(AB_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_j (A|B_j)P(B_j)}$$

3 Random Variables

3.1 A First Look

In addition to basic outcomes themselves, we are often interested in various numerical values derived from the outcomes.

Definition 6 (Random Variable). *Let Ω be a sample space. A* ***random variable*** *is a* ***function*** *from Ω into the real number.*

Some conventions:
Random variables, not variables but functions, are usually denoted by capital letters such as X, Y and Z. The value of a random variable X at sample point ω is $X(\omega)$.

Definition 7 (Probability Distribution). *Let X be a random variable. The* ***probability distribution*** *of the random variable X is the collection of probabilities $P(X \in B)$ for sets B of real numbers.*

The probability distribution of a random variable is an assignment of probability to subsets of $\mathbb{R}$ that satisfies again the axioms of probability.

Definition 8 (Discrete Random Variable). *A random variable X is a* ***discrete random variable*** *is there exists a finite or countably infinite set $\{k_1, k_2, k_3, \ldots\}$ of real numbers such that*

$$\sum_i P(X = k_i) = 1$$

where the sum ranges over the entire set of points $\{k_1, k_2, k_3, \ldots\}$.

That said, if the range of the random variable X is finite or countably infinite, then X is a discrete variable. Those k for which $P(X = k) > 0$ are the possible values of X.

3.2 *Different kinds of random variables*

Bernoulli random variable

A bernoulli random variable is related to the occurrence (or non-occurrence) of a certain event E. If event E occurs, then X = 1; otherwise, X = 0.

Examples:

- Making application for a job
- Tossing a coin
- Getting tested for Covid-19

Binomial random variable

Before introducing the Binomial Distribution, we need to define the Binomial Experiment.

Definition 9 (Binomial experiment)**.** *A experiment is called a binomial experiment if it satisfies the following conditions:*

- *it consists of n independent Bernoulli trials*
- *the probability of success p remain constant from trial to trial*
- *We are interested in x successes out of n trials.*
  Where $x = 0, 1, 2, \ldots, n$.

Bernoullil trial is a trial with two outcomes, success and failure. A success is the outcome of interest. Let's denote the probability of success as p.

Definition 10 (Binomial random variable)**.** *Let X be the random variable that counts the number of successes in n Bernoulli trials, where the probability of success in each trial is p. Then X is a Binomial random variable with the parameters n and p, and it's probability distribution is called the Binomial distribution. We say $X \sim B(n, p)$.*

Theorem 11. *For a binomial random variable $X \sim B(n, p)$, the probability mass function is given by*

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

x measures the number of successes in n independent Bernoulli trials.

Geometric random variable

Sometimes we are interested in the number of trials needed to get the first success.

Definition 11 (Geometric random variable)**.** *A random variable X is said to have a geometric distribution with parameter p if it's probability mass function is given by*

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

where $x = 1, 2, 3, \ldots$, and $0 < p < 1$.

The random variable X is the number of trials at which the first success occurs.

Theorem 12. *Let X be a random variable with a geometric distribution with parameter p. Then*

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

where $x = 1, 2, 3, \ldots$ and $0 < p < 1$.

Note that

- The Binomial random variable gives the number of successes in the fixed number of trials.
- The Geometric random variable gives the number of trials at which the first success occurs, where the number of trials is not fixed.

Negative Binomial random variable

Definition 12 (Negative binomial random variable)**.** *The negative binomial random variable X gives the trial on which the rth success occurs in a sequence of independent Bernoulli trials. Each trial has two possible outcomes, success and failure. THe probability of success remains constant from trial to trial.*

Theorem 13. *Let X be negative binomial random variable, then*

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

where $x = r, r+1, r+2, \ldots$ and $0 < p < 1$.

Poisson random variable

Definition 13 (Poisson random variable)**.** *A random variable X is said to have a Poisson distribution with parameter λ if it's probability mass function is given by*

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

where $x = 0, 1, 2, \ldots$ and $\lambda > 0$.

Theorem 14. *Let $X \sim Binom(n, p)$, where $n \to \infty$ and $p \to 0$ and $np = \lambda$ (constant). Then*

$$P(X = x) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

Proof.

$$\lim_{n\to\infty} P(X = x) = \lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} \tag{2}$$

$$= \lim_{n\to\infty} \binom{n}{x} (\frac{\lambda}{n})^x (1 - \frac{\lambda}{n})^{n-x} \tag{3}$$

$$\tag{4}$$

For $x = 0, 1, 2, \ldots$ and $\lambda > 0$.

Hypergeometric random variable

Definition 14 (Hypergeometric random variable). *The random variable X is a hypergeometric random variable with parameters A hypergeometric random variable X represents the number of successes in n draws without replacement from a finite population of size N that contains exactly K successes. It models situations where sampling is done without replacement, and each draw changes the probabilities of subsequent draws.*

The probability mass function is:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where $k = 0, 1, 2, \ldots, n$ *and* $0 \leq k \leq \min(n, K)$.

3.3 Probability Distributions of Random Variables

Definition 15 (Probability Mass Function). *The* ***probability mass function*** *(p.m.f) of a discrete random variable X is the function p (or p_X) defined by*

$$p(k) = P(X = k)$$

for possible values k of X.

The function p_X gives the probability of each possible value of X. Probabilities of other events of X then come by additivity: for any subset $B \subset \mathbb{R}$

$$P(X \in B) = \sum_{k \in B} P(X = k) = \sum_{k \in B} p_X(k)$$

Definition 16 (Probability Density Function). *Let X be a random variable. If a function f satisfies*

$$P(X \leq b) = \int_{-\infty}^{b} f(x)dx$$

for all real values b, then f is the ***probability density function****(p.d.f) of X.*

In fact, if f satisfies this definition, then

$$P(X \in B) = \int_B f(x)dx$$

for any subset B of the real line for which integration makes sense.

Theorem 15. *If a random variable X has density function f then point values have probability zero:*

$$P(X = c) = \int_c^c f(x)dx = 0 \qquad \text{for any real } c$$

It follows that a random variable with a density function is not discrete, and the probabilities of interval are not changed by including or excluding endpoints.

Remark. A random variable X can not have two different density functions.

3.4 Cumulative Distribution Function

Definition 17 (Cumulative Distribution Function). *The cumulative distribution function (c.d.f) of a random variable X is defined by*

$$F(s) = P(X \leq s) \qquad \textit{for all } s \in \mathbb{R}$$

Be mindful of the convention that the inequality is $\leq$ in the equation.

The cumulative distribution function gives a way to describe the probability distribution of any random variable, including those that do not fall into the discrete or continuous categories. The cumulative distribution function give probabilities of left-open right-closed intervals of the form (a,b]:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Note that:

- the domain of the CDF is the real line $(-\infty, +\infty)$ with the range $[0, 1]$.
- CDF is a non decreasing function, that is, $F(x) \leq F(y)$ for $x \leq y$.
- The CDF is right continuous. It does not jump at x when you approach x from above ($\lim_{x \to a^+} F(x) = F(a)$).

However, contrary to discrete random variables, the CDF of a continuous random variable is a continuous function (there are no jumps).

Knowing these probabilities is enough to determin the distribution of X completely.

Cumulative distribution function of a discrete random variable

$$F(s) = P(X \leq s) = \sum_{k:k \leq s} P(X = k)$$

where the sum extends over those possible values k of X that are less than or equal to s.

Cumulative distribution function of a continuous random variable

$$F(s) = P(X \leq s) = \int_{-\infty}^{s} f(x)dx$$

This equation comes from the definition of probability density function.

Theorem 16. *Let the random variable X have cumulative distribution function F.*

1. *Suppose F is piecewise constant. Then X is a discrete random variable. The possible values of X are the locations where F has jumps, and if x is such a point, then P(X = x) equals the magnitude of the jump of F at X.*

2. *Suppose F is continuous and the derivative $F'(x)$ exists everywhere on the real line, except possibly at finitely many points.*
   Then X is continuous random variable and f(x) = F′(x) is the density function of X. If F is not differentiable at x, then the value f(x) can be set arbitrarily.

3.5 Expectation

Expectation of a discrete random variable

Definition 18. *The expectation or mean of a discrete random variable X is defined by*

$$E(X) = \sum_k kP(X = k)$$

where the sum ranges over all the possible values k of X.

The expectation is also called the first moment, conventionally denoted as $\mu = E(X)$. The expectation is the weighted average of the possible outcomes, where the weights are given by probabilities.

Expectation of a continuous random variable

In continuous case averaging is naturally done via integrals. The weighting is given by the density function.

Definition 19. *The expectation or mean of a continuous random variable X with density function f is*

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

An alternative symbol is $\mu = E[X]$.

It is important to keep separate the random variable (X on the left) and the integration variable(x on the right).

Expectation of a function of a random variable

Taking a function of an existing random variable creates a new random variable.

Theorem 17. *Let g be a real-valued function defined on the range of a random variable X. If X is a discrete random variable then*

$$E[g(X)] = \sum_k g(k)P(X = k)$$

while if X is a continuous random variable with density function f then

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

Proof. The key is that the event g(X)=y is the disjoint union of the events X=k over those values k that satisfy g(k)=y:

$$\begin{aligned}
E[g(X)] &= \sum_y yP(g(X) = y) = \sum_y y \sum_{k:g(k)=y} P(X = k) \\
&= \sum_y \sum_{k:g(k)=y} yP(X = k) = \sum_y \sum_{k:g(k)=y} g(k)P(X = k) \\
&= \sum_k g(k)P(X = k)
\end{aligned}$$

Theorem 18. *The n* ***th moment*** *of the random variable X is the expectation $E(X^n)$. In the discrete case the nth moment is calculated by*

$$E(X^n) = \sum_k k^n P(X = k)$$

If X has density functino f its nth moment is given by

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

The second moment, $E(X^2)$, is also called the mean square.

4 Conditional Distributions and Expectation

4.1 Conditioning on an event

First new definition comes by applying $P(A|B) = \frac{P(AB)}{P(B)}$ to an event $A = \{X = k\}$ for a discrete random variable X.

Definition 20 (Conditional probability mass function of X, given B). *Let X be a discrete random variable and B an event with P(B)>0, Then the conditional probability mass function of X, given B is the function $p_{X|B}$ defined as follows for all possible values k of X:*

$$p_{X|B}(k) = P(X = k|B) = \frac{P(\{X = k\} \cap B)}{P(B)}.$$

Just like a regular probability mass function, its values are nonnegative and sum up to one.

The key point above was that the events $\{X = k\} \cap B$ are disjoint for different values of k and their union over k is B.
We can use the conditional probability mass function to compute an expectation.

Definition 21 (Conditional expectation of X, given the event B). *Let X be a discrete random variable and B an event with P(B)>0, Then the conditional expectation of X, given the event B is the function, is denoted by $E[X|B]$and defined as*

$$E[X|B] = \sum_k k p_{X|B}(k) = \sum_k kP(X = k|B)$$

where the sum ranges over all possible values k of X.

Applying the averaging principle $P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$ to an event $A = \{X = k\}$ gives the following identity:

Theorem 19. *Let Ω be a sample space, X a discrete random variable on Ω, and $B_1, \ldots, B_n$ a partition of Ω such that each $P(B_i) > 0$. Then the (unconditional) probabilities mass function of X can be calculated by averaging the conditional probabilities mass function:*

$$p_X(k) = \sum_{i=1}^n p_{X|B_i} P(B_i).$$

The averaging idea extends to expectations.

Theorem 20. *Let Ω be a sample space, X a discrete random variable on Ω, and $B_1, \ldots, B_n$ a partition of Ω such that each $P(B_i) > 0$. Then*

$$E[X] = \sum_{i=1}^{n} E[X|B_i]P(B_i).$$

4.2 *Conditioning on a random variable*

Let the partition in "Conditioning on an event" part come from another discrete randomvariable Y, then we followings.

Definition 22 (Conditional Probability Mass Function)**.** *Let X and Y be discrete random variables. The conditional probability mass function of Y given X = x is the folloing two-variable function:*

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

The conditional expectation of Y given X = x is

$$\mathbb{E}[Y|X = x] = \sum_y y \cdot P(Y = y \mid X = x) = \sum_y y \cdot p_{Y|X}(y|x).$$

The definition above are valid for y such that $P(X = x) > 0$.

As y varies, the events Y=y form a partition of Ω. Hence, we have

Theorem 21. *Let X and Y be discrete random variables. Then*

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y)$$

and

$$E(X) = \sum_y E[X|Y = y]p_Y(y).$$

The sums extend over those values y such that $p_Y(y) > 0$

The conditional probability mass function $p_{Y|X}(y|x)$ is just a probability mass function in x for each fixed value of y, whenever $p_X(x) > 0$.
The conditional expectation also satisfies familiar properties of usual expectation. For example:

$$\mathbb{E}[g(Y)|X = x] = \sum_y g(y) \cdot p_{Y|X}(y|x)$$

4.3 *Conditional distribution for jointly continuous random variables*

Definition 23 (Conditional Probability Density Function)**.** *Let X and Y be jointly continuous random variables with joint density function $f_{X,Y}(x,y)$. The conditional probability density function of Y given X = x is,*

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

Just as an ordinary density function, a conditional one can also be used to calculated conditional probabilities and expectations. The definition below gives the continuous counterpart of the discrete formula.

Definition 24. *The conditional probability that $X \in A$, given $Y = y$, is*

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

The conditional expectation of g(X), given Y = y, is

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

The quantities above are defined for y such that $f_Y(y) > 0$.

The averaging identities also works in the continuous case.

Theorem 22. *Let X and Y be jointly continuous. Then*

$$f_X(x) = \int_{\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)dy.$$

For any function g for which the expectations below make sense,

$$E[g(X)] = \int_{\infty}^{\infty} E[g(X)|Y = y]f_Y(y)dy.$$

Summary of conditional probability

- Total Probability
$$P(A) = \mathbb{E}[P(A|X)]$$
- Total Expectation
$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$
- Total Conditional Expectation
$$P(Y|A) = \mathbb{E}[P(Y|X, A)|A]$$
- Total Conditional Probability
$$\mathbb{E}[Y|A] = \mathbb{E}[\mathbb{E}[Y|X, A]|A]$$

4.4 *Conditional expectation*

In this section we discuss a conditional expectation that achieves some degree of unification of treatment of discrete and continuous random variables. A quick recap:

Definition 25 (Conditional Expectation). *Let X and Y be discrete or jointly continuous random variables. The conditional expectation of Y given X = x, denoted by $\mathbb{E}[Y|X = x](x)$, is a function of x,*

$$\mathbb{E}[Y|X = x](x) = \begin{cases} \sum_y y \cdot P(Y = y \mid X = x) & \Omega \text{ is discrete} \\ \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)dy & \Omega \text{ is continuous} \end{cases}$$

Before, we have the conditional expectation of X given Y = y, denoted by $E[X|Y = y]$. **For each legitimate y-value,** $E[X|Y = y]$ **is a real number.** We think of it as a function of y, denoted by $v(y) = E[X|Y]$. We can summarize the construction also by saying that the random variable $E(X|Y)$ takes the value $E[X|Y = y]$ when Y = y.

The key idea is that $E[X|Y = y]$ is a real number, and it's a possible value of the function $E(X|Y)$.

Definition 26 (Conditional expection as a random variable). *Let X and Y be discrete or jointly continuous random variables. The conditional expectation of X given Y, denoted by $E(X|Y)$, is by definition the random variable $v(Y)$ where the function v is defined by $v(y) = E(X|Y = y)$.*

Definition 27. *The **conditional expectation** of Y given A, for discrete case, is,*

$$E(Y|A) = \frac{1}{P(A)} \sum_y yP(\{Y = y\} \cap A) = \sum_y yP(Y = y|A)$$

Definition 28 (Law of Total Expectation). *If $A_1, \ldots, A_k$ partitions Ω and Y is a random variable, then the* ***law of total expectation*** *states that,*

$$\mathbb{E}[Y] = \sum_{i=1}^{k} \mathbb{E}[Y|A_i]P(A_i)$$

More generally, $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$

Proof. For the discrete case,

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y|X]] &= \sum_x \mathbb{E}[Y|X=x] \cdot P(X=x) \\
&= \sum_x \left(\sum_y y \cdot P(Y=y \mid X=x) \right) P(X=x) \\
&= \sum_y y \sum_x P(Y=y \mid X=x) \cdot P(X=x) \\
&= \sum_y y \sum_x P(Y=y, X=x) \\
&= \sum_y y \cdot P(Y=y) \\
&= \mathbb{E}(Y)
\end{aligned}$$

4.5 *Conditioning on multiple random variables*

A prelude to stochastic processes! Finally we're about to get there.

A stochastic process in discrete time is a sequence of random variables $X_0, X_1, X_2, \ldots$. One can think of this sequence as the time evolution of a random quantity. The random variable X_n is called the state of the process at time n.

$$\begin{aligned}
P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_0 = x_0)P(X_1 = x_1|X_0 = x_0) \\
\cdots P(X_n = x_n|X_0 = x_0, \ldots, X_{n-1} = x_{n-1})
\end{aligned}$$

A larger important class of stochastic processes have the property that, at any given time, the past influences the future only through the present state. Concretely speaking, all but the last state can be drop from the conditioning side of each conditional probability in the equation above.

Definition 29 (Markov Chain). *Let $X_0, X_1, X_2, \ldots$ be a stochastic process oif discrete random variables. This process is a Markov chain if*

$$P(X_{n+1} = x_{n+1}|X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

for all $n \geq 0$ and all $x_0, \ldots, x_n$ such that $P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n)$.

5 Time-Homogeneous Markov Chains

5.1 Finite State, Time-Homogeneous Chains

Definition 30 (Finite State Stochastic Process). *A **finite state stochastic process** $(X_n)_{n\geq 0}$ has time steps in $\mathbb{N}$ and values in $S = [N-1]$.*

Definition 31 (Markov Property). *The **Markov property** claims that for every $n \in \mathbb{N}$ and every sequence of states $(i_0, i_1, \dots)$ where $i_j \in S$, the behavior of a system depends only on the previous state,*

$$P(X_n = i_n | X_0 = i_0, \dots, X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1})$$

Definition 32 (Time Homogeneity). *A markov chain is **time-homogenous** if the probabilities in definition above do not depend on n,*

$$P(X_n = i_n | X_{n-1} = i_{n-1}) = P(X_1 = i_1 | X_0 = i_0) \quad (n \in \mathbb{N})$$

Definition 33 (Transition Matrix). *The **transition matrix P** for a time-homogeneous Markov chain is the $N \times N$ matrix whose (i,j)th entry P_{ij} is the one-step transition probability $p(i,j) = P(X_1 = j | X_0 = i)$*

Remark. The transition matrix P is stochastic, that is,

- (Non-Negative Entries) $0 \leq P_{ij} \leq 1$ for $1 \leq i, j \leq N$.
- (Row Sum Equal to 1) $\sum_{j=1}^{N} P_{ij} = 1$ for $1 \leq i \leq N$.

Example 1

Let $(X_n)_{n\geq 0}$ denote a sequence of coin flips where,

$$P(X_{n+1} = H \mid X_n) = \begin{cases} 0.51 & \text{if } X_n = H \\ 0.49 & \text{if } X_n = T \end{cases}$$

and,

$$P(X_{n+1} = T \mid X_n) = \begin{cases} 0.51 & \text{if } X_n = T \\ 0.49 & \text{if } X_n = H \end{cases}$$

Then,

$$\mathbf{P} = \begin{pmatrix} 0.51 & 0.49 \\ 0.49 & 0.51 \end{pmatrix} = \begin{pmatrix} P_{HH} & P_{HT} \\ P_{TH} & P_{TT} \end{pmatrix}$$

5.2 Transition Probabilities

Definition 34 (Probability Distribution Vector). *The **distribution** of a discrete random variable X is the vector $\vec{\phi}$ if,*

$$\phi_j = P(X = j) \quad \forall j \in \mathbb{N}$$

Definition 35 (Initial Distribution Vector). *The **initial probability distribution** of a Markov chain $(X_n)_{n\geq 0}$ is the distribution $\vec{\phi}$ of X_0.*

Definition 36 (Transition Probabilities). *The n-step transition probability* $p_n(i,j) = P(X_n = i|X_0 = j)$ *is the (i,j)th entry in the matrix* $\mathbf{P}^n$.