Probability - Math 394/395/396 Notes

Brett Saiki

March 2022

This is a compilation of notes from the probability sequence, Math 394/395/396, at the University of Washington. Math 394 was taught by Alexander Giessing and borrows material from *Introduction to Probability*, 1982, by D. Anderson et. al. and *A First Course in Probability*, 10th edition, 2018 by S. Ross.

Contents

1 Introduction

1.1 Fundamental Concepts

Definition 1.1. An *experiment* is any activity or process whose outcome is subject to uncertainty.

Definition 1.2. An *sample space* of an experiment is the set of all possible outcomes of the experiment. We denote the sample space by Ω.

Definition 1.3. An *event*, A, is a subset of a sample space, Ω, that is, $A \subseteq \Omega$. Let $\omega \in \Omega$ be the outcome of an experiment. We say that the *event A occurs* if $\omega \in A$.

Definition 1.4. A *simple event* is a subset of the sample space that contains only one outcome.

1.2 Laplace Distribution

Definition 1.5. Let N give the number of simple events in an event. Suppose all outcomes of an experiment with finite sample space Ω are equally likely. Then, for all events $A \subseteq \Omega$,

$$\mathbb{P}(A) = \frac{N(A)}{N(\Omega)}.$$

We call $\mathbb{P}$ the *Laplace distribution (over Ω).*

Lemma 1.6. The Laplace distribution $\mathbb{P}$ over Ω has the following properties:

(i) $\mathbb{P}(\Omega) = 1$.

(ii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for disjoint events A and B.

1.3 Probability and Set Theory

Theorem 1.7 (DeMorgan's Law). For any events A and B we have

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

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

Definition 1.8. Given $A_1, \ldots, A_n \subseteq \Omega$ we define

$$\bigcup_{k=1}^{n} A_k = A_1 \cup \cdots \cup A_n = \{\omega \in \Omega \mid \exists k \in \{1, \ldots, n\} : \omega \in A_k\},$$
$$\bigcap_{k=1}^{n} A_k = A_1 \cap \cdots \cap A_n = \{\omega \in \Omega \mid \forall k \in \{1, \ldots, n\} : \omega \in A_k\}.$$

Theorem 1.9. Given $A_1, \ldots, A_n \subseteq \Omega$,

(i) $\left(\bigcup_{k=1}^{n} A_k\right)^c = \bigcap_{k=1}^{n} A_k^c$

(ii) $\left(\bigcap_{k=1}^{n} A_k\right)^c = \bigcup_{k=1}^{n} A_k^c$

Definition 1.10. Let $(A_k)_{k=1}^{\infty}$ be a sequence of subsets in Ω and define

$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty} A_k = \{\omega \in \Omega \mid \exists n \geq 1 : \forall k \geq n : \omega \in A_k\},$$
$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k = \{\omega \in \Omega \mid \forall n \geq 1 : \exists k \geq n : \omega \in A_k\}.$$

1.4 Axioms of Probability Theory

Definition 1.11. Let Ω be a finite sample space and $\mathcal{A}$ be the collection of all subsets of Ω. A *probability measure on* $(\Omega, \mathcal{A})$ is a function $\mathbb{P}$ from $\mathcal{A}$ into the real numbers that satisfies

(i) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$;

(ii) $\mathbb{P}(\Omega) = 1$;

(iii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for all pairwise disjoint $A, B \in \mathcal{A}$.

The number of $\mathbb{P}(A)$ is called the probability that event A occurs. These properties are called *non-negativity*, *normalization*, and *additivity*.

Definition 1.12. A collection $\mathcal{A}$ of subsets of Ω is called a σ-algebra if it satisfies the following conditons:

(i) $\emptyset \in \mathcal{A}$;

(ii) if $A \in \mathcal{A}$, then $A^c \in \mathcal{A}$;

(iii) if $A_1, A_2, \ldots \in \mathcal{A}$, then $\bigcup_{k=1}^{\infty} A_k \in \mathcal{A}$.

Properties (ii) and (iii) are called *closed under complement* and *countable additivity.*

Theorem 1.13. The smallest σ-algebra associated with Ω is $\mathcal{A} = \{\emptyset, \Omega\}$.

Theorem 1.14. If Ω is finite, then the power set 2^{Ω} is a σ-algebra.

Theorem 1.15. If A is any subset of Ω, then $\mathcal{A} = \{\emptyset, A, A^c, \Omega\}$ is a σ-algebra.

Definition 1.16. Let Ω be a sample space and $\mathcal{A}$ be a σ-algebra on Ω. A *probability measure on* $(\Omega, \mathcal{A})$ is a function $\mathbb{P}$ from $\mathcal{A}$ into the real numbers that satisfies

(i) $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{A}$;

(ii) $\mathbb{P}(\Omega) = 1$;

(iii) if $A_1, A_2, \ldots \in \mathcal{A}$ is a collection of pairwise disjoint events, in that $A_j \cap A_k = \emptyset$ for all pairs j, k satisfying $j \neq k$, then
$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k).$$

The triplet $(\Omega, \mathcal{A}, \mathbb{P})$ is called a *probability space.*

Lemma 1.17. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $A, B \subseteq \Omega$.

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;

(ii) if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$;

(iii) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Lemma 1.18 (Inclusion-Exclusion Formula). For any events $A_1, \dots, A_n$ we have

$$\mathbb{P}(A_1 \cup \dots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{1 \le i_1 < \dots < i_k \le n} \mathbb{P}(A_1 \cap \dots \cap A_k).$$

For $n = 2$, this equation simplifies to (iii) of Lemma 1.17.

Theorem 1.19. Let $A_1, A_2, \cdots$ be an increasing sequence of events, i.e. $A_1 \subset A_2 \subset A_3 \subset \cdots$, then

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \bigcup_{k=1}^{\infty} \mathbb{P}(A_k).$$

Let $B_1, B_2, \cdots$ be an decreasing sequence of events, i.e. $B_1 \supset B_2 \supset B_3 \supset \cdots$, then

$$\lim_{n\to\infty} \mathbb{P}(B_n) = \bigcap_{k=1}^{\infty} \mathbb{P}(B_k).$$

2 Combinatorics

2.1 Urn Models

Definition 2.1 (Falling Factorial). For $r \in \mathbb{R}$ and $k \in \mathbb{N}$ we define $(r)_k$, "r falling k", as

$$(r)_k = r \cdot (r-1) \cdots (r-k+1).$$

Definition 2.2 (Factorial). For $n \in \mathbb{N}$ we define $n!$, "n factorial", as

$$n! = \begin{cases} n \cdot (n-1) \cdots 2 \cdot 1 & \text{for } n > 1, \\ 1 & \text{for } n = 0. \end{cases}$$

Definition 2.3 (Binomial Coefficient).

For $r \in \mathbb{R}$ and $n \in \mathbb{N}$ we define binomial coefficient $\binom{r}{n}$, "r choose n" as

$$\binom{r}{n} = \frac{r \cdot (r-1) \cdots (r-n+1)}{n!} = \frac{r!}{n!(r-n)!}.$$

For $r \in \mathbb{R}$ and $n \in \mathbb{Z}$, $n \ge 0$ we define the binomial coefficient $\binom{r}{n}$ as

$$\binom{r}{n} = \begin{cases} 1 \text{ if } n = 0, \\ 0 \text{ if } n < 0. \end{cases}$$

Theorem 2.4 (Vandermonde's identity). For non-negative integers $m, n, r, k \in \mathbb{N}_0$,

$$\binom{m+n}{r} = \sum_{k=0}^{r} \binom{m}{k} \binom{n}{r-k}.$$

Definition 2.5 (Urn Model of Laplace experiments). Consider an urn with n balls which are labeled $1, \dots, n$. An urn model is an experiment in which k times a ball is drawn at random from the urn and its number is noted.

Definition 2.6 (Urn Model I, "Ordered Sampling with Replacement"). Draw k times from an urn with n balls. The number and the order of the ball are noted and the ball is put back into the urn. The outcome is $\omega = (a_1, \dots, a_k)$ where a_i is the number of the ith draw (i.e. a k-tuple with values $\{1, \dots, n\}$). The sample space is

$$\Omega_I = \{(a_1, \dots, a_k) \mid a_1, \dots, a_k \in \{1, \dots, n\}\}.$$

(i.e. all possible k-tuples with values in $\{1, \dots, n\}$).

Lemma 2.7. The cardinality of the set Ω_I is $|\Omega_I| = n^k$.

Definition 2.8 (Urn Model II, "Ordered Sampling without Replacement"). Draw k times from an urn with n balls. The number and the order of the ball are noted and the ball is not returned to the urn. The outcome is $\omega = (a_1, \dots, a_k)$ where a_i is the number of the ith draw (i.e. an arrangement of k elements of $\{1, \dots, n\}$). The sample space is

$$\Omega_{II} = \{(a_1, \dots, a_k) \mid a_1, \dots, a_k \in \{1, \dots, n\}, a_i \neq a_j \text{ for } i \neq j\}.$$

Lemma 2.9. The cardinality of the set Ω_{II} is $|\Omega_{II}| = (n)_k = n \cdot (n-1) \cdots (n-k+1)$.

Definition 2.10 (Urn Model III, "Unordered Sampling without Replacement"). Draw k times from an urn with n balls. The number of the ball is noted but not the order, and the ball is not returned to the urn. The outcome is $\omega = (a_1, \dots, a_k)$ (i.e. subsets of $\{1, \dots, n\}$ of size k). The sample space is

$$\Omega_{III} = \{\omega \subseteq \{1, \dots, n\} \mid |\omega| = k\}$$

(i.e. all possible subsets of $\{1, \dots, n\}$ of size k).

Lemma 2.11. The cardinality of the set Ω_{III} is

$$|\Omega_{III}| = \binom{n}{k} = \frac{(n)_k}{k!} = \frac{n \cdots (n-1) \cdots (n-k+1)}{k!}.$$

Definition 2.12 (Urn Model IV, "Unordered Sampling with Replacement"). Draw k times from an urn with n balls. The number of the ball is noted but not the order, and the ball is returned to the urn. The outcome is $\omega = (k_1, \dots, k_n)$ where k_i denotes how often the ith ball was drawn (i.e. a tuple whose values sum up to k). The sample space is

$$\Omega_{IV} = \{(k_1, \dots, k_n) \mid k_i \in \mathbb{N}_0, k_1 + \ldots + k_n = k\}.$$

Lemma 2.13. The cardinality of the set Ω_{IV} is

$$|\Omega_{IV}| = \binom{k+n-1}{n-1} = \binom{k+n-1}{k}.$$

2.2 Discrete Probability Spaces

Definition 2.14. A probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is called *discrete* if there exists a finite or countable infinite subset $D \subseteq \Omega$ such that $\mathbb{P}(D) = 1$. The associated probability measure is also called *discrete*.

Lemma 2.15. Any discrete probability measure, $\mathbb{P}$ satisfies

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}),$$

that is, a discrete probability measure $\mathbb{P}$ is fully characterized by its values on simple events.

Lemma 2.16. Let $p : \Omega \to \mathbb{R}$ be a function that satisfies the following:

(i) $p(\omega) = 0$, except for countable many $\omega \in \Omega$,

(ii) $p(\omega) \geq 0$ for all $\omega \in \Omega$,

(iii) $\sum_{\omega \in \Omega} p(\omega) = 1$.

Then p is a probability measure on $(\Omega, \mathcal{A})$ and we call p the *probability (mass) function.*

Definition 2.17 (Urn Model with Colored Balls). Consider an urn with n balls which are labeled $1, \ldots, N$ with balls $\{1, \ldots, R\}$ being one color and $\{R+1, \ldots, N\}$ being another color. We draw n times a ball at random from the urn and note its number and/or color.

2.3 Hypergeometric Distribution

Definition 2.18 (Hypergeometric Distribution). Under the urn model with colored balls, draw n balls at once from the urn. Consider the event E_r where exactly r balls are the first color, then

$$E_r = \{A \subseteq \{1, \ldots, N\} : |A| = n, |A \cap \{1, \ldots, R\}| = r, |A \cap \{R+1, \ldots, N\}| = n - r\},$$

and

$$\Omega = \{\omega \subset \{1, \ldots, N\} : |\omega| = n\}.$$

Lemma 2.19. Define the probability mass function of the hypergeometric distribution as

$$p(r) = \frac{\binom{R}{r}\binom{N-R}{n-r}}{\binom{N}{n}} \quad \text{for} \quad r \in \{0, 1, \ldots, n\}.$$

Then $\mathbb{P}(E_r) = p(r)$.

2.4 Binomial Distribution

Definition 2.20 (Binomial Distribution). Under the urn model with colored balls, draw n times from the urn with replacement. Consider the event E_r where exactly r balls are the first color, then

$$E_r = \{(a_1, \ldots, a_n) : |\{i : a_i \in \{1, \ldots, R\}\}| = r\},$$

and

$$\Omega = \{(a_1, \ldots, a_n) : a_1, \ldots, a_n \in \{1, \ldots, N\}\}.$$

Lemma 2.21. Define the probability mass function of the binomial distribution as

$$p(r) = \binom{n}{r} \left(\frac{R}{N}\right)^n \left(1 - \frac{R}{N}\right)^{n-r} \quad \text{for} \quad r \in \{0, 1, \ldots, n\}.$$

Then $\mathbb{P}(E_r) = p(r)$.

2.5 Multinomial Distribution

Definition 2.22 (Urn Model With Many Colored Balls). Consider an urn with N balls which are labeled $1, \ldots, N$ with the first N_1 balls of color 1, the second N_2 balls of color 2, $\ldots$, the last N_r balls of color r. We draw n times a ball at random from the urn and its number and/or color is noted.

Lemma 2.23. The number of possible ways in which a set A with cardinality $|A| = k$ can be partitioned into n subsets $A_1, \ldots, A_n$ with cardinalities $k_1, \ldots, k_n$ such that $k_1 + \ldots + k_n = n$ is given by

$$\frac{k!}{k_1! \cdots k_n!}.$$

Definition 2.24. For $k, k_1, \ldots, k_n \in \mathbb{Z}$ we define *multinomial coefficient* as

$$\binom{k}{k_1, \ldots, k_n} = \begin{cases} \frac{k!}{k_1! \cdots k_n!} & \text{if } k_1 \geq 0, \text{ and } \text{ and } \sum_{i=1}^n k_i = k, \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.25 (Multinomial Distribution). Under the urn model with many colored balls, draw n balls with r colors with replacement. Consider the event $E_{n_1,\ldots,n_r}$, where exactly n_1 balls are of one color, n_2 balls are of the second color, and so on, can be written as

$$E_{n_1,\ldots,n_r} = \{(a_1, \ldots, a_n) : |\{i : a_i \in \{N_{k-1} + 1, \ldots, N_k\}\}| = n_k, k \in \{1, \ldots, r\}\},$$

where $N_0 = 0, N_1 + \cdots N_r = N$ and $n_1 + \cdots + n_r = n$, and

$$\Omega = \{(a_1, \ldots, a_n) : a_1, \ldots, a_n \in \{1, \ldots, N\}\}.$$

Lemma 2.26. Define the probability mass function of the multinomial distribution as

$$p(n_1, \ldots, n_r) = \binom{n}{n_1, \ldots, n_r} \prod_{k=1}^{r} \left(\frac{N_k}{N}\right)^{n_k},$$

for $n_1, \ldots, n_r \in \mathbb{N}_0$ and $n_1 + \cdots + n_r = n$. Then $\mathbb{P}(E_{n_1,\ldots,n_r}) = p(n_1, \ldots, n_r)$.

3 Independence and Conditional Events

3.1 Independence

Definition 3.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability triple. Two events A, B on $(\Omega, \mathcal{A}, \mathbb{P})$ are called *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Definition 3.2. The events $A_1, \ldots, A_n$ are called *independent* if for each $k \in \{1, \ldots, n\}$ and each collection of indices $1 \leq i_1 < \ldots < i_k \leq n$

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k}).$$

Lemma 3.3. Let $A_1, \ldots, A_n$ be independent events. Consider events $B_1, \ldots, B_n$ such that

$$B_i = A_i \qquad \text{or} \qquad B_i = A_i^c.$$

Then the events $B_1, \ldots, B_n$ are independent.

Definition 3.4. Let $(\Omega_i, \mathcal{A}_i, \mathbb{P}_i)$ be discrete probability spaces with $\mathbb{P}_i$ characterized by the probability mass function $p_i : \Omega_i \to [0, 1]$, $i = 1, \ldots, n$. The *product space* $(\Omega, \mathbb{P})$ is the discrete probabilty space with sample space

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \ldots, \omega_n) : \omega_i \in \Omega_i, 1 \leq i \leq n\},$$

and *product measure* $\mathbb{P}$ defined by the probability mass function

$$p(\omega_1, \ldots, \omega_n) = p_1(\omega_1) \cdots p_n(\omega_n).$$

Lemma 3.5. Let $A_i \in \Omega_i$ be any event concerning only the *ith* experiment and let A_i' be defined by

$$A_i' = \{\omega : \omega \in \Omega, \omega_i \in A_i\},$$

for $1 \leq i \leq n$. Then

$$\mathbb{P}(A_i') = \mathbb{P}_i(A_i) \quad \text{for all} \quad i = 1, \ldots, n,$$

and the events $A_1', \ldots, A_n'$ are stochastically independent.

3.2 Conditional Probability

Definition 3.6. Let $A, B \subseteq \Omega$ be events such that $\mathbb{P}(A) > 0$. The conditional probability of B given A is defined by as

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Lemma 3.7. Events $A, B \subset \Omega$ are independent if and only if $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

Lemma 3.8 (Multiplication Rule). Let $A_1, \ldots, A_n \subseteq \Omega$ be events with $\mathbb{P}(A_1 \cap \ldots \cap A_{n-1} \neq 0)$. Then,

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2 \mid A_1) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1}).$$

Definition 3.9. Events $A_1, \ldots, A_n \subseteq \Omega$ are a *disjoint partition of* Ω when $B_1 \cup \cdots \cup B_n$ and $B_i \cap B_j = \emptyset$ for $i \neq j$.

Lemma 3.10 (Law of Total Probability). Let $B_1, \ldots, B_n$ be a disjoint partition of Ω. If $\mathbb{P}(B_i) > 0$ for all $1 \leq i \leq n$, then for any event $A \subseteq \Omega$,

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i).$$

Lemma 3.11 (Bayes' Rule). Let $B_1, \ldots, B_n$ be a disjoint partition of Ω. If $\mathbb{P}(B_i) > 0$ for all $1 \leq i \leq n$, then for any events $A \subseteq \Omega$ and $B_k \subseteq \Omega$,

$$\mathbb{P}(B_k \mid A) = \frac{\mathbb{P}(A \mid B_k)\mathbb{P}(B_k)}{\sum_{i=1}^{n} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)}.$$

Definition 3.12. In the previous lemma, Lemma 3.11, $\mathbb{P}(B)$ is called the *prior* probability of B and $\mathbb{P}(B \mid A)$ is called the *posterior* probability of B given A.

Lemma 3.13 (Gambler's Ruin). Choose p to be some number such that $0 < p < 1$, choose an integer x such that $0 \leq x \leq K$ for some bound K, and let $q = 1 - p$. Consider a sequence $\{a_n\}$ generated by the following method:

$$a_n = \begin{cases} 0, & a_{n-1} = 0 \\ 1, & a_{n-1} = K \\ a_{n-1} + 1, & \text{with probability } p \\ a_{n-1} - 1, & \text{with probability } q. \end{cases}$$

That is, a_n moves by one in either direction but terminates once it reaches 0 or K. Let A_x be the event that a_n terminates at 0.

(i) If $p \neq q$, then the probability that A_x occurs is

$$\mathbb{P}(A_x) = \frac{(q/p)^x - (q/p)^K}{1 - (q/p)^K}.$$

(ii) If $p = q = 1/2$, then the probability that A_x occurs is

$$\mathbb{P}(A_x) = 1 - \frac{x}{K}.$$

Definition 3.14. A *linear first-order difference equation* is a recursive formula of the form

$$x_{t+1} = ax_t + b, \quad \text{for} \quad t = 0, 1, \ldots$$

where $a \neq 1$ and b are constants.

Lemma 3.15. The solution to the first-order linear difference equation is

$$x_t = a\left(x_0 - \frac{b}{1-a}\right) + \frac{b}{1-a}.$$

4 Discrete Random Variables

4.1 Random Variables

Definition 4.1 (Random Variable). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is called *measureable* if for all $\alpha \in \mathbb{R}$

$$\{\omega \in \Omega : X(\omega) \leq \alpha\} \in \mathcal{A}.$$

We call such a function a *random variable.*

Remark. In discrete probability spaces, the σ-algebra $\mathcal{A}$ is usually the power set 2^Ω, and therefore every function $X : \Omega \to \mathbb{R}$ is a random variable. For more generate probability spaces, this is not generally true.

Definition 4.2. If $X(\omega) = x$ for some $\omega \in \Omega$, we call x the *realization* or *observed value* of $X(\omega)$.

Remark. We often drop ω and write X instead of $X(\omega)$ and thus denote events of Ω by

$$\{X = a\} = \{\omega \in \Omega : X(\omega) = a\}.$$

Lemma 4.3. A random variable X defines a probability measure $\mathbb{P}_X$ on $\mathbb{R}$ by assigning each $A \subset \mathbb{R}$ the probability that X takes a value in A:

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}).$$

When $X^{-1}(A)$ is an event in $\mathcal{A}$,

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)).$$

Lemma 4.4. Given a random variable X and a set $A \subset \mathbb{R}$, $X^{-1}(A) \in \mathcal{A}$ if X is measureable and A is Borel-measureable subset of $\mathbb{R}$.

Lemma 4.5. For our purposes it suffices to know that all intervals and all open and closed subsets of $\mathbb{R}$ are Borel-measureable.

Definition 4.6. Let X be a random variable on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The probability distribution $\mathbb{P}_X$ on $\mathbb{R}$ defined by

$$\mathbb{P}_X(A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\}) \qquad A \subset \mathbb{R} \text{ is measureable}$$

is called the distribution of X. We generally denote $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$ by $\mathbb{P}(X \in A)$.

4.2 Discrete Random Variables

Definition 4.7. A random variable X is called *discrete*, if there exists a finite or countably infinite subset $D \subseteq \mathbb{R}$ such that $\mathbb{P}(X \in D) = 1$.

Definition 4.8. Let X be a discrete random variable with range $\{x_1, x_2, \ldots\}$. The function $p : X(\Omega) \to \mathbb{R}$ defined by

$$p(x_i) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x_i\}) = \mathbb{P}(X = x_i).$$

is called the *probability mass function* of X. It is convenient to extend p to all of $\mathbb{R}$ by assigning $p(x) = 0$ for $x \in \mathbb{R} \setminus X(\Omega)$.

Lemma 4.9. Let X be a discrete random variable with range $X(\Omega) = \{x_1, x_2, \ldots\}$. Then x has a probability mass function that satisfies the following

(i) $p(x_i) \geq 0$,

(ii) $\sum_{i=1}^{\infty} p(x_i) = 1$.

Lemma 4.10. If a function $p : \mathbb{R} \to \mathbb{R}$ satisfies properties (i) and (ii) from Lemma 4.9, then it is a probability mass function for some random variable.

4.3 Distributions of Discrete Random Variables

Definition 4.11 (Laplace Distribution). A discrete random variable X has a *Laplace distribution* (or uniform distribution) on $\{1, 2, \ldots, N\}$ if its probability mass function is given by

$$p_X(k) = \mathbb{P}(X = k) = \frac{1}{N} \quad \text{for} \quad k \in \{1, 2, \ldots, N\}.$$

Definition 4.12. A *Bernoulli trial* (or binomial trial), X on $\Omega = \{S, F\}$ by

$$X(\omega) = \begin{cases} 1, & \omega = S, \\ 0, & \omega = F. \end{cases}$$

Usually, S is called a "success" and F is a "failure".

Definition 4.13 (Bernoulli Distribution). A Bernoulli trial X has a *Bernoulli distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(1) = \mathbb{P}(X = 1) = p \quad \text{and} \quad p_X(0) = \mathbb{P}(X = 0) = 1 - p.$$

We denote this distribution by $\text{Ber}(p)$.

Definition 4.14 (Binomial Distribution). A discrete random variable X has a *binomial distribution* with parameters n and p if its probability mass function is given by

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

for $k = 0, 1, \ldots, n$. We denote this distribution by $\text{Binom}(n, p)$.

Lemma 4.15. Let $0 \leq p \leq 1$ be some probability and let $n \in \mathbb{N}$ be an integer. Suppose $X = Y_1 + Y_2 + \ldots Y_n$ is a discrete random variable where each Y_i is an independent and identically distributed random variable with a Bernoulli distribution of parameter p. Then X has a binomial distribution with parameters n and p.

Definition 4.16 (Geometric Distribution). A discrete random variable X has a *geometric distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

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

for $k = 1, 2, \dots$. We denote this distribution by $\mathrm{Geo}(p)$.

Remark. The geometric distribution is obtained by running an infinite sequence of independent Bernoulli trials. X is the random variable defined by the number of trials conducted until the first "success" occurs.

Definition 4.17 (Negative Binomial Distribution). A discrete random variable X has a *negative binomial distribution* with parameters r and p, where $r \in \mathbb{N}$ and $0 \leq p \leq 1$ if its probability mass function is given by

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

for $k = 0, 1, 2, \dots$. We denote this distribution by $\mathrm{NB}(r, p)$.

Remark. The negative binomial distribution is obtained by counting the number of "failures" before r "successes" occur.

Definition 4.18 (Hypergeometric Distribution). A discrete random variable X has a *hypergeometric distribution* with parameter N, M, and n if its probability mass function is given by

$$p_X(x) = \mathbb{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where $\max\{0, n - N + M\} \leq x \leq \min\{n, M\}$. We denote this distribution by $\mathrm{Hypergeo}(N, M, n)$.

Theorem 4.19 (Poisson Limit Theorem). Let $X_1, X_2, \dots$ be a sequence of $\mathrm{Binom}(n, p_n)$ distributed random variables. Suppose for some $\lambda \in (0, \infty)$, $np_n \to \lambda$ as $n \to \infty$. Then for all $k = 0, 1, 2, \dots,$

$$\lim_{n\to\infty} \mathbb{P}(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Moreover, $p_\lambda(k) = e^{-\lambda}\lambda^k/k!$ is a probability mass function on $k = 0, 1, 2, \dots$.

Definition 4.20 (Poisson Distribution). A discrete random variable X has a *Poisson distribution* with parameter $\lambda > 0$, if its probability mass function is given by

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

for $k = 0, 1, 2, \dots$. We denote this distribution by $\mathrm{Pois}(\lambda)$.

Remark. If $X \sim \mathrm{Binom}(n, p)$ is a random variable where n is sufficiently large, then X can be approximated by $\mathrm{Pois}(np)$.

4.4 Expectation, Variance, and Transformations

Definition 4.21 (Expected Value of a Discrete Random Variable). Let X be a discrete random variable with probability mass function p. We define the *expected value* (also called the expectation or the mean) of X to be

$$\mathbb{E}[X] = \sum_{x \to X(\Omega)} x \cdot p(x).$$

We say that the expected value of X exists if $\sum_x |x|p(x) < \infty$.

Theorem 4.22. Let X be a discrete random variable with probability mass function p and let $g : X(\Omega) \to R$ be a map such that $\sum_x |g(x)|p(x) < \infty$. Then

$$\mathbb{E}[g(X)] = \sum_{x \in X(\Omega)} g(x) \cdot p(x).$$

Theorem 4.23 (Triangle Inequality for the Expected Value). Let X be a discrete random variable whose expected value exists. Then

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|].$$

Theorem 4.24 (Linearity of the Expected Value). Let X, Y be two discrete random variables whose expected values exists. Then for arbitrary $a, b \in \mathbb{R}$,

(i) $\mathbb{E}[aX] = a\mathbb{E}[X]$,

(ii) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$,

(iii) $\mathbb{E}[b] = b$.

Definition 4.25. Let X be a random variable such that $\mathbb{E}[X^2] < \infty$. We define the *variance* of X as

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

The square root of the variance is called the *standard deviation.*

Theorem 4.26. Let X be a random variable. The following holds true:

(i) $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$ for all $a, b \in \mathbb{R}$,

(ii) $\mathrm{Var}(X) = \mathbb{E}[X^2] + (\mathbb{E}[X])^2$.

Theorem 4.27. Let X be a random variable and $a \in \mathbb{R}$ be an arbitrary number. Then,

$$\mathbb{E}[(X - a)^2] \geq \mathrm{Var}(x),$$

and equality holds if and only if $a = \mathbb{E}[X]$.

Theorem 4.28 (Markov's Inequality). Let X be a random variable and $a > 0$ be arbitrary. Then

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

Theorem 4.29 (Chebychev's Inequality). Let X be a random variable and $a > 0$ be arbitrary. Then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(x)}{a^2}.$$

Corollary 4.30. Let X be a random variable and $a > 0$ be arbitrary. Then

$$\mathbb{P}\left(|X - \mathbb{E}[X]| < a\sqrt{\mathrm{Var}(X)}\right) > 1 - \frac{1}{a^2}.$$

Theorem 4.31 (Weak Law of Large Numbers for Bernoulli Experiments). Let S_n be the number of successes in n independent Bernoulli Experiments with success probability p. Given $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{\epsilon^2 n},$$

and the right-hand side converges to 0 as $n \to \infty$.

Definition 4.32. For a continuous function $f : [0,1] \to \mathbb{R}$ defined the *Bernstein polynomial* as

$$B_n^f(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$

Theorem 4.33. For every continuous function $f : [0,1] \to \mathbb{R}$,

$$\sup_{0 \le x \le 1} \left| B_n^f(x) - f(x) \right| \to 0 \quad \text{as} \quad n \to \infty,$$

i.e. the sequence of Bernstein polynomials converges uniformly to f.

5 Continuous Random Variables

5.1 Continuous Random Variables

Definition 5.1. An integrable, non-negative function f is called *probabilty density function* of the random variable X (or of its distribution $\mathbb{P}_X$), if for all $a, b \in \mathbb{R}$ with $a \le b$,

$$\mathbb{P}(a < X \le b) = \mathbb{P}_X((a,b]) = \int_a^b f(x)\, dx.$$

A distribution with a probability density function is called a *continuous distribution.*

Lemma 5.2. If f is a probability density function, then

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Lemma 5.3. Let X be a continuous random variable. The distribution of X does not uniquely determine the probability density function f.

Definition 5.4. A continuous random variable X has the *uniform distribution over the interval* $[a, b]$ if it has the probability density function

$$f(x) = \begin{cases} 1/(b-a), & a \le x \le b \\ 0, & \text{otherwise.} \end{cases}$$

The uniform distribution is denoted by $\mathrm{Unif}(a, b)$ and is the continuous analog to the Laplace distribution.

Definition 5.5. A continuous random variable X has the *exponential distribution with rate parameter* $\lambda > 0$ if it has the probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

The exponential distribution is denoted by $\mathrm{Exp}(\lambda)$ and is the continuous analog to the geometric distribution.

Lemma 5.6. Let $X \sim \mathrm{Exp}(\lambda)$ be a continuous random variable with the exponential distribution. Then X has the memoryless-ness property, that is,

$$\mathbb{P}(X \ge s + t \mid X \ge s) = \mathbb{P}(X \ge t)$$

for all $s, t \ge 0$.

Definition 5.7. A continuous random variable X has the *normal (Gaussian) distribution with mean* μ *and variance* σ^2 if it has the probability density function

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

The normal distribution is denoted by $\mathcal{N}(\mu, \sigma^2)$.

5.2 Cumulative Distribution Function

Definition 5.8. Let X be a random variable. We define its *cumulative distribution function* $F : \mathbb{R} \to [0, 1]$ as

$$F(x) = \mathbb{P}(X \geq x),$$

i.e. $F(x)$ is the probability that the observed value of X is less or equal to x. If X is discrete with PMF p, then

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

If X is continuous with PDF f, then

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

Theorem 5.9. The cumulative distribution function F of a random variable X has the following properties:

(i) F is monotone increasing, i.e for all $s, t \in \mathbb{R}$, $F(s) \leq F(t)$ whenever $s \leq t$;

(ii) F is right-continuous, i.e. for all $x \in \mathbb{R}$, $\lim_{y \to x^+} F(y) = F(x)$;

(iii) F has the following behavior at infinities: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

5.3 Expectation and Variance

Lemma 5.10. Let X be a random variable with values in $I \subseteq \mathbb{R}$ and PDF f_X. Let $u : I \to J$ and suppose that u, u^{-1} are continuously differentiable on I and J, respectively. Then, the random variable $Y = u(X)$ has PDF

$$f_Y(y) = \begin{cases} f_X(u^{-1}(y)) \left| \frac{d}{dy} u^{-1}(y) \right|, & y \in J, \\ 0, & y \in \mathbb{R} \setminus J. \end{cases}$$

Definition 5.11. Let X be a continuous random variable with PDF f. We say that the expected value of X exists if $\int |x| f(x)\, dx < \infty$, and we define the expected value of X as

$$\mathbb{E}[X] = \int x f(x)\, dx.$$

Theorem 5.12. Let X be a continuous random variable with PDF f and $g : \mathbb{R} \to \mathbb{R}$ be a measurable map. If $\int |g(x)| f(x)\, dx < \infty$, then we have

$$\mathbb{E}[g(X)] = \int g(x) f(x)\, dx.$$

Lemma 5.13. Let X be a random variable (continuous or discrete) and $p \geq 0$ be arbitrary. If $\mathbb{E}[|X|^p] < \infty$, then

$$\mathbb{E}[|X|^q] < \infty \qquad \text{for all} \qquad q \in [0, p].$$

6 Joint Distributions

6.1 Definition

Definition 6.1. Let $X_1, \dots, X_n$ be random variables on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The probability distribution $\mathbb{P}_X$ (or $\mathbb{P}_{X_1,\dots,X_n}$) on $\mathbb{R}^n$ defined by

$$\mathbb{P}_X(A) = \mathbb{P}_{X_1,\dots,X_n}(A) = \mathbb{P}((X_1, \dots, X_n) \in A)$$

for measurable $A \subseteq \mathbb{R}^n$ is called the *joint distribution* of $X_1, \dots, X_n$.

Definition 6.2. Let $X_1, \dots, X_n$ be random variables on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The probability distribution $p_X : X(\Omega) \to \mathbb{R}$ defined by

$$p_X(x) = p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n),$$

is called the *joint probability mass function* of $X_1, \dots, X_n$ or the probability mass function of the random vector $X = (X_1, \dots, X_n)$.

Definition 6.3. We call random variables $X_1, \dots, X_n$ *independent* if, for all intervals $I_1, \dots, I_n \subseteq \mathbb{R}$,

$$\mathbb{P}(X_1 \in I_1, \dots, X_n \in I_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in I_i).$$

Lemma 6.4. The random variables $X_1, \dots, X_n$ are independent if and only if the events $\{X_1 \in I_1\}, \dots, \{X_n \in I_n\}$ are independent for all intervals $I_1, \dots, I_n \subseteq \mathbb{R}$.

6.2 Discrete Joint Distributions

Lemma 6.5. Let $X_1, \dots, X_n$ be discrete random variable with joint probability mass function $p_x(x_1, \dots, x_n)$. Then the marginal probability mass function of $X_{i_1}, \dots, X_{i_k}$ is

$$p_{i_1,\dots,i_k}(x_{i_1}, \dots, x_{i_k}) = \sum_{x_{j_1},\dots,x_{j_{n-k}}} p_X(x_1, \dots, x_n),$$

where the indices $\{j_1, \dots, j_{n-k}\}$ are the complement of the indices $\{i_1, \dots, i_k\}$ in $\{1, \dots, n\}$.

Theorem 6.6. Let $X_1, \dots, X_n$ be discrete random variables with joint probability mass functions $p_X(x_1, \dots, x_n)$ and let $g : \mathbb{R}^n \to \mathbb{R}$. Then,

$$\mathbb{E}[g(X_1, \dots, X_n)] = \sum_{(x_1,\dots,x_n) \in X(\Omega)} g(x_1, \dots, x_n) \cdot p_X(x_1, \dots, x_n).$$

Theorem 6.7. Discrete random variables X, Y are independent if and only if

$$p_{(X,Y)}(x, y) = p_X(x) \cdot p_Y(y) \qquad \text{for all} \qquad x, y \in \mathbb{R}.$$

Theorem 6.8 (Convolution Formula for Discrete Random Variables). Let X, Y be two independent, discrete random variables with PMFs p and q. Then the random variable $Z = X + Y$ has PMF

$$r(z) = \sum_{x \in X(\Omega)} p(x) q(z - x) = \sum_{y \in Y(\Omega)} p(z - y) q(y).$$

6.3 Continuous Joint Distributions

Definition 6.9. An integrable, non-negative $f : \mathbb{R}^n \to \mathbb{R}$ is called joint probability density function (joint PDF) of the random variables $X_1, \ldots, X_n$ (or of its distribution $\mathbb{P}_{X_1,\ldots,X_n}$), it for all rectangles $R = (a_1, b_1] \times \cdots \times (a_n, b_n] \subseteq \mathbb{R}^n$,

$$\mathbb{P}((X_1, \ldots, X_n) \in R) = \mathbb{P}_{X_1,\ldots,X_n}(R) = \int_R f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Theorem 6.10. The formula in Definition 6.9 is valid for all regular domains $A \subseteq \mathbb{R}^n$:

$$\mathbb{P}((X_1, \ldots, X_n) \in A) = \int_A f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Lemma 6.11. Let X and Y be two continuous random variables with joint probability density function $f(x, y)$. Then the marginal probability density function of X is

$$f_X(x) = \int f(x, y)\, dy.$$

Theorem 6.12. Let $X_1, \ldots, X_n$ be continuous random variables with joint probability density function f and let $g : \mathbb{R}^n \to \mathbb{R}$. Then,

$$\mathbb{E}[g(X_1, \ldots, X_n)] = \int g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n.$$

Theorem 6.13. The continuous random variables X and Y are independent if and only if

$$f_{(X,Y)}(x, y) = f_X(x) \cdot f_Y(y) \qquad \text{for all} \qquad x, y \in \mathbb{R}.$$

Corollary 6.14. Random variables $X_1, \ldots, X_n$ are independent if and only if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

Corollary 6.15. Consider maps $g_1, \ldots, g_n$ where $g_i : \mathbb{R} \to \mathbb{R}$. If random variables $X_1, \ldots, X_n$ are independent, then $g_1(X_1), \ldots, g_n(X_n)$ are independent, too.

Theorem 6.16 (Convolution Formula for Continuous Random Variables). Let X and Y be two independent, continuous random variables with PDFs f and g, respectively. Then the random variable $Z = X + Y$ has PDF

$$h(z) = \int f(z - y) g(y)\, dy = \int f(x) g(z - x)\, dx.$$

7 Covariance and Correlation

7.1 Weak Law of Large Numbers

Lemma 7.1. Let X and Y be two independent random variable whose expectations exist. Then,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Lemma 7.2. Let $X_1, \ldots, X_n$ be independent random variables whose variances exist. Then,

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

Theorem 7.3. Let $\{X_n\}_{n\geq 1}$ be a sequence of independent and identically distributed random variables with finite mean μ and finite variance σ^2. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) \to 0$$

as $n \to \infty$.

7.2 Covariance and Correlation

Definition 7.4. Let X and Y be two random variables. We defined the *covariance* $\operatorname{Cov}(X,Y)$ by

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

and the *correlation coefficient* $\rho_{X,Y}$ as

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}.$$

The random variables X and Y are called uncorrelated if $\rho_{X,Y} = 0$. The coerrelation coefficient satisfies $-1 \leq \rho_{X,Y} \leq 1$.

Theorem 7.5. The coerrelation coefficient is scale invariant, i.e. for all $\alpha > 0$,

$$\rho_{\alpha X,Y} = \rho_{X,\alpha Y} = \rho_{X,Y}.$$

Lemma 7.6. Let X and Y be two random variables. Then,

(i) $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$,

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

(iii) if X and Y are independent, then $\rho_{X,Y} = 0$, i.e. X and Y are uncorrelated.

7.3 Central Limit Theorem

Theorem 7.7 (Central Limit Theorem). Let $X_1, \ldots, X_n$ independent and identically distributed random variables with finite mean μ and finite variance σ^2. For $n \geq 1$, define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

Then for any $a \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{Z_n}(a) = \Phi(a),$$

where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$.