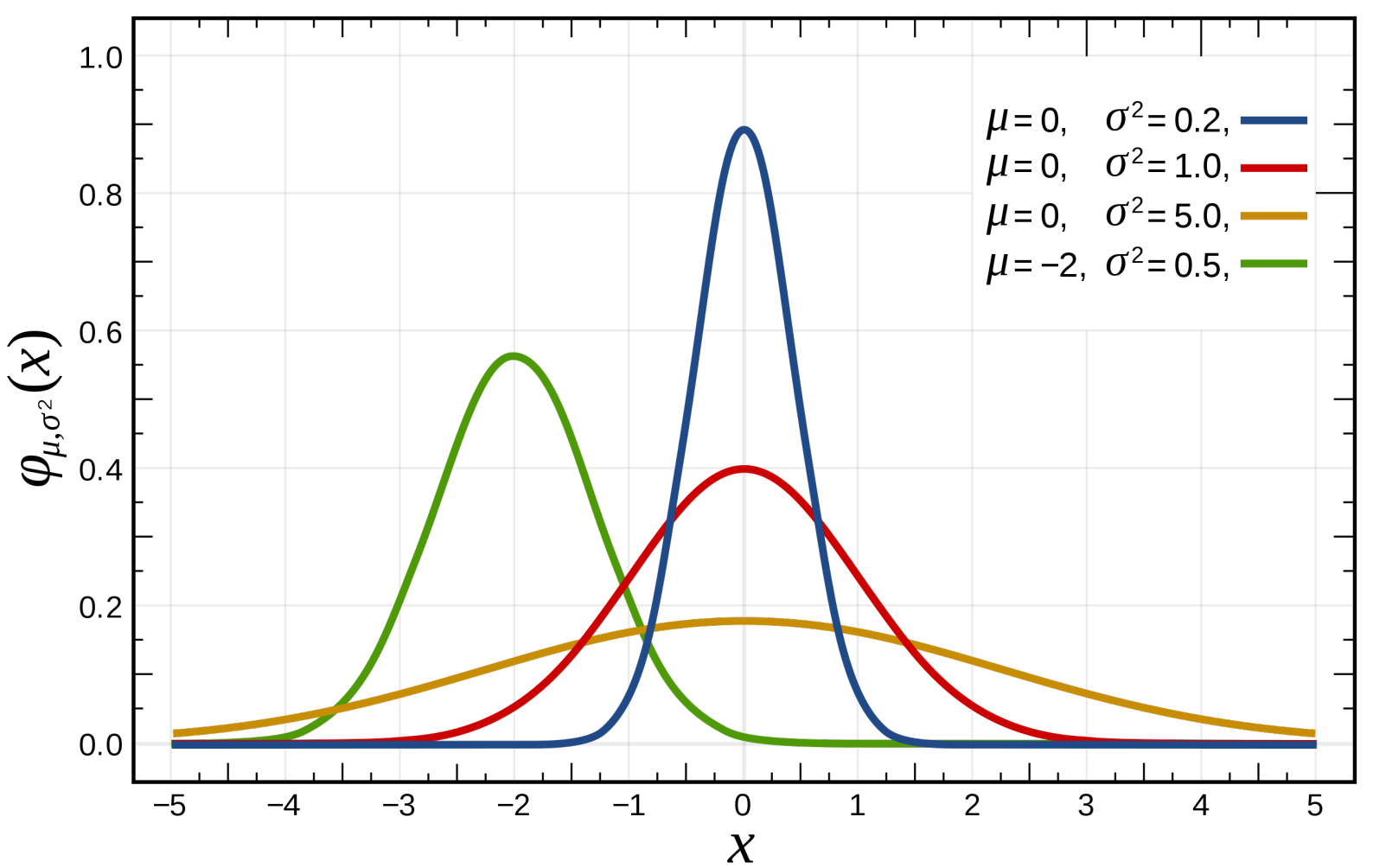

Introduction to Probability Theory

Quick reference book created from Math 531 Lecture Notes

Author: Chris Cai/ Linrong Cai

Institute: University of Wisconisn Madison

Date: May 15, 2023

Version: 1.0

Instructor: Benedek Valkó

The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account.

Contents

Chapter 1 Events and Probability

1.1 Kolmogorov's axioms and probability space

These are Kolmogorov's axioms for probability theory.

Definition 1.1

*A **probability space** is a triple $(\Omega, \mathcal{F}, P)$, with the following components.*

*(a) Ω is a set, called the **sample space**.*

(b) $\mathcal{F}$ is a collection of subsets of Ω. Members of $\mathcal{F}$ are called events. $\mathcal{F}$ is assumed to be a σ-algebra, which means that it satisfies the following three properties.

(b.1) $\Omega \in \mathcal{F}$. That is, the whole sample space itself is an event.

(b.2) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

(b.3) If $\{A_k\}_{1 \le k < \infty}$ is a sequence of members of $\mathcal{F}$, then their union $\bigcup_{k=1}^{\infty} A_k$ is also a member of $\mathcal{F}$.

*(c) P is a function from $\mathcal{F}$ into real numbers, called the **probability measure** . P satisfies the following axioms.*

(c.1) $0 \le P(A) \le 1$ for each event $A \in \mathcal{F}$.

(c.2) $P(\varnothing) = 0$ and $P(\Omega) = 1$.

(c.3) If $\{A_k\}_{1 \le k < \infty}$ is a sequence of pairwise disjoint events then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P\left(A_k\right).$$

Note *In mathematical analysis and in probability theory, a σ-algebra (also σ-field) on a set X is a nonempty collection Σ of subsets of X closed under complement, countable unions, and countable intersections. The ordered pair (X, Σ) is called a measurable space.*

1.2 Inclusion Exclusion Principle

Theorem 1.1 (inclusion-exclusion principle)

Let $A_1, A_2, A_3, \ldots$ be events in some probability space $(\Omega, \mathcal{F}, P)$. Then for each integer $n \ge 2$,

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

This is called the inclusion-exclusion identity.

1.3 Monotonicity and Countable Subadditivity

Proposition 1.1

Let $A, B, A_1, A_2, A_3, \ldots$ be events in some probability space $(\Omega, \mathcal{F}, P)$
(i) Monotonicity: if $A \subset B$ then $P(A) \leq P(B)$.
(ii) Countable subadditivity: for any sequence of events $\{A_k\}$,

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} P\left(A_k\right).$$

Countable subadditivity generalizes the countable additivity axiom in a natural way. Its truth should be fairly obvious because the union $\bigcup_{k=1}^{\infty} A_k$ can have overlaps whose probabilities are then counted several times over in the sum $\sum_{i=1}^{\infty} P\left(A_k\right)$. By taking $A_k = \varnothing$ for all $k > n$ we get a finite version of subadditivity:

$$P\left(A_1 \cup \cdots \cup A_n\right) \leq P\left(A_1\right) + \cdots + P\left(A_n\right)$$

valid for all events $A_1, \ldots, A_n$.

♣

Corollary 1.1

Let $\{A_k\}$ be a sequence of events on $(\Omega, \mathcal{F}, P)$.
(i) If $P\left(A_k\right) = 0$ for all k, then $P\left(\bigcup_k A_k\right) = 0$.
(ii) If $P\left(A_k\right) = 1$ for all k, then $P\left(\bigcap_k A_k\right) = 1$.

♡

1.4 Continuity of Probability

Definition 1.2

Suppose $\{A_k\}_{k \in \mathbb{Z}_{>0}}, \{B_k\}_{k \in \mathbb{Z}_{>0}}, A$, and B are events in a probability space $(\Omega, \mathcal{F}, P)$. We say that A_k increases up to A and use the notation

$$A_k \nearrow A$$

if the events A_k are nested nondecreasing, which means that $A_1 \subset A_2 \subset A_3 \subset \cdots \subset A_k \subset \cdots$, and $A = \bigcup_k A_k$. Figure 1 illustrates. Analogously, we say that B_k decreases down to B and use the notation

$$B_k \searrow B$$

if the events B_k are nested nonincreasing, which means that $B_1 \supset B_2 \supset B_3 \supset \cdots \supset B_k \supset \cdots$, and $B = \bigcap_k B_k$

♠

Theorem 1.2

If $A_k \nearrow A$ or $A_k \searrow A$, then the probabilities converge: $\lim_{k \to \infty} P\left(A_k\right) = P(A)$

♡

1.5 Conditional Probability

Definition 1.3 (conditional probability)

Let B be an event on the probability space $(\Omega, \mathcal{F}, P)$ such that $P(B) > 0$. Then for all events $A \in \mathcal{F}$ the conditional probability of A given B is defined as

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

♠

Proposition 1.2

Let B be an event on the probability space $(\Omega, \mathcal{F}, P)$ such that $P(B) > 0$. Then as a function of the event A, $P(A \mid B)$ is a probability measure on $(\Omega, \mathcal{F})$

♣

Theorem 1.3

In each statement below all events are on the same probability space$(\Omega, \mathcal{F}, P)$

(a) Let A and B be two events and assume $P(B) > 0$. Then

$$P(AB) = P(B)P(A \mid B)$$

Let $A_1, \ldots, A_n$ be events and assume $P(A_1 \cdots A_{n-1}) > 0$. Then

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

(b) Let $\{B_i\}$ be a countable partition of Ω and A any event. Then

$$P(A) = \sum_{i:P(B_i)>0} P(A \mid B_i)\, P(B_i)\,.$$

The sum above ranges over those indices i such that $P(B_i) > 0$.

♡

1.6 Bayes' Formula

Theorem 1.4 (Bayes' formula)

Let $\{B_k\}$ be a countable partition of the sample space Ω. Then for any event A with $P(A) > 0$ and each k such that $P(B_k) > 0$,

$$P(B_k \mid A) = \frac{P(AB_k)}{P(A)} = \frac{P(A \mid B_k)\, P(B_k)}{\sum_{i:P(B_i)>0} P(A \mid B_i)\, P(B_i)}$$

♡

1.7 Independent Events

Definition 1.4

Two events A and B are independent if

$$P(AB) = P(A)P(B)$$

♠

Theorem 1.5

Suppose that A and B are independent events. Then the same is true for each of these pairs: A^c and B, A and B^c, and A^c and B^c.

♡

The definition of independence of more than two events requires that the product property hold for any subcollection of events.

Definition 1.5

(a) Events $A_1, \ldots, A_n$ are independent (or mutually independent) if for every collection $A_{i_1}, \ldots, A_{i_k}$, where $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$

$$P\left(A_{i_1} A_{i_2} \cdots A_{i_k}\right) = P\left(A_{i_1}\right) P\left(A_{i_2}\right) \cdots P\left(A_{i_k}\right)$$

(b) Let $\{A_k\}_{k \in \mathbb{Z}_{>0}}$ be an infinite sequence of events in a probability space $(\Omega, \mathcal{F}, P)$. Then events $\{A_k\}_{k \in \mathbb{Z}_{>0}}$ are independent if for each $n \in \mathbb{Z}_{>0}$, events $A_1, \ldots, A_n$ are independent.

♠

Theorem 1.6

(a) Suppose events $A_1, \ldots, A_n$ are mutually independent. Then for every collection $A_{i_1}, \ldots, A_{i_k}$, where $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, we have

$$P\left(A_{i_1}^* A_{i_2}^* \cdots A_{i_k}^*\right) = P\left(A_{i_1}^*\right) P\left(A_{i_2}^*\right) \cdots P\left(A_{i_k}^*\right)$$

where each A_i^* can represent either A_i or A_i^c.

(b) Let $\{A_k\}_{k \geq 1}$ be a finite or infinite sequence of independent events. Let $0 = k_0 < k_1 < \cdots < k_n$ be integers. Let $B_1, \ldots, B_n$ be events constructed from the A_ks so that, for each $j = 1, \ldots, n$, B_j is formed by applying set operations to $A_{k_{j-1}+1}, \ldots, A_{k_j}$. Then the events $B_1, \ldots, B_n$ are independent.

♡

Definition 1.6

Let $A_1, \ldots, A_n$ and B be events on $(\Omega, \mathcal{F}, P)$ and assume $P(B) > 0$. Then events $A_1, \ldots, A_n$ are conditionally independent, given B, if for every collection $A_{i_1}, \ldots, A_{i_k}$, where $2 \leq k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$,

$$P\left(A_{i_1} A_{i_2} \cdots A_{i_k} \mid B\right) = \prod_{j=1}^{k} P\left(A_{i_j} \mid B\right)$$

♠

Chapter 2 Random Variables and Probability Distributions

2.1 Random Variables

Often we are interested in some numerical value associated to the outcome of a random experiment. This just means that we are interested in the value of a function that maps the elements of the sample space into the real numbers. These functions are called random variables.

Definition 2.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable on Ω is a real valued function $X : \Omega \to \mathbb{R}$, for which $\{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F}$ for any $c \in \mathbb{R}$.

There is a simple way to encode an events as a random variable.

Definition 2.2

Let B be an event in a probability space. Then the indicator function (or indicator random variable) of B is defined as the function

$$I_B(\omega) = \begin{cases} 1, & \text{if } \omega \in B \\ 0, & \text{if } \omega \notin B \end{cases}$$

Note *Note that I_B is a random variable.*

2.2 Probability Distributions

Through the probabilities of events of type $\{X \in B\}$, a random variable induces a probability measure on the real line.

Definition 2.3

Let X be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The probability distribution of X is the probability measure μ on $\mathbb{R}$ defined by

$$\mu(B) = P(X \in B)$$

for Borel subsets B of $\mathbb{R}$.

Note *In mathematics, a Borel set is any set in a topological space that can be formed from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and relative complement. Borel sets are named after Émile Borel.*

2.3 Discrete Random Variables

Definition 2.4

A random variable X is discrete if there exists a finite or countably infinite set $B \subset \mathbb{R}$ such that $P(X \in B) = 1$.

We say that those values k for which $P(X = k) > 0$ are the possible values of the discrete random variable X.
As for functions in general, the range of a random variable X is the set of all its values: the range of X is the set $\{X(\omega) : \omega \in \Omega\}$. In particular, if the range of the random variable X is finite or countably infinite, then X is a discrete random variable.

2.4 Probability Mass Functions

Definition 2.5

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 the possible values k of X. (In some cases it is convenient to extend the function p for other values as well, we can use the same definition even if $P(X = k) = 0$.)
Probabilities of events involving X come by summing values of the probability mass function: for any subset $B \subseteq \mathbb{R}$

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

where the sum is over the possible values k of X that lie in B.

2.5 Bernoulli Distribution

Definition 2.6

Let $0 \leq p \leq 1$. We say that a random variable X has Bernoulli distribution with parameter p if X is a discrete random variable with probability mass function

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

We abbreviate this as $X \sim \text{Ber}(p)$.

Note *The distribution of an indicator random variable I_B is always Bernoulli, its parameter is $P(B)$.*

2.6 Cumulative distribution functions

Definition 2.7

Let X be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The cumulative distribution function (c.d.f.) of X is defined by $F(s) = P(X \leq s)$ for all $s \in \mathbb{R}$.

Proposition 2.1

(a) Let $F : \mathbb{R} \to [0, 1]$ be the cumulative distribution function of a random variable X. Then F has the following properties.

(i) Monotonicity: if $s < t$ then $F(s) \leq F(t)$.
(ii) Right continuity: $F(t) = F(t+)$ for each $t \in \mathbb{R}$.
(iii) $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to\infty} F(t) = 1$.
(b) Conversely, if a function $F : \mathbb{R} \to [0,1]$ has properties (i)-(iii) above, then F is the cumulative distribution function of some random variable.
(c) Let X be a random variable with cumulative distribution function F. Then for any $s \in \mathbb{R}$ we have these identities:

$$P(X < s) = F(s-)$$

and

$$P(X = s) = F(s) - F(s-).$$

2.7 Probability Density Function

Definition 2.8

Let X be a random variable on $(\Omega, \mathcal{F}, P)$. If a function f on $\mathbb{R}$ satisfies $f(x) \geq 0$ for all x and

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

for all real values b, then f is the probability density function (p.d.f.) of X. When X has a density function, we call X an absolutely continuous random variable.

Theorem 2.1

Suppose the cumulative distribution function F of the random variable X is continuous and the derivative $F'(x)$ exists everywhere on the real line, except possibly at countably many points. Then X is an absolutely continuous random variable and $f(x) = F'(x)$ is the density function of X. If F is not differentiable at a point x, then the value $f(x)$ can be set arbitrarily.

Definition 2.9

Let f be a piecewise continuous function on $\mathbb{R}$. Then f is the density function of a random variable if and only if $f(x) \geq 0$ for all $x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} f(x)dx = 1$

Corollary 2.1

Random variables can be neither discrete nor absolutely continuous.

Example 2.1 Fix $a < b$ and define $F : \mathbb{R} \to [0,1]$ by

$$F(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{3} \cdot \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b. \end{cases}$$

Note *It is natural to generalize the above to random vectors.*

2.8 Uniform Distribution

Definition 2.10 (Uniform distribution on an interval.)

Let $[a, b]$ be a bounded interval on the real line. Random variable X has the uniform distribution on the interval $[a, b]$ if X has density function

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

Abbreviate this by $X \sim \text{Unif}[a, b]$ If $X \sim \text{Unif}[a, b]$ and $[c, d] \subset [a, b]$, then

$$P(c \leq X \leq d) = \int_c^d \frac{1}{b-a} dx = \frac{d-c}{b-a}$$

♠

Definition 2.11

Let Ω be a subset of d-dimensional Euclidean space $\mathbb{R}^d$ with finite volume. Then the random point $\mathbf{X}$ is uniformly distributed on Ω if its joint density function is

$$f(\mathbf{x}) = \begin{cases} \frac{1}{\text{vol}(\Omega)} & \text{if } \mathbf{x} \in \Omega \\ 0 & \text{if } \mathbf{x} \notin \Omega \end{cases}$$

♠

2.9 Notion of equality

Definition 2.12 (Almost sure equality)

Let X and Y be random variables defined on $(\Omega, \mathcal{F}, P)$. Then X and Y are equal almost surely if $P(X = Y) = 1$. This is abbreviated by $X = Y$ a.s.

♠

Note *Almost sure equality is also expressed by saying $X = Y$ with probability one, abbreviated $X = Y$ w.p.1. Below is a discrete and an absolutely continuous example of almost sure equality $X = Y$ where pointwise equality fails.*

Example 2.2 Let $\Omega = \{1, 2, 3\}$ with probability measure $P\{1\} = P\{2\} = \frac{1}{2}$ and $P\{3\} = 0$. Define random variables X and Y on Ω by

$$X(1) = Y(1) = 1, X(2) = Y(2) = 2, X(3) = 3 \text{ and } Y(3) = 0.$$

Then $P(X = Y) = P\{1, 2\} = 1$.

Definition 2.13 (Equality in distribution)

Random variables X and Y are equal in distribution if $P(X \in B) = P(Y \in B)$ for all Borel subsets B of $\mathbb{R}$. This is abbreviated by $X \stackrel{d}{=} Y$.

♠

Theorem 2.2

Suppose X and Y are random variables on the same probability space $(\Omega, \mathcal{F}, P)$. Then $P(X = Y) = 1$ implies $X \stackrel{d}{=} Y$.

♡

Chapter 3 Independent and dependent random variables

3.1 Independence

Definition 3.1

(a) Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be random vectors defined on the same probability space. Let $\mathbf{X}_i$ be $\mathbb{R}^{d_i}$-valued. Then $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent if

$$P\left(\mathbf{X}_1 \in B_1, \mathbf{X}_2 \in B_2, \ldots, \mathbf{X}_n \in B_n\right) = \prod_{k=1}^{n} P\left(\mathbf{X}_k \in B_k\right)$$

for all Borel subsets $B_i \subset \mathbb{R}^{d_i}, i = 1, \ldots, n$.

(b) Let $\{\mathbf{X}_k\}_{k \in \mathbb{Z}_{>0}}$ be an infinite sequence of random vectors defined on some probability space $(\Omega, \mathcal{F}, P)$. Then the random vectors $\{\mathbf{X}_k\}_{k \in \mathbb{Z}_{>0}}$ are independent if, for each $n \in \mathbb{Z}_{>0}$, the random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent.

♠

Theorem 3.1

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be random vectors defined on the same probability space. For $1 \leq i \leq n$ let d_i be the dimension of $\mathbf{X}_i$ and set $d = d_1 + \cdots + d_n$. Define the d-dimensional random vector $\mathbf{Y} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ by putting all the coordinates of the $\mathbf{X}_i$ s together. Denote the joint cumulative distribution functions of these random vectors by $F_{\mathbf{Y}}, F_{\mathbf{X}_1}, \ldots, F_{\mathbf{X}_n}$. Then $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent if and only if

$$F_{\mathbf{Y}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_n\right) = F_{\mathbf{X}_1}\left(\mathbf{x}_1\right) \cdot F_{\mathbf{X}_2}\left(\mathbf{x}_2\right) \cdots F_{\mathbf{x}_n}\left(\mathbf{x}_n\right)$$

for all vectors $\mathbf{x}_1 \in \mathbb{R}^{d_1}, \mathbf{x}_2 \in \mathbb{R}^{d_2}, \ldots, \mathbf{x}_n \in \mathbb{R}^{d_n}$.

♡

3.2 Independent and identically distributed random variables

Definition 3.2

Random variables $X_1, X_2, X_3, \ldots$ are independent and identically distributed (abbreviated i.i.d.) if they are independent and each X_k has the same probability distribution. That is, $X_k \stackrel{d}{=} X_\ell$ for any two indices k, ℓ.

♠

3.3 Independence of jointly absolutely continuous random variables.

Theorem 3.2

Theorem 3.11. Let $X_1, \ldots, X_d$ be random variables on the same sample space. Assume that for each $j = 1, 2, \ldots, d$, X_j has density function f_{X_j}. (a) If $X_1, \ldots, X_d$ have joint density function f given by

$$f\left(x_1, x_2, \ldots, x_d\right) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right) \cdots f_{X_d}\left(x_d\right)$$

then $X_1, \ldots, X_d$ are independent.

(b) Suppose $X_1, \ldots, X_d$ are independent. Then they are jointly absolutely continuous with joint density function

$$f(x_1, x_2, \ldots, x_d) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_d}(x_d)$$

♡

3.4 Finding Independence

Theorem 3.3

(a) Suppose $X_1, \ldots, X_n$ are independent random variables and for each index i, g_i is a function on the range of X_i. Then the random variables $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are independent.
One needs to assume that $g_1, \ldots, g_n$ are measurable, but this does not come up usually in applications.

(b) Let $\{X_k\}_{k \geq 1}$ be a finite or infinite sequence of independent random variables. Let $0 = k_0 < k_1 < \cdots < k_n$ be integers. Let $g_1, \ldots, g_n$ be functions such that g_j is defined on the range of the random vector $(X_{k_{j-1}+1}, \ldots, X_{k_j})$. Define new random variables $Y_j = g_j(X_{k_{j-1}+1}, \ldots, X_{k_j})$ for $j = 1, \ldots, n$. Then the random variables $Y_1, \ldots, Y_n$ are independent.

♡

3.5 Binomial Distribution

Definition 3.3 (Binomial Distribution)

Let n be a positive integer and $0 \leq p \leq 1$. A random variable X has the binomial distribution with parameters n and p if the possible values of X are $\{0, 1, \ldots, n\}$ and the probabilities are

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

Abbreviate this by $X \sim \text{Bin}(n, p)$.

♠

3.6 Geometric Distribution

Definition 3.4 (Geometric Distribution)

*Let $0 < p \leq 1$. A random variable X has the **geometric distribution** with success parameter p if the set of possible values of X is $\mathbb{Z}_>0$ and X satisfies $P(X = k) = (1-p)^{k-1}p$ for positive integers k. Abbreviate this by $X \sim \text{Geom}(p)$.*

♠

3.7 Negative Binomial Distribution

Definition 3.5 (Negative binomial distribution)

Let k be a positive integer and $0 < p \leq 1$. A random variable X has the negative binomial distribution with parameters (k, p) if the set of possible values of X is the set of integers $\mathbb{Z}_{\geq k} = \{k, k+1, k+2, \ldots\}$ and the probability mass function is

$$P(X = n) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \quad \text{for } n \in \mathbb{Z}_{\geq k}$$

Abbreviate this by $X \sim \text{Negbin}(k, p)$.

Note *The Negbin $(1, p)$ distribution is the same as the $\text{Geom}(p)$ distribution.*

Corollary 3.1

Consider a sequence of independent trials with success probability $0 < p \leq 1$. Let N_k be the number of trials needed for the k th success. Set $Y_1 = N_1$ and $Y_k = N_k - N_{k-1}$ for $k \geq 2$. Then the random variables $Y_1, Y_2, Y_3, \ldots$ are i.i.d. In particular, $N_k - N_{k-1} \sim \text{Geom}(p)$ for each $k \geq 2$.

Note *Can think of negative binomial as sum of geometric random variables.*

3.8 Possion Distribution

Definition 3.6 (Possion Distribution)

Let $\lambda > 0$. A random variable X has the Poisson distribution with parameter λ if the possible values of X are the nonnegative integers and the probability mass function is

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for } k \in \{0, 1, 2, \ldots\}$$

Abbreviate this by $X \sim$ Poisson (λ).

Theorem 3.4

Fix $\lambda > 0$. For positive integers n for which $\lambda/n < 1$, let $S_n \sim \text{Bin}(n, \lambda/n)$. Then

$$\lim_{n \to \infty} P(S_n = k) = e^{-\lambda} \frac{\lambda^k}{k!} \quad \text{for all} \quad k \in \mathbb{Z}_{\geq 0}$$

3.9 Exponential Distributions

Definition 3.7

Let $0 < \lambda < \infty$. A random variable X has the exponential distribution with parameter λ if X has density function

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

on the real line. Abbreviate this by $X \sim \text{Exp}(\lambda)$.

Exponential Distribution

Theorem 3.5

Suppose that $X \sim \text{Exp}(\lambda)$. *Then for any* $s, t > 0$, (3.27)

$$P(X > t + s \mid X > t) = P(X > s).$$

♡

Note *This is also called the memoryless property.*

Theorem 3.6

Fix $\lambda > 0$. *Consider* n *large enough so that* $\lambda/n < 1$. *Suppose that for each* n, *the random variable* T_n *satisfies* $nT_n \sim \text{Geom}(\lambda/n)$. *Then*

$$\lim_{n\to\infty} P\left(T_n > t\right) = e^{-\lambda t} \quad \textit{for all real } t \geq 0$$

♡

3.10 Multinomial distribution

Definition 3.8

Let n *and* r *be positive integers and let* $p_1, p_2, \ldots, p_r$ *be positive reals such that* $p_1 + p_2 + \cdots + p_r = 1$. *The random vector* $\mathbf{X} = (X_1, \ldots, X_r)$ *has the multinomial distribution with parameters* n, r *and* $p_1, \ldots, p_r$ *if the possible values of* $\mathbf{X}$ *are integer vectors* $(k_1, \ldots, k_r)$ *such that* $k_j \geq 0$ *and* $k_1 + \cdots + k_r = n$, *and the joint probability mass function is given by*

$$P\left(X_1 = k_1, X_2 = k_2, \ldots, X_r = k_r\right) = \binom{n}{k_1, k_2, \ldots, k_r} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$$

where the multinomial coefficient is defined by

$$\binom{n}{k_1, k_2, \ldots, k_r} = \frac{n!}{k_1! k_2! \cdots k_r!}$$

Abbreviate this by $(X_1, \ldots, X_r) \sim \text{Mult}\left(n, r, p_1, \ldots, p_r\right)$.

♠

3.11 Gamma Distribution

Definition 3.9

Let $r, \lambda > 0$. *A random variable* X *has the gamma distribution with parameters* (r, λ) *if* X *is nonnegative and has probability density function*

$$f_X(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} \quad \textit{for } x > 0,$$

and $f_X(x) = 0$ *for* $x \leq 0$. *We abbreviate this* $X \sim \text{Gamma}(r, \lambda)$.

♠

3.12 Convolution

Definition 3.10

*As a mathematical concept, convolution is a way of multiplying functions to produce new functions. The operation is denoted by *. The convolution of two functions on the real line is the function $f * g$ whose value at x is defined by*

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

provided of course that these integrals are well-defined for all $x \in \mathbb{R}$. Change of variable from y to $x - y$ shows that convolution is symmetric: $f * g = g * f$.

♠

Theorem 3.7 (Discrete)

If X and Y are independent $\mathbb{Z}$-valued random variables with probability mass functions p_X and p_Y, then the probability mass function of $X + Y$ is

$$p_{X+Y}(n) = p_X * p_Y(n) = \sum_{k\in\mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{\ell\in\mathbb{Z}} p_X(n-\ell)p_Y(\ell)$$

♡

Theorem 3.8 (Continuous)

If X and Y are independent absolutely continuous random variables with density functions f_X and f_Y then the density function of $X + Y$ is (3.39) $f_{X+Y}(z) = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{\infty} f_X(z-x)f_Y(x)dx$.

♡

3.13 Exchangebale Random Variables

Definition 3.11

Random variables $X_1, \ldots, X_n$ are exchangeable if for any permutation σ on $\{1, 2, \ldots, n\}$, the joint distribution of $(X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)})$ is the same as the joint distribution of $(X_1, X_2, \ldots, X_n)$. In other words, permuting the random variables does not change the joint distribution.

♠

Note *Verifying exchangeability boils down to checking that either the joint cumulative distribution function, the joint probability mass function, or the joint density function is a symmetric function. (A function is symmetric if its value is not altered by permuting its arguments.) These cases are collected in the next theorem.*

Theorem 3.9

(i) The random variables $X_1, \ldots, X_n$ are exchangeable if for any permutation σ on $\{1, \ldots, n\}$ and for any choice of real numbers $x_1, x_2, \ldots, x_n$ we have

$$P(X_1 \le x_1, \ldots, X_n \le x_n) = P\left(X_1 \le x_{\sigma(1)}, \ldots, X_n \le x_{\sigma(n)}\right)$$

(ii) Suppose $X_1, \ldots, X_n$ are discrete. Then exchangeability is equivalent to

$$P(X_1 = x_1, \ldots, X_n = x_n) = P\left(X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)}\right)$$

for all permutations σ on $\{1, \ldots, n\}$ and for all choices of real numbers $x_1, x_2, \ldots, x_n$.

(iii) Suppose $X_1, \ldots, X_n$ are jointly absolutely continuous with joint density function $f_{X_1,\ldots,X_n}$. Then

exchangeability is equivalent to having the identity

$$f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = f_{X_1,\ldots,X_n}\left(x_{\sigma(1)},\ldots,x_{\sigma(n)}\right)$$

for all permutations σ on $\{1,\ldots,n\}$ and for all choices of real numbers $x_1, x_2, \ldots, x_n$, except possibly on a set of zero volume.

♡

Corollary 3.2 (Producing new exchangeable random variables)

Suppose that $X_1,\ldots,X_n$ are exchangeable.
(i) If $1 \leq m \leq n$ then $X_1,\ldots,X_m$ are also exchangeable.
(ii) Let $g : \mathbb{R} \to \mathbb{R}$ be a function and define $Y_k = g(X_k)$. Then the random variables $Y_1,\ldots,Y_n$ are also exchangeable.

♡

Corollary 3.3 (Applications of exchangeability)

If $X_1,\ldots X_n$ are exchangeable and $k_1,\ldots,k_m$ are distinct numbers from $\{1,2,\ldots,n\}$ then the following statements hold. (i) $X_{k_1},\ldots,X_{k_m}$ have the same joint distribution as $X_1,\ldots,X_m$. (ii) For all Borel sets $B \subset \mathbb{R}^m$,

$$P\left((X_{k_1},\ldots,X_{k_m}) \in B\right) = P\left((X_1,\ldots,X_m) \in B\right)$$

♡

Chapter 4 Simple Random Walk

[Upon proving that the best betting stragegy for Gambler's Ruin was to bet all on the first trial.] It is true that a man who does this is a fool. I have only proved that a man who does anything else is an even bigger fool. ——- Coolidge, Julian Lowell

4.1 Ganbler's Ruin

Example 4.1 Gambler's Ruin You play repeatedly the following gamble. A fair coin is tossed: heads you win a dollar, tails you lose a dollar. You start playing with x dollars in your pocket. You choose a target $M > x$. Then you play until you either reach M dollars or lose all your money. A question of obvious interest: what is the probability that you reach M dollars before going broke?

$$\begin{aligned} P_x(\text{ reach } M \text{ before } 0) &= \frac{1}{2}P_x(\text{ reach } M \text{ before } 0 \mid \text{ first flip heads }) \\ &\quad + \frac{1}{2}P_x(\text{ reach } M \text{ before } 0 \mid \text{ first flip tails }) \\ &= \frac{1}{2}P_{x+1}(\text{ reach } M \text{ before } 0) + \frac{1}{2}P_{x-1}(\text{ reach } M \text{ before } 0). \end{aligned}$$

$$1 = p_M - p_0 = \sum_{x=1}^{M} (p_x - p_{x-1}) = M\,(p_k - p_{k-1})$$

from which $p_k - p_{k-1} = 1/M$. Then for any x,

$$p_x = p_x - p_0 = \sum_{k=1}^{x} (p_k - p_{k-1}) = \frac{x}{M}$$

Definition 4.1

Fix $p \in (0,1)$. Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with $P(X_i = 1) = p$ and $P(X_i = -1) = 1 - p$. Let S_0 be an integer. (If S_0 is also random, then S_0 is independent of the random variables $\{X_i\}$.) For $n \geq 1$ define

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n$$

The random sequence $S_0, S_1, S_2, \ldots$ is the simple random walk (SRW) with initial position S_0. If an initial position is not specified, then $S_0 = 0$.
If $p = \frac{1}{2}$ then $\{S_n\}$ is called symmetric simple random walk (SSRW), while if $p \neq \frac{1}{2}$, then $\{S_n\}$ is asymmetric simple random walk.

♠

Theorem 4.1

Fix integers $0 < x < M$. Consider $SSRW$ $\{S_n\}_{n \geq 0}$ with nonrandom initial state $S_0 = x$. Then

$$P(S_n \text{ visits point } M \text{ before visiting } 0) = \frac{x}{N}.$$

♡

Theorem 4.2

For times $0 \leq m < n$ and points $a, b \in \mathbb{Z}$, SRW with initial point $S_0 = 0$ satisfies

$$P(S_{n+m} = a + b \mid S_m = a) = P(S_n = b)$$

♡

4.2 Symmetric simple random walk

Theorem 4.3 (Reflection principle)

Let a, b be integers with $b > \max(0, a)$. The number of paths of length n that go from 0 to a and visit point b along the way is equal to $N_{n,0\to 2b-a}$. In particular, for $SSRW$,

$$P\left(S_n = a, S_k = b \text{ for some } k = 0, \ldots, n\right) = P\left(S_n = 2b - a\right).$$

♡

4.3 Distribution of the running maximum

The running maximum of the random walk is defined by $M_n = \max\left(0, S_1, \ldots, S_n\right)$ for $n = 0, 1, 2, \ldots$ It is always nonnegative. We find its distribution.

Theorem 4.4

For $r \in \mathbb{Z}_{\geq 0}$, the running maximum of symmetric SRW satisfies

$$P\left(M_n = r\right) = P\left(S_n = r\right) + P\left(S_n = r + 1\right).$$

Note that one of the terms on the right is always zero depending on the parity of $n - r$.

♡

Note *Will include more on this later in stochastic process.*

Chapter 5 Expectation

5.1 Definition of expectation

Definition 5.1 (Arbitrary real-valued random variables)

All definitions of the expected value may be expressed in the language of measure theory. In general, if X is a real-valued random variable defined on a probability space $(\Omega, \Sigma, \mathrm{P})$, then the expected value of X, denoted by $\mathrm{E}[X]$, is defined as the Lebesgue integral

$$\mathrm{E}[X] = \int_{\Omega} X d\mathrm{P}$$

Note *This is a generalized definition which required Lebesgue Integral. I will direct the readers who are interested in this definiton to here . Another definiton is given below.*

Definition 5.2 (3 step construction)

(Definition of the expectation $\mathrm{E}[X]$). Consider random variables on some probability space $(\Omega, \mathcal{F}, P)$.

Step 1. A simple random variable is a discrete random variable with finitely many values. Let X be a nonnegative simple random variable. Then X is of the form

$$X(\omega) = \sum_{i=1}^{m} \alpha_i I_{A_i}(\omega)$$

where $\alpha_1, \ldots, \alpha_m$ are its distinct nonnegative real values and the events $A_i = \{X = \alpha_i\}$ form a partition of Ω. Define the expectation of X by

$$\mathrm{E}[X] = \sum_{i=1}^{m} \alpha_i P(A_i) = \sum_{i=1}^{m} \alpha_i P(X = \alpha_i)$$

Step 2. Let X be a $[0, \infty]$-valued random variable on Ω. That is, the values $X(\omega)$ are nonnegative reals and possibly ∞. Then the expectation of X is defined by $\mathrm{E}[X] = \sup\{EY : Y$ is a simple random variable on Ω such that

$$0 \leq Y(\omega) \leq X(\omega) \text{ for all } \omega \in \Omega\}$$

This defines a value $\mathrm{E}[X] \in [0, \infty]$.

Step 3. Let X be a $[-\infty, \infty]$-valued random variable on Ω. That is, the values $X(\omega)$ are real numbers, ∞ or $-\infty$. The random variable X is decomposed into its positive part X^+ and negative part X^- by defining $X^+(\omega) = X(\omega) \vee 0$ and $X^-(\omega) = (-X(\omega)) \vee 0$. Random variables $X, |X|, X^+$ and X^- satisfy the identities

$$X(\omega) = X^+(\omega) - X^-(\omega) \text{ and } |X(\omega)| = X^+(\omega) + X^-(\omega)$$

The expectations $E(X^+)$ and $E(X^-)$ were defined in Step 2. If at least one of them is finite, the expectation of X is defined by

$$\mathrm{E}[X] = E(X^+) - E(X^-)$$

In this case we say that $\mathrm{E}[X]$ *is well-defined. If both* $E\left(X^{+}\right)$*and* $E\left(X^{-}\right)$*are infinite, then* $\mathrm{E}[X]$ *is not defined.*

Note $\vee$ *is an alternative notation for maximum*

Theorem 5.1

(a) Suppose X *is a discrete random variable. Then the following expectations have well-defined values in* $[0, \infty]$ *:*

$$E\left(X^{+}\right)=\sum_{k\geq 0} kP(X=k) \quad \text{and} \quad E\left(X^{-}\right)=\sum_{k\leq 0}(-k)P(X=k).$$

If at least one of these values is finite, then EX *is well defined as the difference:*

$$EX=\sum_{k\geq 0} kP(X=k)-\sum_{k\leq 0}(-k)P(X=k)$$

(b) Suppose X *is an absolutely continuous random variable with density function* f*. Then the following expectations have well-defined values in* $[0, \infty]$ *:*

$$E\left(X^{+}\right)=\int_{0}^{\infty} xf(x)dx \quad \text{and} \quad E\left(X^{-}\right)=\int_{-\infty}^{0}(-x)f(x)dx$$

If at least one of these values is finite, then EX *is well defined as the difference:*

$$EX=\int_{0}^{\infty} xf(x)dx-\int_{-\infty}^{0}(-x)f(x)dx$$

Corollary 5.1

When well-defined the below definition also holds.

(a) For a discrete random variable

$$E(X)=\sum_{k} kP(X=k)$$

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

(b) For an absolutely continuous random variable X *with density function* f

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

5.2 Theorems and propositions for expectation

Theorem 5.2

Suppose $P(X\geq 0)=1$*. Then*

$$E(X)=\int_{0}^{\infty} P(X>s)ds$$

In the particular discrete case where $P\left(X\in\mathbb{Z}_{\geq 0}\right)=1$*, the formula can also be expressed as*

$$E(X)=\sum_{k=0}^{\infty} P(X>k)$$

Theorem 5.3

Let $\mathbf{X} = (X_1, \ldots, X_d)$ *be a random vector (or a scalar random variable, in case* $d = 1$ *) and* g *a real-valued function defined on the range of* $\mathbf{X}$*. Assume that either* $g \geq 0$ *or that* $g(\mathbf{X})$ *is absolutely integrable, so that the expectation* $E[g(\mathbf{X})]$ *is well-defined. Then we have the following formulas for this expectation. (a) Suppose* $X_1, \ldots, X_d$ *are discrete random variables. Then*

$$E[g(\mathbf{X})] = \sum_{\mathbf{k}} g(\mathbf{k}) P(\mathbf{X} = \mathbf{k})$$

where $\mathbf{k}$ *ranges over the possible values of* $\mathbf{X}$*. (b) Suppose* $X_1, \ldots, X_d$ *are jointly continuous random variables with joint density function* f*. Then*

$$E[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

♡

Proposition 5.1

Let X *and* Y *be random variables on* $(\Omega, \mathcal{F}, P)$*. Assume their expectations are well-defined. Then their expectations have the following properties.*

(i) Linearity: if EX *and* EY *are finite, then for any real numbers a,b,c,*

$$E[aX + bY + c] = aE(X) + bE(Y) + c.$$

(ii) Monotonicity: if $P(X \leq Y) = 1$*, then* $EX \leq EY$*. If* $P(X = Y) = 1$ *then* $EX = EY$

(iii) $|EX| \leq E|X|$*.*

(iv) Suppose $P(X \geq 0) = 1$ *and* $EX = 0$*. Then* $P(X = 0) = 1$*.*

♣

Theorem 5.4

Let X *be a nonnegative random variable and* $r \in (1, \infty)$*. Assume that* $E[X^r] < \infty$*. Then* $EX < \infty$*.*

♡

Note *This is an analogy for convergence in* L_p *space.*

5.3 Variance

Definition 5.3

The variance is important enough to be highlighted as a formula. If X *has finite mean* $\mu = EX$*, then its variance is*

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

Expansion of the square inside the brackets and linearity of expectation give an alternative formula for the variance:

$$\begin{aligned} \mathrm{Var}(X) &= E\left[X^2 - 2\mu X + \mu^2\right] = E\left[X^2\right] - 2\mu EX + \mu^2 \\ &= E\left[X^2\right] - \mu^2 \end{aligned}$$

♠

Theorem 5.5

(i) Suppose X *has finite variance and* a, b *are real numbers. Then* $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$

(ii) $\mathrm{Var}(X) = 0$ *if and only if there exists a real number* c *such that* $P(X = c) = 1$*. When this happens,* c *is the mean of* X*.*

♡

5.4 Linearity of expectation

Theorem 5.6

let $\mathbf{X}_i, 1 \leq i \leq n$, be random vectors defined on $(\Omega, \mathcal{F}, P)$ and for each index i let g_i be a real-valued function defined on the range of X_i. Then, as long as the expectations below are finite,

$$\begin{aligned} &E\left[g_1\left(\mathbf{X}_1\right)+g_2\left(\mathbf{X}_2\right)+\cdots+g_n\left(\mathbf{X}_n\right)\right] \\ &\quad=E\left[g_1\left(\mathbf{X}_1\right)\right]+E\left[g_2\left(\mathbf{X}_2\right)\right]+\cdots+E\left[g_n\left(\mathbf{X}_n\right)\right] \end{aligned}$$

♡

Corollary 5.2

Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space. Then

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

provided all the expectations on both sides are finite.

♡

Note *The caveat in the theorem about all expectations being finite is needed.*

5.5 Expectation and independence

Theorem 5.7

Suppose $X_1, \ldots, X_n$ are independent random variables. Then for all functions $g_1, \ldots, g_n$ for which the expectations below are well-defined,

$$E\left[\prod_{k=1}^{n} g_k\left(X_k\right)\right]=\prod_{k=1}^{n} E\left[g_k\left(X_k\right)\right]$$

♡

Theorem 5.8

Assume the random variables $X_1, \ldots, X_n$ are independent and have finite variances. Then

$$\operatorname{Var}\left(X_1+X_2+\cdots+X_n\right)=\operatorname{Var}\left(X_1\right)+\operatorname{Var}\left(X_2\right)+\cdots+\operatorname{Var}\left(X_n\right)$$

♡

5.6 Covariance and correlation

Definition 5.4

Let X and Y be random variables defined on the same sample space with expectations μ_X and μ_Y. The covariance of X and Y is

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

if the expectation on the right is finite. By expanding the product inside the expectation, we get an alternative formula for the covariance:

$$\begin{aligned} \operatorname{Cov}(X, Y) &= E\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right]=E\left[XY-\mu_X Y-\mu_Y X+\mu_X \mu_Y\right] \\ &= E[XY]-\mu_X E[Y]-\mu_Y E[X]+\mu_X \mu_Y \\ &= E[XY]-\mu_X \mu_Y \end{aligned}$$

♠

Proposition 5.2

The following statements hold when the covariances are welldefined.

(i) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

(ii) $\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$ *for real numbers* a, b.

(iii) Covariance is bilinear: namely, for random variables X_i *and* Y_j *and real numbers* a_i *and* b_j

$$\mathrm{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \,\mathrm{Cov}\,(X_i, Y_j)$$

♣

Theorem 5.9

Let $X_1, \ldots, X_n$ *be random variables with finite variances and covariances. Then*

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

♡

Definition 5.5

Let X *and* Y *be random variables such that* $\mathrm{Cov}(X, Y)$ *is finite,* $0 < \mathrm{Var}(X) < \infty$ *and* $0 < \mathrm{Var}(Y) < \infty$. *The correlation coefficient of* X *and* Y *is defined by*

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

♠

Proposition 5.3

Assume that $\mathrm{Cov}(X, Y)$ *is finite,* $0 < \mathrm{Var}(X) < \infty$ *and* $0 < \mathrm{Var}(Y) < \infty$. *The correlation coefficient has these properties.*

(i) $-1 \le \mathrm{Corr}(X, Y) \le 1$.

(ii) $\mathrm{Corr}(X, Y) = 1$ *if and only if there exist* $a > 0$ *and* $b \in \mathbb{R}$ *such that* $Y = aX + b$.

(iii) $\mathrm{Corr}(X, Y) = -1$ *if and only if there exist* $a < 0$ *and* $b \in \mathbb{R}$ *such that* $Y = aX + b$

♣

Note *(i) follows from Cauchy-Schwarz inequality.*

Chapter 6 Law of large numbers

6.1 Markov and Chebyshev inequalities

Lemma 6.1 (Markov's inequality)

Let X be a $[0,\infty]$-valued random variable and $c>0$. Then

$$P(X \geq c) \leq \frac{EX}{c}.$$

This inequality is useful only if $c > EX$.

Lemma 6.2 (Chebyshev's inequality)

Let X be a random variable with finite mean μ and finite variance σ^2. Then for any real $c>0$ we have

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

Theorem 6.1 (Weak law of large numbers, WLLN)

Let $\{X_k\}_{k\geq 1}$ be i.i.d. random variables with finite mean $\mu = E[X_k]$ and finite variance $\sigma^2 = \mathrm{Var}(X_k)$. Let $S_n = X_1 + \cdots + X_n$. Then for any fixed $\varepsilon > 0$ we have

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

Note *This is the same with $\frac{S_n}{n}$ converge in probability to μ.*

6.2 Convergence in probability and Almost surely convergence

Definition 6.1

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables and X another random variable, all defined on the same probability space $(\Omega, \mathcal{F}, P)$.

(a) We say that X_n converges to X in probability as $n\to\infty$ if, for each $\varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

A common abbreviation is $X_n \xrightarrow{P} X$.

(b) We say that X_n converges to X almost surely or with probability one if there is an event $\Omega_0 \subset \Omega$ such that $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$, $X_n(\omega) \to X(\omega)$ as $n\to\infty$. Common abbreviations are $X_n \to X$ a.s. and $X_n \to X$ w.p.1.
The definition of almost sure convergence $X_n \to X$ can be stated succinctly as $P\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$ or even more tersely as $P(X_n \to X) = 1$.

Theorem 6.2

Suppose $X_n \to X$ almost surely. Then also $X_n \to X$ in probability.

6.3 Borel-Cantelli lemma

Lemma 6.3 (Borel-Cantelli lemma)

Let $\{A_n\}_{n\geq 1}$ *be a sequence of events, all defined on the same sample space. Suppose* $\sum_{n=1}^{\infty} P(A_n) < \infty$. *Then*

$$P\{\omega : \omega \in A_n \text{ for infinitely many } n\} = 0$$

♡

6.4 Almost sure convergence from the Borel-Cantelli lemma

The next theorem encapsulates the most common strategy for proving almost sure convergence with the Borel-Cantelli lemma.

Theorem 6.3

Let $\{X_n\}_{n\geq 1}$ *and* X *be random variables on* $(\Omega, \mathcal{F}, P)$. *Suppose that for all* $\varepsilon > 0$

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \varepsilon) < \infty.$$

Then $X_n \to X$ *almost surely.*

♡

6.5 Strong law of large numbers

Theorem 6.4

Let $\{X_k\}_{k\geq 1}$ *be i.i.d. random variables with finite mean* $\mu = E[X_k]$. *Let* $S_n = X_1 + \cdots + X_n$. *Then*

$$P\left\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\right\} = 1.$$

In other words, $S_n/n \to \mu$ *almost surely.*

♡

Chapter 7 Limits in distribution

7.1 Converge in distribution

The definition of a limit in distribution is given below. In contrast with almost sure convergence and convergence in probability, for convergence in distribution the random variables do not have to be defined on the same sample space.

Definition 7.1

Suppose that for each positive integer n, X_n is a random variable with cumulative distribution function F_n. Let X be a random variable with cumulative distribution function F. Then X_n converges to X in distribution if $F_n(x) \to F(x)$ as $n \to \infty$ at each point x where F is continuous.

♠

Theorem 7.1

Suppose $X_n \to X$ in probability. Then also $X_n \to X$ in distribution.

♡

7.2 Gaussian Distribution

In preparation for the central limit theorem, we introduce the Gaussian, or normal distribution.

Definition 7.2

Let μ be real and $\sigma > 0$. A random variable X has the normal distribution with mean μ and variance σ^2 if X has density function

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

on the real line. Abbreviate this by $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

♠

Note *There is a great 3blue1brown video explaining why this is a PDF.*

Theorem 7.2

Let $Z \sim \mathcal{N}(0,1)$. Then $E(Z) = 0$ and $\mathrm{Var}(Z) = E\left(Z^2\right) = 1$.

♡

Proposition 7.1

Let μ be real, $\sigma > 0$, and $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$.
(i) Let $a \neq 0$, b real, and $Y = aX + b$. Then $Y \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right)$.
(ii) $Z = \frac{X-\mu}{\sigma}$ is a standard normal random variable.

♣

7.3 Central limit theorem

Theorem 7.3 (Central limit theorem)

Suppose $X_1, X_2, X_3, \ldots$ *are i.i.d. random variables with finite mean* $E[X_1] = \mu$ *and finite variance* $\mathrm{Var}(X_1) = \sigma^2$. *Let* $S_n = X_1 + \cdots + X_n$. *Then for any fixed* $-\infty \le a \le b \le \infty$ *we have*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$$

7.4 Normal Approximation of the binomial

Theorem 7.4 (Central limit theorem for Bernoulli random variables)

Let $0 < p < 1$ *be fixed and suppose that* $S_n \sim \mathrm{Bin}(n, p)$. *Then for any fixed* $-\infty \le a \le b \le \infty$

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

Another name for this theorem is the de Moivre-Laplace theorem, and it goes back to the early 1700s. We use this theorem as an imprecise approximation of binomial probabilities with Gaussian probabilities:

$$P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) \approx \Phi(b) - \Phi(a)$$

Note *A commonly used rule of thumb is that the approximation is good if* $np(1-p) > 10$.

7.5 Continuity correction

Corollary 7.1

A random variable $S_n \sim \mathrm{Bin}(n, p)$ *takes only integer values. Thus, if* k_1, k_2 *are integers then*

$$P(k_1 \le S_n \le k_2) = P(k_1 - 1/2 \le S_n \le k_2 + 1/2)$$

7.6 Confidence interval

$$\begin{aligned} P(|\widehat{p} - p| < \varepsilon) &= P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = P(-n\varepsilon < S_n - np < n\varepsilon) \\ &= P\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right). \end{aligned}$$

Up to this point we have used only algebra. Now comes the normal approximation:

$$\begin{aligned} P\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) &\approx \Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) \\ &= 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \\ &\ge 2\Phi(2\varepsilon\sqrt{n}) - 1 \end{aligned}$$

7.7 Possion Approximation

Theorem 7.5 (Poisson approximation of the binomial with error term)

Let $S \sim \mathrm{Bin}(n, p)$ *and* $Y \sim \mathrm{Poisson}(np)$. *Then*

$$\sum_{k=0}^{\infty} |P(S = k) - P(Y = k)| \leq 2np^2.$$

♡

Chapter 8 Generating functions

8.1 3 Probability Generating functions

Definition 8.1

The probability generating function of a random variable X is defined as the function $G_X(s) = Es^X$ (for whichever s this is defined).

The moment generating function of a random variable X is defined as the function $M_X(t) = Ee^{tX}$ (for whichever t this is defined).

The characteristic function of a random variable X is defined as the function $\phi_X(t) = Ee^{itX}$ Note: e^{itX} is a complex valued random variable. The expectation of a complex random variable Z is defined as $E\Re Z + iE\Im Z$. (You just take the expectation of real and imaginary parts separately.)

♠

Proposition 8.1 (nth derivative)

(a) When X is nonnegative and integer valued.

$$\frac{d^k}{s^k} G_X(s)\Big|_{s=0} = k!P(X = k)$$

Thus $G_X(s)$ identifies the distribution if X is nonnegative and integer valued.

(b) If $M_X(t)$ is defined in a small neighborhood of 0 then the derivatives at zero are equal to the moments (if they exist).

$$\frac{d}{dt} Ee^{tX} = E\left(e^{tX}X\right), \quad M'(0) = EX$$

(One would need to justify the fact that we can differentiate inside the expectation. If $M_X(t)$ is finite in a neighborhood of 0 then this is justified.) Same works for higher order derivatives (if the moments exist):

$$M_X^{(n)}(0) = EX^n$$

(c) The derivatives at zero will produce the moments (if they exist).

$$\frac{d}{dt} Ee^{itX} = E\left(e^{itX}X\right), \quad \phi'(0) = iEX$$

(One would again need to justify the fact that we can differentiate inside the expectation.) Same works for higher order derivatives (if the moments exist):

$$\phi_X^{(n)}(0) = i^n EX^n$$

♣

Theorem 8.1

When the moment generating function $M(t)$ of a random variable X is finite in an interval around the origin, then all moments of X are finite and are given by

$$E\left(X^n\right) = M^{(n)}(0)$$

♡

8.2 Identification of distributions with moment generating functions

Theorem 8.2

Let X and Y be two random variables with moment generating functions $M_X(t) = E\left(e^{tX}\right)$ and $M_Y(t) = E\left(e^{tY}\right)$. Suppose there exists $\delta > 0$ such that for $t \in (-\delta, \delta)$, $M_X(t) = M_Y(t)$ and these are finite numbers. Then X and Y are equal in distribution.

♡

8.3 Moment generating function of a sum of independent random variables

Theorem 8.3

Suppose that X and Y are independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then for all real numbers t

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

♡

Note *Moment generating function can help to prove central limit theorem.*

Chapter 9 Conditional Expectation

9.1 Conditional distributions

Definition 9.1 (Discrete case)

Let X and Y be discrete random variables. Let $y \in \mathbb{R}$ be point such that $P(Y = y) > 0$. Then the conditional probability mass function of X given $Y = y$ is the function $p_{X|Y}(x \mid y)$ of possible values x of X, defined as follows:

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

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

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

♠

Definition 9.2 (Continuous case)

Let X and Y be jointly continuous random variables with joint density function $f_{X,Y}(x, y)$. For those $y \in \mathbb{R}$ such that $f_Y(y) > 0$ we make the following definitions.
The conditional density function of X, given $Y = y$, is denoted by $f_{X|Y}(x \mid y)$ and defined as

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

The conditional probability that $X \in A$, given $Y = y$, is

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

The conditional expectation of X, given $Y = y$, is

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

♠

Theorem 9.1

For the conditional expectation of $g(X)$ given that $Y = y$ we have the following formulas provided the expectations are well-defined. (i) In the discrete case

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

for y such that $P(Y = y) > 0$.
(ii) In the jointly continuous case

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

for y such that $f_Y(y) > 0$.

♡

Proposition 9.1

We have the following averaging identities. Assume that the expectations of $g(X)$ below are well-defined.
(i) In the discrete case

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

and

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

(ii) In the jointly continuous case

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

and

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

♣

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

9.2 Conditioning and independence

Theorem 9.2

Discrete random variables X and Y are independent if and only if $p_{X|Y}(x \mid y) = p_X(x)$ for all possible values x of X, whenever $p_Y(y) > 0$.
Jointly continuous random variables X and Y are independent if and only if $f_{X|Y}(x \mid y) = f_X(x)$ for all x, whenever $f_Y(y) > 0$.

♡

Chapter 10 Reference

Math 531: Probability Theory