

# Probability Theory

Daniel Mao



# Contents

<b>1</b>	<b>Theory in General</b>	<b>1</b>
1.1	Probability Models . . . . .	1
1.2	Random Variables . . . . .	2
1.3	Cumulative Distribution Function . . . . .	2
<b>2</b>	<b>Probability Functions</b>	<b>5</b>
2.1	Probability Function of Events . . . . .	5
2.2	Probability Function of Random Variables . . . . .	7
<b>3</b>	<b>Joint Probability Distributions</b>	<b>9</b>
3.1	Joint Cumulative Distribution Functions . . . . .	9
3.2	Joint Probability Mass Functions . . . . .	9
3.3	Joint Probability Density Functions . . . . .	9
3.4	Marginal Distributions . . . . .	10
<b>4</b>	<b>Expectation</b>	<b>11</b>
4.1	Definition . . . . .	11
4.2	Properties of the Expectation Operator . . . . .	12
4.3	Variance . . . . .	12
4.4	Moment . . . . .	12
4.5	Moment Generating Function . . . . .	13
<b>5</b>	<b>Joint Expectation</b>	<b>15</b>
5.1	Joint Expectation . . . . .	15
5.2	Covariance . . . . .	15
5.3	Joint Moment . . . . .	16
5.4	Joint Moment Generating Function . . . . .	17
5.5	Theory in Higher Dimensions . . . . .	17

<b>6</b>	<b>Conditional Probability Distributions</b>	<b>19</b>
6.1	Conditional Probability of Events . . . . .	19
6.2	Conditional Distribution . . . . .	20
6.3	Conditional Expectations . . . . .	21
<b>7</b>	<b>Independence</b>	<b>23</b>
7.1	Independent Events . . . . .	23
7.2	Independent Random Variables . . . . .	24
<b>8</b>	<b>Discrete Random Variables</b>	<b>27</b>
8.1	Discrete Uniform Distribution . . . . .	27
8.2	Bernoulli Distribution . . . . .	27
8.3	Binomial Distribution . . . . .	28
8.4	Negative Binomial Distribution . . . . .	29
8.5	Geometric Distribution . . . . .	30
8.6	Hypergeometric Distribution . . . . .	30
8.7	Poisson Distribution . . . . .	30
8.8	Multinomial Distribution . . . . .	32
8.9	Bivariate Discrete Distributions . . . . .	33
<b>9</b>	<b>Continuous Random Variables</b>	<b>35</b>
9.1	Continuous Uniform Distribution . . . . .	35
9.2	Beta Distribution . . . . .	35
9.3	Exponential Distribution . . . . .	35
9.4	Erlang Distribution . . . . .	36
9.5	Gamma Distribution . . . . .	36
9.6	Normal Distribution . . . . .	38
9.7	Bivariate Normal Distribution . . . . .	39
9.8	Weibull Distribution . . . . .	39
9.9	Chi-squared Distribution . . . . .	39
9.10	t Distribution . . . . .	40
9.11	Properties . . . . .	40
<b>10</b>	<b>Unclassified</b>	<b>41</b>

# 1

## Theory in General

### 1.1 Probability Models

Random Experiment, two criteria

- outcome is random. i.e., the process can have multiple different outcomes, and before observing we don't know which one of them will happen.
- the random experiment must be theoretically repeatable.

**Definition** (Random Experiment). *A phenomenon or process that is repeatable, at least in theory.*

**Definition.** *A single repetition of the experiment as a trial.*

Two types:

- collecting raw data.
- summarizing raw data

**Definition** (Sample Space). *For a random experiment in which all possible outcomes are known, The set of all distinct outcomes for a random experiment, with the property that in a single trial, exactly one of these outcomes occurs, is called the **sample space**, denoted by  $\Omega$ .*

**Definition** (Event). *We define an **event**, denoted by  $A$ , to be a subset of the sample space.*

**Definition** (Probability Model). *A **probability model** consists of 3 essential components, a sample space, a collection of event, and a probability function.*

Probability Model: describes a random experiment.

## 1.2 Random Variables

**Definition** (Random Variables). *Let  $S$  be a sample space. We define a **random variable**, denoted by  $X$ , to be a function from  $S$  to  $\mathbb{R}$  such that  $\forall x \in \mathbb{R}$ , the set  $\{s \in S : X(s) \leq x\}$  is a valid event.*

## 1.3 Cumulative Distribution Function

**Definition** (Cumulative Distribution Function). *Let  $X$  be a random variable. We define the **cumulative distribution function** of  $X$ , denoted by  $F$ , to be a function from  $\mathbb{R}$  to  $\mathbb{R}$  given by*

$$F(x) = P(X \leq x).$$

**Definition** (Joint Cumulative Distribution Function). *Let  $S$  be a sample space. Let  $X_1, \dots, X_n$  be random variables on  $S$ . We define the **joint cumulative distribution function** of  $X_1, \dots, X_n$ , denoted by  $F(x_1, \dots, x_n)$ , to be a function given by*

$$F(x_1, \dots, x_n) := P(X_1 \leq x_1, \dots, X_n \leq x_n) = P\left(\bigcap_{i=1}^n \{X_i \leq x_i\}\right),$$

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

**Proposition 1.3.1.** *Properties of cumulative distribution function. Say  $F$  takes  $n$  variables  $x_1, \dots, x_n$ .*

(1) *Non-decreasing.*

*$F$  is non-decreasing in each of its variables. i.e.,  $\forall i \in \{1, \dots, n\}$ , we have*

$$x_i \leq x'_i \implies F(x_1, \dots, x_i, \dots, x_n) \leq F(x_1, \dots, x'_i, \dots, x_n).$$

(2)  *$\forall i \in \{1, \dots, n\}$ , we have*

$$\lim_{x_i \rightarrow -\infty} F(x_1, \dots, x_i, \dots, x_n) = 0.$$

(3)  *$\forall i \in \{1, \dots, n\}$ , we have*

$$\lim_{x_i \rightarrow +\infty}$$

(4) *Right Continuity.*

$$\forall a \in \mathbb{R}, \quad \lim_{x \rightarrow a^+} F(x) = F(a).$$

(5)

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

(6)

$$\forall a \in \mathbb{R}, \quad P(X < a) = \lim_{x \rightarrow a^+} F(x) - \lim_{x \rightarrow a^-} F(x).$$

(7)

$$\forall z \in \mathbb{R}, \quad P(X = a) = \text{jump at } a.$$

*Proof.***Proof of (1).**

Since  $x_1 \leq x_2$ ,  $\{X \leq x_1\} \subseteq \{X \leq x_2\}$ .

Since  $\{X \leq x_1\} \subseteq \{X \leq x_2\}$ ,  $P(X \leq x_1) \leq P(X \leq x_2)$ .

That is,  $F(x_1) \leq F(x_2)$ .

**Proof of (2).**

$$x \rightarrow +\infty \implies \{X \leq x\} \rightarrow S.$$

$$x \rightarrow -\infty \implies \{X \leq x\} \rightarrow \emptyset.$$

■





## 2

# Probability Functions

## 2.1 Probability Function of Events

**Definition** (Probability Function). *Let  $\Omega$  be a sample space. We define a **probability function**, denoted by  $P$ , to be a function from  $\Omega$  to  $\mathbb{R}$  that satisfies all of the following conditions:*

(1) *Non-negativity.*

$$P(A) \geq 0 \text{ for any } A.$$

(2)  $P(\Omega) = 1$ .

(3) *Countable Additivity.*

*Let  $\{A_i\}_{i \in \mathbb{N}}$  be a countable collection of events. Then if the  $A_i$ 's are mutually exclusive, we have*

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i).$$

**Proposition 2.1.1** (Properties of Probability Functions). *Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Then*

(1)  $P(\emptyset) = 0$ .

(2)  $A \subseteq B \implies P(A) \leq P(B)$ .

(3)  $P(A) \in [0, 1]$  for any event  $A$ .

*Proof.*

**Proof of (1):**

By the countable additivity, we have

$$P(\emptyset) = P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset).$$

Hence

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

**Proof of (2).**

$$P(B) = P(B \setminus A) + P(A).$$

So

$$P(B) - P(A) = P(B \setminus A) \geq 0.$$

**Proof of (3).**

$$P(A) \leq P(S) = 1.$$

■

**Proposition 2.1.2** (Set Operations). *Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Then*

(1)

$$\forall A, B \in \Omega, \quad P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

(2)

$$\forall A, B \in \Omega, \quad P(A \cap \overline{B}) = P(A) - P(A \cap B).$$

(3)

$$\forall A, B \in \Omega, \quad P(\overline{A}) = 1 - P(A).$$

*Proof of (3).* Note that

$$P(\overline{A}) + P(A) = P(\overline{A} \cup A) = P(\Omega) = 1.$$

So

$$P(\overline{A}) = 1 - P(A).$$

■

**Remark.**  $P(A) = 0$  does not imply  $A = \emptyset$  in general.

## 2.2 Probability Function of Random Variables

### 2.2.1 Probability Mass Functions

**Definition** (Probability Mass Function). *Let  $X$  be a discrete random variable. We define the **probability mass function**  $f$  of  $X$  to be a function from  $\mathbb{R}$  to  $[0, 1]$  given by*

$$f(x) := \begin{cases} P(X = x), & x \in \text{range}(X) \\ 0, & \text{otherwise} \end{cases}.$$

**Proposition 2.2.1.** *Let  $X$  be a discrete random variable. Let  $f$  be the probability mass function of  $X$ . Let  $\mathcal{S}$  be the support of  $f$ .*

$$\sum_{x \in \mathcal{S}} f(x) = 1.$$

### 2.2.2 Probability Density Functions

**Definition** (Probability Density Function). *Let  $X$  be a continuous random variable. We define the **probability density function** of  $X$  to be a function from  $\mathbb{R}$  to  $\mathbb{R}$  given by*

$$f(x) = \begin{cases} F'(x), & \text{if } F(x) \text{ is differentiable at } x \\ 0, & \text{otherwise} \end{cases}.$$

**Definition** (Support Set). *Let  $X$  be a continuous random variable. We define the **support set** of  $X$ , denoted by  $A$ , to be a subset of the reals given by*

$$A := \{x \in \mathbb{R} : f(x) > 0\}$$

where  $f$  is the probability density function of  $X$ .

**Proposition 2.2.2.** *The probability density of a singleton set is 0.*

**Proposition 2.2.3.**  $\forall x \in \mathbb{R}, f(x) \geq 0$ .

**Proposition 2.2.4.**

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



# 3

## Joint Probability Distributions

### 3.1 Joint Cumulative Distribution Functions

**Definition** (Joint Cumulative Distribution Function). *Let  $X$  and  $Y$  be random variables. We define the **joint cumulative distribution function**  $F$  of  $X$  and  $Y$  to be a function from  $\mathbb{R}^2$  to  $[0, 1]$  given by*

$$F(x, y) := P(X \leq x, Y \leq y).$$

### 3.2 Joint Probability Mass Functions

**Definition** (Joint Probability Mass Function). *Let  $X$  and  $Y$  be two discrete random variables. We define the **joint probability mass function**  $f$  of  $X$  and  $Y$  to be a function from  $\text{range}(X) \times \text{range}(Y)$  to  $[0, 1]$  given by*

$$f(x, y) := P(X = x, Y = y).$$

**Proposition 3.2.1.** *Let  $S$  be a sample space. Let  $X_1, \dots, X_n$  be random variables on  $S$ . Let  $f$  be the joint probability mass function of  $X_1, \dots, X_n$ . Let  $f_i$  be the marginal probability mass function of  $X_i$ , for some  $i \in \{1, \dots, n\}$ . Then*

$$f_i(x) = \sum_{X_i=x} f(X_1, \dots, X_n).$$

### 3.3 Joint Probability Density Functions

**Definition** (Joint Probability Density Functions). *Let  $X$  and  $Y$  be continuous random variables. Let  $F$  be the joint cumulative distribution function of  $X$  and  $Y$ . We define the*

**joint probability density function**  $f$  of  $X$  and  $Y$  to be a function from  $\text{range}(X) \times \text{range}(Y)$  to  $[0, 1]$  given by

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

### 3.4 Marginal Distributions

**Definition** (Marginal Cumulative Distribution Function). *Let  $S$  be a sample space. Let  $X_1, \dots, X_n$  be random variables on  $S$ . Let  $F$  be the joint cumulative distribution function of  $X_1, \dots, X_n$ . We define the **marginal cumulative distribution function** of  $X_i$ , for some  $i \in \{1, \dots, n\}$ , denoted by  $F_{X_i}$ , to be a function given by*

$$F_{X_i}(x) := \lim_{X_j \rightarrow \infty, j \neq i} F(X_1, \dots, X_n) = P(X_i \leq x).$$

# 4

## Expectation

### 4.1 Definition

**Definition** (Expectation of a Discrete Random Variable). *Let  $X$  be discrete random variable. Let  $f$  be the probability mass function of  $X$ . Let  $A$  be the support of  $f$ . Let  $g$  be a real-valued function on  $X$ . We define the **expectation** of  $g(X)$ , denoted by  $\mathbb{E}[g(X)]$ , to be a number given by*

$$\mathbb{E}[g(X)] := \sum_{x \in A} g(x)f(x),$$

*if the absolute summation  $\sum_{x \in A} |g(x)f(x)|$  converges; and we say that the expectation of  $g(X)$  does not exist otherwise.*

**Definition** (Expectation of a Continuous Random Variable). *Let  $X$  be continuous random variable. Let  $f$  be the probability density function of  $X$ . Let  $A$  be the support of  $f$ . Let  $g$  be a real-valued function on  $X$ . We define the **expectation** of  $g(X)$ , denoted by  $\mathbb{E}[g(X)]$ , to be a number given by*

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

*if the absolute integral  $\int_A |g(x)f(x)|dx$  converges; and we say that the expectation of  $g(X)$  does not exist otherwise.*

**Definition** (Expectation of a Random Vector). *Let  $X = (X_1, \dots, X_n)$  be a random vector. We define the **expectation** of  $X$  to be a vector given by*

$$\mathbb{E}[X] := \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}.$$

## 4.2 Properties of the Expectation Operator

**Proposition 4.2.1** (Linearity). *Expectation is a linear operator. i.e., Let  $X = (X_1, \dots, X_n)$  be a random vector. Let  $\vec{\lambda} = (\lambda_1, \dots, \lambda_n)$  be a constant. Then*

$$\mathbb{E}\left[\sum_{i=1}^n \lambda_i X_i\right] = \sum_{i=1}^n \lambda_i \mathbb{E}[X_i].$$

Or,

$$\mathbb{E}[\vec{\lambda}X] = \vec{\lambda} \cdot \mathbb{E}[X].$$

**Proposition 4.2.2.** *Let  $X$  be a random vector. Let  $g_1, \dots, g_n$  be real-valued functions on  $X$ . Let  $\lambda_1, \dots, \lambda_n$  be constants. Then*

$$\mathbb{E}\left[\sum_{i=1}^n \lambda_i g_i(X)\right] = \sum_{i=1}^n \lambda_i \mathbb{E}[g_i(X)].$$

## 4.3 Variance

**Definition** (Variance). *Let  $X$  be a random variable. We define the **variance** of  $X$ , denoted by  $\text{var}[X]$ , to be the number given by*

$$\text{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2],$$

or equivalently,

$$\text{var}(X) = \text{cov}(X, X).$$

**Proposition 4.3.1.**

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

**Proposition 4.3.2.**

$$\text{var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2.$$

## 4.4 Moment

**Definition** (Moment). *Let  $X$  be a random variable. Let  $n$  be a natural number. We define the  $k^{\text{th}}$  **moment** of  $X$  to be the number given by*

$$\mathbb{E}[X^k].$$

**Definition** (Central Moment). *We define the  $k^{\text{th}}$  **central moment** of  $X$  for  $k \in \mathbb{N}$  to be the number given by*

$$\mathbb{E}[(X - \mathbb{E}[X])^k].$$



**Remark.** *The first moment is the mean.*

**Proposition 4.4.1.**

$$\text{var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

*provided that  $\mathbb{E}[X^2]$  exists.*

*Proof.*

$$\begin{aligned} \text{var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

■

## 4.5 Moment Generating Function

**Proposition 4.5.1.**

$$M(0) = 1.$$

**Proposition 4.5.2** (Expansion of the Moment Generating Function). *Let  $X$  be a random variable. Let  $\Phi_X$  be the moment generating function of  $X$ . Then*

$$\Phi_X(t) = \sum_{i=0}^{\infty} \mathbb{E}[X^i] \frac{t^i}{i!}.$$

*Proof.*

$$\begin{aligned} \Phi_X(t) &= \mathbb{E}[e^{tX}] = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\right] \\ &= \sum_{i=0}^{\infty} \mathbb{E}\left[\frac{(tX)^i}{i!}\right] = \sum_{i=0}^{\infty} \mathbb{E}[X^i] \frac{t^i}{i!}. \end{aligned}$$

That is,

$$\Phi_X(t) = \sum_{i=0}^{\infty} \mathbb{E}[X^i] \frac{t^i}{i!}.$$

The  $i^{\text{th}}$  moment of the random variable  $X$  is the coefficient of the term  $\frac{t^i}{i!}$ . ■

**Proposition 4.5.3.** *Let  $X$  be a random variable. Let  $\Phi_X$  be the moment generating function of  $X$ . Given the moment generating function of  $X$ , we can extract its  $n^{\text{th}}$  moment, for  $n \in \mathbb{N}$ , via*

$$\Phi_X^{(n)}(0) = \mathbb{E}[X^n].$$

**Proposition 4.5.4** (Linear Transformations). *Let  $X$  be a random variable. Let  $M_X$  be the moment generating function for  $X$  on  $(-h, h)$  for some  $h > 0$ . Let  $\alpha, \beta \in \mathbb{R}$  and  $\alpha \neq 0$ . Then the moment generating function  $M_{\alpha X + \beta}$  for the random variable  $\alpha X + \beta$  is*

$$M_{\alpha X + \beta}(t) = e^{\beta t} M_X(\alpha t),$$

*defined on  $(-\frac{h}{|\alpha|}, \frac{h}{|\alpha|})$ .*

**Proposition 4.5.5** (Uniqueness Property). *Let  $X$  and  $Y$  be random variables. Let  $M_X$  be the moment generating function for  $X$ . Let  $F_X$  be the cumulative distribution function of  $X$ . Let  $M_Y$  be the moment generating function for  $Y$ . Let  $F_Y$  be the cumulative distribution function of  $Y$ . Then  $M_X = M_Y$  if and only if  $F_X = F_Y$ .*

# 5

## Joint Expectation

### 5.1 Joint Expectation

**Definition** (Joint Expectation of Discrete Random Variables). *Let  $X$  be a discrete random vector. Let  $f$  be the joint probability mass function of  $X$ . Let  $A$  be the support of  $f$ . Let  $g$  be a real-valued function on  $X$ . We define the **joint expectation** of  $g(X)$ , denoted by  $\mathbb{E}[g(X)]$ , to be a number given by*

$$\mathbb{E}[g(X)] = \sum_{\vec{x} \in A} g(\vec{x})f(\vec{x}),$$

*if  $\sum_{\vec{x} \in A} |g(\vec{x})f(\vec{x})| < +\infty$ ; and we say that the expectation of  $g(X)$  does not exist otherwise.*

**Definition** (Joint Expectation of Continuous Random Variables). *Let  $X$  be a continuous random vector. Let  $f$  be the joint probability density function of  $X$ . Let  $A$  be the support of  $f$ . Let  $g$  be a function on  $X$ . We define the **joint expectation** of  $g(X)$ , denoted by  $\mathbb{E}[g(X)]$ , to be a number given by*

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

*if  $\int_A |g(x)f(x)|dx < +\infty$ ; and we say that the expectation of  $g(X)$  does not exist otherwise.*

### 5.2 Covariance

**Definition** (Covariance). *Let  $X$  and  $Y$  be random variables. We define the **covariance** of  $X$  and  $Y$ , denoted by  $\text{cov}(X, Y)$ , to be the number given by*

$$\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

**Definition** (Uncorrelated). *Let  $X$  and  $Y$  be two random variables. We say that  $X$  and  $Y$  are **uncorrelated** if  $\text{cov}(X, Y) = 0$ .*

**Proposition 5.2.1.** *If  $X$  and  $Y$  are independent, then  $\text{cov}(X, Y) = 0$ . i.e. independent random variables are uncorrelated.*

**Proposition 5.2.2.** *Let  $X$  and  $Y$  be two random variables. Then*

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y].$$

*Proof.*

$$\begin{aligned} \text{cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X] \mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] - \mathbb{E}[Y] \mathbb{E}[X] + \mathbb{E}[X] \mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]. \end{aligned}$$

That is,

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y].$$

■

**Proposition 5.2.3** (Bilinearity of the Covariance Operator). *Let  $X = (X_1, \dots, X_n)$  be a random vector. Let  $Y := \vec{a}X = \sum_{i=1}^n a_i X_i$  and  $Z := \vec{b}X = \sum_{i=1}^n b_i X_i$  where  $\vec{a}$  and  $\vec{b}$  are constant vectors. Then*

$$\text{cov}\left(\sum_{i=1}^n a_i X_i, \sum_{i=1}^n b_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i b_j \text{cov}(X_i, X_j).$$

Or,

$$\text{cov}(Y, Z) = \vec{a}^T \text{var}(Y, Z) \vec{b}.$$

### 5.3 Joint Moment

**Definition** (Joint Moment). *Let  $X$  and  $Y$  be random variables. Let  $m$  and  $n$  be natural numbers. We define the  $(m, n)^{\text{th}}$  **joint moment** of  $X$  and  $Y$  to be a number given by*

$$\mathbb{E}[X^m Y^n] = \Phi^{(m, n)} = \frac{\partial^{m+n}}{\partial s^m \partial t^n} \Phi(s, t) \big|_{s=0, t=0}.$$

## 5.4 Joint Moment Generating Function

**Definition** (Joint Moment Generating Function). Let  $X_1, \dots, X_n$  be random variables. We define the **joint moment generating function** of  $X_1, \dots, X_n$ , denoted by  $\Phi$ , to be a function from  $\mathbb{R}^n$  to  $\mathbb{R}$  given by

$$\Phi(t_1, \dots, t_n) := \mathbb{E} \left[ \exp \left\{ \sum_{i=1}^n t_i X_i \right\} \right],$$

if  $\exists h_1, \dots, h_n > 0$  such that the RHS is defined on  $(-h_1, h_1) \times \dots \times (-h_n, h_n)$ . The domain of  $\Phi$  is the set of all tuples  $(t_1, \dots, t_n)$  such that the RHS is defined.

## 5.5 Theory in Higher Dimensions

**Definition** (Variance of a Random Vector). Let  $X = (X_1, \dots, X_n)$  be a random vector. We define the **variance** of  $X$  to be a matrix given by

$$\text{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T].$$

**Proposition 5.5.1.**

$$\begin{aligned} \text{var}(X) &= \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \dots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \dots & \text{cov}(X_n, X_n) \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \dots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \dots & \text{var}(X_n) \end{bmatrix}. \end{aligned}$$

**Proposition 5.5.2.** Covariance matrices are symmetric.

*Proof.*  $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$ . ■

**Proposition 5.5.3.** Let  $X$  be a random vector. Then  $\text{var}(X)$  is positive definite. i.e.,  $\forall a \in \mathbb{R}^n : a^T \text{var}(X) a \geq 0$ .



## 6

# Conditional Probability Distributions

### 6.1 Conditional Probability of Events

**Definition** (Conditional Probability). Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Let  $A$  and  $B$  be two events in the sample space. We define the **conditional probability** of event  $A$  given event  $B$  occurs, denoted by  $P(A | B)$ , to be the number given by

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

provided that  $P(B) \neq 0$ .

**Proposition 6.1.1** (Multiplication Rule). Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Then

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

provided that  $P(B) \neq 0$ .

Let  $\{A_i\}_{i=1}^{i=n}$  be a sequence of events. Then

$$P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i | \bigcap_{j=0}^{j=i-1} A_j)$$

where  $A_0$  is defined to be  $\Omega$ .

*Proof.* Since  $P(A | B)$  is defined to be  $\frac{P(A \cap B)}{P(B)}$ , we get

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

■

**Proposition 6.1.2** (Law of Total Probability). *Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Let  $A$  be an event in  $\Omega$ . Let  $\{B_i\}_{i \in \mathbb{N}}$  be a countable collection of events in  $\Omega$ . Suppose that  $\bigcup_{i \in \mathbb{N}} B_i = \Omega$  and that  $\forall i, j \in \mathbb{N}$ , we have  $B_i \cap B_j = \emptyset$ . Then*

$$P(A) = \sum_{i \in \mathbb{N}} P(A \mid B_i)P(B_i).$$

*Proof.*

$$\begin{aligned} P(A) &= P(A \cap \Omega) \\ &= P(A \cap \bigcup_{i \in \mathbb{N}} B_i) \\ &= P(\bigcup_{i \in \mathbb{N}} A \cap B_i), \text{ by the distributivity property} \\ &= \sum_{i \in \mathbb{N}} P(A \cap B_i), \text{ since mutually exclusive} \\ &= \sum_{i \in \mathbb{N}} P(A \mid B_i)P(B_i). \text{ by the multiplication rule} \end{aligned}$$

That is,

$$P(A) = \sum_{i \in \mathbb{N}} P(A \mid B_i)P(B_i).$$

Think of this as distributing the event  $A$  over all  $B_i$ 's. Then the probability  $P(A)$  is a weighted sum of the conditional probabilities of event  $A$  where the weights are the corresponding probabilities of the given events  $B_i$ . ■

**Proposition 6.1.3** (Bayes' Formula).

$$\forall j \in \mathbb{N}, \quad P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{i \in \mathbb{N}} P(A \mid B_i)P(B_i)}.$$

*Proof.*

$$P(B_j \mid A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j \cap A)}{\sum_{i \in \mathbb{N}} P(A \mid B_i)P(B_i)}.$$

■

## 6.2 Conditional Distribution

**Definition** (Conditional Probability Mass/Density Function). *Let  $X$  and  $Y$  be discrete/continuous random variables. Let  $f$  be the joint probability mass/density function of  $X$  and  $Y$ . We define the **conditional probability mass/density function** of  $X$  given  $Y = y$ , denoted by  $f_X(\cdot \mid y)$ , to be a function given by*

$$f_X(x \mid y) = \frac{f(x, y)}{f_Y(y)} = \frac{P(X = x, Y = y)}{P(Y = y)} = P(X = x \mid Y = y)$$



where  $f_Y$  is the marginal probability mass/density function of  $Y$ , provided that  $f_Y(y) \neq 0$ .

**Proposition 6.2.1.** *Let  $X$  and  $Y$  be discrete/continuous random variables. Let  $f_X$  and  $f_Y$  be the marginal probability mass/density functions of  $X$  and  $Y$ , respectively. Let  $f_X(\cdot | y)$  and  $f_Y(\cdot | x)$  be the conditional probability mass/density functions of  $X$  and  $Y$ , respectively. Let  $A_X$  and  $A_Y$  be the marginal support of  $X$  and  $Y$ , respectively. Then  $X$  and  $Y$  are independent if and only if*

$$f_X(\cdot | y) = f_X \text{ and } f_Y(\cdot | x) = f_Y.$$

*Proof.*  $X$  and  $Y$  are independent if and only if  $f(x, y) = f_X(x)f_Y(y)$ . ■

## 6.3 Conditional Expectations

**Definition** (Conditional Expectation). *Let  $X$  and  $Y$  be random variables. Let  $g$  be a function on  $X$ . We define the **conditional expectation** of  $g(X)$  given  $Y = y$  to be a number given by*

$$E[g(X) | Y = y] = \begin{cases} \sum_{\text{all } x} g(x)f_X(x | y), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{+\infty} g(x)f_X(x | y)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

if  $\sum_{\text{all } x} |g(x)f_X(x | y)| \neq +\infty$  or  $\int_{-\infty}^{+\infty} |g(x)f_X(x | y)|dx \neq +\infty$ .

**Definition** (Conditional Mean). *Let  $X$  and  $Y$  be random variables. Let  $g$  be a function on  $X$ . We define the **conditional mean** of  $X$  given  $Y = y$  to be the number  $E[X | Y = y]$ .*

**Definition** (Conditional Variance). *Let  $X$  and  $Y$  be random variables. Let  $g$  be a function on  $X$ . We define the **conditional variance** of  $X$  given  $Y = y$ , denoted by  $\text{Var}[X | Y = y]$ , to be the number given by*

$$\mathbb{E}[(X - \mathbb{E}[X | Y = y])^2 | Y = y].$$

**Proposition 6.3.1** (Substitution Rule).

$$E[h(X, Y) | Y = y] = E[h(X, y) | Y = y].$$

**Theorem 1** (Law of Total Expectation).

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

*Proof.*

$$\begin{aligned}
& E [E[g(X) | Y]] \\
&= E \left[ \int_{-\infty}^{+\infty} g(x) f_X(x | Y) dx \right] \\
&= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} g(x) f_X(x | y) dx \right] f_Y(y) dy \\
&= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} g(x) f_X(x | y) f_Y(y) dx \right] dy \\
&= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} g(x) f(x, y) dx \right] dy \\
&= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} g(x) f(x, y) dy \right] dx \\
&= \int_{-\infty}^{+\infty} g(x) \left[ \int_{-\infty}^{+\infty} f(x, y) dy \right] dx \\
&= \int_{-\infty}^{+\infty} g(x) f_X(x) dx \\
&= E[g(X)].
\end{aligned}$$

■

**Proposition 6.3.2** (Law of Total Variance).

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

# 7

## Independence

### 7.1 Independent Events

#### 7.1.1 Definitions

**Definition** (Independent Events). Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Let  $A$  and  $B$  be two events in  $\Omega$ . We say that  $A$  and  $B$  are **independent** if  $P(A \cap B) = P(A)P(B)$ .

**Definition** (Independent Events). Let  $A$  and  $B$  be two events with positive probabilities. We say that  $A$  and  $B$  are **independent** if both  $P(A | B) = P(A)$  and  $P(B | A) = P(B)$ .

**Proposition 7.1.1.** The two definitions of independence are equivalent.

*Proof.*

For one direction, assume that  $P(A \cap B) = P(A)P(B)$ .

Since  $P(A \cap B) = P(A)P(B)$  and  $P(B)P(A | B) = P(A \cap B)$ ,  $P(A)P(B) = P(A | B)P(B)$ .

Since  $P(B) \neq 0$  and  $P(A)P(B) = P(A | B)P(B)$ ,  $P(A | B) = P(A)$ .

Since  $P(A \cap B) = P(A)P(B)$  and  $P(A)P(B | A) = P(A \cap B)$ ,  $P(A)P(B) = P(B | A)P(A)$ .

Since  $P(A) \neq 0$  and  $P(A)P(B) = P(B | A)P(A)$ ,  $P(B | A) = P(B)$ .

For the reverse direction, assume that  $P(A | B) = P(A)$  and  $P(B | A) = P(B)$ .

Since  $P(A | B) = \frac{P(A \cap B)}{P(B)}$  and  $P(A | B) = P(A)$ ,  $P(A)P(B) = P(A \cap B)$ .

■

**Definition** (Pairwise Independent). Let  $\mathcal{A} = \{A_i\}_{i=1}^n$  be a finite collection of events where  $n \in \mathbb{N}$ . We say that the events in  $\mathcal{A}$  are **pairwise independent** if any pair of events are independent. i.e.,  $\forall i, j \in \{1, \dots, n\}$ , we have  $P(A_i \cap A_j) = P(A_i)P(A_j)$ .

**Definition** (Mutually Independent). Let  $\mathcal{A} = \{A_i\}_{i=1}^n$  be a finite collection of events where  $n \in \mathbb{N}$ . We say that the events in  $\mathcal{A}$  are **mutually independent** if any event

is independent of the intersection of any other events. i.e.,  $\forall I \subseteq \{1, \dots, n\}$ , we have  $P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ .

### 7.1.2 Properties

**Proposition 7.1.2** (Self-Independence). *An event  $A$  is independent of itself if and only if  $P(A) = 0$  or  $P(A) = 1$ .*

*Proof.*

$$P(A) = P(A \cap A) = P(A)P(A) \iff P(A) \in \{0, 1\}.$$

■

**Proposition 7.1.3.** *A zero-probability event is independent of any any other event.*

*Proof.* Let  $\Omega$  be a sample space. Let  $P$  be a probability function defined on the sample space. Let  $A$  and  $B$  be two events in  $\Omega$ . Suppose that  $P(A) = 0$ . Since  $A \cap B \subseteq A$ , we get  $P(A \cap B) \leq P(A)$ . Note that  $P(A \cap B) \geq 0$  and that  $P(A) = 0$ . So  $P(A \cap B) = 0$ . So  $P(A \cap B) = P(A)P(B)$ . So  $A$  and  $B$  are independent. ■

## 7.2 Independent Random Variables

### 7.2.1 Definitions

**Definition** (Independence 1). *Let  $X$  and  $Y$  be two random variables. We say that  $X$  and  $Y$  are **independent** if*

$$\forall A, B \subseteq \mathbb{R}, \quad P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

**Definition** (Independence 2). *Let  $X$  and  $Y$  be two random variables. Let  $f$  be the joint probability function of  $X$  and  $Y$ . Let  $f_X$  be the marginal probability function of  $X$ . Let  $f_Y$  be the marginal probability function of  $Y$ . We say that  $X$  and  $Y$  are **independent** if*

$$f = f_X f_Y.$$

i.e., if

$$\forall (x, y) \in \mathcal{S}_X \times \mathcal{S}_Y, \quad f(x, y) = f_X(x)f_Y(y).$$

where  $\mathcal{S}_X$  is the support of  $X$  and  $\mathcal{S}_Y$  is the support of  $Y$ .

**Definition** (Independence 3). *Let  $X$  and  $Y$  be two random variables. Let  $F$  be the joint cumulative distribution function of  $X$  and  $Y$ . Let  $F_X$  be the marginal cumulative distribution function of  $X$ . Let  $F_Y$  be the marginal cumulative distribution function of  $Y$ . We say that  $X$  and  $Y$  are **independent** if*

$$F = F_X F_Y.$$

**Definition** (Independence 4). Let  $X$  and  $Y$  be two random variables. Let  $M$  be the joint moment generating function of  $X$  and  $Y$ . Let  $M_X$  be the marginal moment generating function of  $X$ . Let  $M_Y$  be the marginal moment generating function of  $Y$ . We say that  $X$  and  $Y$  are **independent** if

$$M = M_X M_Y.$$

**Proposition 7.2.1.** The 4 definitions of independence are equivalent.

### 7.2.2 Properties

**Proposition 7.2.2.** Let  $X$  and  $Y$  be random variables. Let  $g$  be a function on  $X$ . Let  $h$  be a function on  $Y$ . Suppose that  $X$  and  $Y$  are independent. Then the random variables  $g(X)$  and  $h(Y)$  are also independent.

**Proposition 7.2.3.** Let  $X$  and  $Y$  be random variables. Let  $g$  be a function on  $X$ . Then if  $X$  and  $Y$  are independent, we have

$$\mathbb{E}[g(X) \mid Y = y] = \mathbb{E}[g(X)].$$

In particular,  $E[X \mid Y = y] = E[X]$  and  $\text{var}[X \mid Y = y] = \text{var}[X]$ .

**Proposition 7.2.4** (Expectation). Let  $X_1, \dots, X_n$  be independent random variables. Let  $g_i$  be a function on  $X_i$  for  $i = 1..n$ . Then

$$\mathbb{E}\left[\prod_{i=1}^n g_i(X_i)\right] = \prod_{i=1}^n \mathbb{E}[g_i(X_i)].$$

**Proposition 7.2.5** (Moment Generating Function). Let  $X_i$  for  $i = 1, \dots, n$  be independent random variables. Let  $\Phi_i$  be the marginal moment generating function of  $X_i$  for  $i = 1..n$ . Let  $a_i$  be real numbers for  $i = 1..n$ . Define a random variable  $X$  by

$$X := \sum_{i=1}^n a_i X_i = \vec{a} \cdot \vec{X}.$$

Then the moment generating function  $\Phi_X$  of  $X$  is

$$\Phi_X(t) = \prod_{i=1}^n \Phi_i(a_i t).$$

*Proof.*

$$\begin{aligned}
\Phi_X(t) &= \mathbb{E}[e^{tX}] \\
&= \mathbb{E}[\exp\{t \sum_{i=1}^n a_i X_i\}] \\
&= \mathbb{E}[\prod_{i=1}^n \exp\{ta_i X_i\}] \\
&= \prod_{i=1}^n \mathbb{E}[e^{ta_i X_i}], \text{ by independence} \\
&= \prod_{i=1}^n \Phi_i(a_i t).
\end{aligned}$$

That is,

$$\Phi_X(t) = \prod_{i=1}^n \Phi_i(a_i t).$$

■

### 7.2.3 Factorization

**Theorem 2** (Factorization Theorem of Independence). *Let  $X$  and  $Y$  be two random variables. Let  $f$  be the joint probability function of  $X$  and  $Y$ . Let  $A_X$  be the support of  $X$ . Let  $A_Y$  be the support of  $Y$ . Then  $X$  and  $Y$  are independent if and only if there exist functions  $g : A_X \rightarrow \mathbb{R}$  and  $h : A_Y \rightarrow \mathbb{R}$  such that  $f = gh$ . i.e.,  $\forall (x, y) \in A_X \times A_Y$ ,  $f(x, y) = g(x)h(y)$ .*

**Corollary.** *If  $A$  is not rectangular, then  $X$  and  $Y$  cannot be independent.*

*Proof.* If  $A$  is not rectangular, then  $\exists x \in A_X, y \in A_Y$  such that  $(x, y) \notin A$ . So  $f(x, y) = 0 < f_X(x)f_Y(y)$ . ■

## 8

# Discrete Random Variables

**Definition** (Discrete Random Variable). *Let  $X$  be a random variable. We say that  $X$  is a **discrete random variable** if the state space of  $S$  is countable.*

## 8.1 Discrete Uniform Distribution

**Definition** (Discrete Uniform Distribution).  *$X$  is equally likely to take on values in the finite set  $\{a, \dots, b\}$ , We say that  $X$  follows a **discrete uniform distribution**, denoted by  $X \sim DU(a, b)$ .*

## 8.2 Bernoulli Distribution

**Definition** (Bernoulli Distribution). *If we consider a Bernoulli trial, which is a random trial with probability  $p$  of being a “success” and probability  $1 - p$  being a “failure”, then we say that  $X$  follows **Bernoulli distribution**, denoted by  $X \sim \text{Bernoulli}(p)$ .*

**Proposition 8.2.1** (Probability Density Function of Bernoulli Distribution).

$$f(x) = \begin{cases} P(X = x), & x \in \{0, 1\} \\ 0, & \text{otherwise} \end{cases} = \begin{cases} p^x(1-p)^{1-x}, & x \in \{0, 1\} \\ 0, & \text{otherwise} \end{cases}$$

**Proposition 8.2.2** (Expectation of Bernoulli Distribution).

$$\mathbb{E}[X] = \sum_{x \in A} xf(x) = (1)(p) + (0)(1-p) = p.$$

**Example 8.2.1.** *Flipping a coin once.*

### 8.3 Binomial Distribution

**Definition** (Binomial Distribution). Let  $X_i \sim \text{Bernoulli}(p)$  for  $i \in \{1, \dots, n\}$ . Define a random variable  $X$  by  $X = \sum_{i=1}^n X_i$ . We say that the random variable  $X$  follows a **binomial distribution**, denoted by  $X \sim \text{Binomial}(n, p)$ . Then  $X$  records the number of “success” trials.

**Proposition 8.3.1** (Probability Density Function of Binomial Distribution).

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

**Proposition 8.3.2** (Moment Generating Function of Binomial Distribution). Let  $X \sim \text{Binomial}(n, p)$ . Then for  $t \in \mathbb{R}$ ,

$$\Phi_X(t) = (pe^t + (1-p))^n.$$

*Proof.* For  $t \in \mathbb{R}$ ,

$$\begin{aligned} \Phi_X(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{x=0}^n e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + (1-p))^n. \end{aligned}$$

That is, for  $t \in \mathbb{R}$ ,

$$\Phi_X(t) = (pe^t + (1-p))^n.$$

■

**Proposition 8.3.3** (Mean of Binomial Distribution). Let  $X \sim \text{Binomial}(n, p)$ . Then

$$\mathbb{E}[X] = np.$$

*Proof Approach 1.*

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = np.$$

■



*Proof Approach 2.*

$$\begin{aligned}
 \mathbb{E}[X] &= \Phi'_X(t)|_{t=0} \\
 &= \frac{d}{dt}((pe^t) + (1-p))^n|_{t=0} \\
 &= n(pe^t + 1-p)^{n-1}pe^t|_{t=0} \\
 &= np.
 \end{aligned}$$

■

**Proposition 8.3.4** (Variance of Binomial Distribution). *Let  $X \sim \text{Binomial}(n, p)$ . Then*

$$\text{var}[X] = np(1-p).$$

*Proof Approach 2.*

$$\begin{aligned}
 \Phi''_X(t)|_{t=0} &= \frac{d^2}{dt^2}((pe^t) + (1-p))^n|_{t=0} \\
 &= n(pe^t + 1-p)^{n-1}pe^t + npe^t(n-1)(pe^t + 1-p)^{n-2}pe^t|_{t=0} \\
 &= np + n(n-1)p^2.
 \end{aligned}$$

$$\begin{aligned}
 \text{var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= \Phi''_X(t)|_{t=0} - (\Phi'_X(t)|_{t=0})^2 \\
 &= np + n(n-1)p^2 - (np)^2 \\
 &= np - np^2 = np(1-p).
 \end{aligned}$$

■

## 8.4 Negative Binomial Distribution

**Definition** (Negative Binomial Distribution). *If  $X$  denotes the number of Bernoulli trials required to observe  $k \in \mathbb{N}$  successes, We say that the random variable  $X$  follows a **negative binomial distribution**, denoted by  $X \sim \text{NB}(k, p)$ .*

$X := \#$  of 0 outcomes before the  $r^{\text{th}}$  outcome of 1 in repeated Bernoulli( $p$ ) experiments

$X \sim \text{NegBin}(r, p)$ .

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

$$X = \sum_{i=1}^r X_i$$

$$X_i \sim \text{Geo}(p).$$

## 8.5 Geometric Distribution

**Definition** (Geometric Distribution).  $X$  denotes the number of Bernoulli trials required to observe the first success. i.e.,  $X \sim NB(1, p)$ . We say that the random variable  $X$  follows a **geometric distribution**, denoted by  $X \sim \text{Geo}(p)$ .

## 8.6 Hypergeometric Distribution

**Definition** (Hypergeometric Distribution).  $X$  denotes the number of success objects in  $n$  draws without replacement from a finite population of size  $N$  containing exactly  $r$  success objects. We say that  $X$  follows a **hypergeometric distribution**, denoted by  $X \sim HG(N, r, n)$ .

**Proposition 8.6.1** (Probability Function of Hypergeometric Distribution). For  $x = \max\{0, n - N + r\}, \dots, \min\{n, r\}$ ,

$$p(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}}.$$

## 8.7 Poisson Distribution

**Definition** (Poisson Distribution). Let  $X \sim \text{Poisson}(\lambda)$  for  $\lambda \in \mathbb{R}_{++}$ . Then the probability mass function of  $X$  is

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

with support  $k \in \mathbb{N}_0$ .

**Remark.** Note that if we force  $\lambda$  to be equal to 0, we get

$$p(x) = \frac{e^{-0} 0^x}{x!} = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise.} \end{cases}$$

**Proposition 8.7.1** (Moment Generating Function). The moment generating function of a  $\text{Poisson}(\lambda)$  distributed random variable is

$$M(t) = e^{\lambda(e^t - 1)} \text{ for } t \in \mathbb{R}.$$

*Proof.*

$$\begin{aligned}
 M(t) &= \mathbb{E}[e^{tX}] \\
 &= \sum_{x=0}^{\infty} e^{tx} f(x) \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x e^{tx}}{x!} \\
 &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\
 &= e^{\lambda(e^t - 1)},
 \end{aligned}$$

for any  $t \in \mathbb{R}$ . ■

**Proposition 8.7.2** (Mean and Variance). *The mean and variance of a  $\text{Poisson}(\lambda)$  distributed random variable are*

$$\begin{cases} \mathbb{E}[X] = \lambda \text{ and} \\ \text{var}[X] = \lambda. \end{cases}$$

*Proof.*

$$\begin{aligned}
 \mathbb{E}[X] &= M'(0) = \lambda. \\
 \text{var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
 &= M''(0) - (M'(0))^2 \\
 &= (\lambda^2 + \lambda) - \lambda^2 = \lambda.
 \end{aligned}$$
■

**Proposition 8.7.3.** *When  $n$  is large and  $p$  is small,  $\text{Poisson}(np)$  can be used to approximate  $\text{Binomial}(n, p)$ .*

*Proof.*

$$\begin{aligned}
 \lim_{n \rightarrow \infty} P(X = x) &= \lim_{n \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} \\
 &= \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
 &= \lim_{n \rightarrow \infty} \frac{n}{n} \frac{n-1}{n} \dots \frac{n-x+1}{n} \frac{\lambda^x}{x!} \frac{(1 - \frac{\lambda}{n})^n}{(1 - \frac{\lambda}{n})^x} \\
 &= 1 \cdot \dots \cdot 1 \cdot \frac{\lambda^x}{x!} \cdot \frac{e^{-\lambda}}{1} \\
 &= \frac{e^{-\lambda} \lambda^x}{x!}.
 \end{aligned}$$
■

## 8.8 Multinomial Distribution

Let  $X_1, \dots, X_k$  be random variables. Let  $p_1, \dots, p_k$  be probabilities such that  $\sum_{i=1}^k p_i = 1$ . Let  $n$  be the number of trials.

$$(X_1, \dots, X_n) \sim \text{Multinomial}(n, p_1, \dots, p_k).$$

**Proposition 8.8.1** (Joint Probability Mass Function).

$$f(x_1, \dots, x_k) = \begin{cases} \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}, & \text{if } x_i = 0, 1, \dots \text{ and } \sum_{i=1}^k x_i = n \\ 0, & \text{otherwise.} \end{cases}$$

**Proposition 8.8.2** (Joint Moment Generating Function).

$$M(t_1, \dots, t_n) = \mathbb{E}[\exp\{\sum_{i=1}^k t_i X_i\}] = (\sum_{i=1}^k p_i e^{t_i})^n$$

for any  $(t_1, \dots, t_k) \in \mathbb{R}^k$ , where  $\mathbb{E}$  denotes the expectation operator and  $\exp$  denotes the exponential function.

**Proposition 8.8.3** (Marginal Distribution). •  $X_i \sim \text{Binomial}(n, p_i)$ .

- $\mathbb{E}[X_i] = np_i$ .
- $\text{var}[X_i] = np_i(1 - p_i)$ .
- 

$$\begin{aligned} M_{X_i}(t_i) &= M(0, \dots, 0, t_i, 0, \dots, 0) \\ &= (p_i e^{t_i} + \sum_{j \neq i} p_j)^n \\ &= (p_i e^{t_i} + (1 - p_i))^n. \end{aligned}$$

**Proposition 8.8.4** (Conditional Distribution). •

$$X_i \mid X_j = x_j \sim \text{Binomial}\left(n - x_j, \frac{p_i}{1 - p_j}\right)$$

for  $i \neq j$ .

$$X_i \mid X_i + X_j = t \sim \text{Binomial}\left(t, \frac{p_i}{p_i + p_j}\right).$$

**Proposition 8.8.5.** Let  $T := X_i + X_j$ . Then  $T \sim \text{Binomial}(n, p_i + p_j)$ .

*Proof.* Idea: use MGF. ■

**Proposition 8.8.6.**  $\text{cov}(X_i, X_j) = -np_i p_j$ .

*Proof.*

$$\begin{aligned}
& \text{cov}(X_i, X_j) \\
&= \frac{1}{2} [2 \text{cov}(X_i, X_j)] \\
&= \frac{1}{2} [\text{cov}(X_i, X_i) + \text{cov}(X_i, X_j) + \text{cov}(X_j, X_i) + \text{cov}(X_j, X_j) - \text{cov}(X_i, X_i) - \text{cov}(X_j, X_j)] \\
&= \frac{1}{2} [\text{cov}(X_i + X_j, X_i + X_j) - \text{cov}(X_i, X_i) - \text{cov}(X_j, X_j)] \\
&= \frac{1}{2} [\text{var}(X_i + X_j) - \text{var}(X_i) - \text{var}(X_j)] \\
&= \frac{1}{2} [n(p_i + p_j)(1 - p_i - p_j) - np_i(1 - p_i) - np_j(1 - p_j)] \\
&= \frac{1}{2} [-2np_i p_j] \\
&= -np_i p_j.
\end{aligned}$$

■

## 8.9 Bivariate Discrete Distributions

**Definition** (Bivariate Discrete Random Variables). *Let  $S$  be a sample space. We define a pair of **bivariate discrete random variables** on  $S$ , to be a pair  $(X, Y)$  of random variables on  $S$  such that there exists some subset  $A$  of  $\mathbb{R}^2$  such that  $P((X, Y) \in A) = 1$ .*

**Definition** (Joint Support). *Let  $S$  be a sample space. Let  $(X, Y)$  be a pair of bivariate discrete random variables. We define the **joint support** of  $(X, Y)$ , denoted by  $A$ , to be a set given by*

$$A := \{(x, y) \in \mathbb{R}^2 : f(x, y) > 0\}.$$



## 9

# Continuous Random Variables

**Definition** (Continuous Random Variable). *Let  $F$  be the cumulative distribution function of  $X$ .*

(1)  *$F$  is continuous on  $\mathbb{R}$ .*

(2)  *$F$  is differentiable almost everywhere on  $\mathbb{R}$ .*

## 9.1 Continuous Uniform Distribution

## 9.2 Beta Distribution

## 9.3 Exponential Distribution

**Definition** (Exponential Distribution). *Let  $X \sim \text{Exponential}(\lambda)$ . Then  $X$  has probability density function*

$$f(x) = \lambda e^{-\lambda x}$$

*with support  $x \in \mathbb{R}_+$ .*

**Proposition 9.3.1** (Mean and Variance). *Then mean and variance of a  $\text{Exponential}(\lambda)$  distributed random variable are*

$$\begin{cases} \mathbb{E}[X] = \frac{1}{\lambda} \text{ and} \\ \text{var}[X] = \frac{1}{\lambda^2}. \end{cases}$$

## 9.4 Erlang Distribution

**Proposition 9.4.1** (Probability Density Function). *For  $x > 0$ ,*

$$f(x) = \frac{\lambda^n x^{n-1} e^{-\lambda x}}{(n-1)!}.$$

**Proposition 9.4.2.**  *$\text{Erlang}(1, \lambda) = \text{Exponential}(\lambda)$ .*

## 9.5 Gamma Distribution

**Probability Density Function**

$$f(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha}, & x > 0 \\ 0, & x \leq 0, \end{cases}$$

for  $\alpha, \beta \geq 0$ .

$$X \sim \text{Gamma}(\alpha, \beta)$$

**Verification of the properties**

$$\begin{aligned} & \int_{-\infty}^{+\infty} f(x) dx \\ &= \int_0^{\infty} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx \\ &= \int_0^{\infty} \frac{(x/\beta)^{\alpha-1} \beta^{\alpha-1} e^{-(x/\beta)}}{\Gamma(\alpha) \beta^\alpha} \beta d(x/\beta) \\ &= \int_0^{\infty} \frac{1}{\Gamma(\alpha)} (x/\beta)^{\alpha-1} e^{-(x/\beta)} d(x/\beta) \\ &= \frac{1}{\Gamma(\alpha)} \int_0^{\infty} y^{\alpha-1} e^{-y} dy \\ &= \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) \\ &= 1. \end{aligned}$$



**Moment**

$$\begin{aligned}
\mathbb{E}[X^p] &= \int_{-\infty}^{+\infty} x^p f(x) dx \\
&= \int_0^{\infty} x^p \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx \\
&= \int_0^{\infty} \frac{x^{p+\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx \\
&= \int_0^{\infty} \frac{\beta^{p+\alpha-1} (x/\beta)^{p+\alpha-1} e^{-(x/\beta)}}{\Gamma(\alpha) \beta^\alpha} \beta d(x/\beta) \\
&= \frac{\beta^p}{\Gamma(\alpha)} \int_0^{\infty} (x/\beta)^{p+\alpha-1} e^{-(x/\beta)} d(x/\beta) \\
&= \frac{\beta^p \Gamma(\alpha + p)}{\Gamma(\alpha)}.
\end{aligned}$$

**Moment Generating Function**

$$\begin{aligned}
\mathbb{E}[e^{tX}] &= \int_0^{\infty} e^{tx} \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha) \beta^\alpha} dx \\
&= \frac{1}{\Gamma(\alpha) \beta^\alpha} \int_0^{\infty} x^{\alpha-1} e^{-x(\frac{1}{\beta} - t)} dx \\
&= \frac{1}{\Gamma(\alpha)} \left( \frac{1}{1 - t\beta} \right)^\alpha \int_0^{\infty} \left[ \left( \frac{1 - t\beta}{\beta} \right) x \right]^{\alpha-1} e^{-\left( \frac{1 - t\beta}{\beta} \right) x} d\left( \frac{1 - t\beta}{\beta} x \right) \\
&= \frac{1}{\Gamma(\alpha)} \left( \frac{1}{1 - t\beta} \right)^\alpha \int_0^{\infty} y^{\alpha-1} e^{-y} dy. \\
&= \frac{1}{\Gamma(\alpha)} \left( \frac{1}{1 - t\beta} \right)^\alpha \Gamma(\alpha) \\
&= \left( \frac{1}{1 - t\beta} \right)^\alpha
\end{aligned}$$

This integral exists when  $t < \frac{1}{\beta}$ . So

$$M(t) = \left( \frac{1}{1 - \beta t} \right)^\alpha,$$

if  $t < \frac{1}{\beta}$ .

**Mean**

From moment:

$$\mathbb{E}[X] = \mathbb{E}[X^p] \Big|_{p=1} = \frac{\beta \Gamma(\alpha + 1)}{\Gamma(\alpha)} = \alpha \beta.$$

From moment generating function:

$$\mathbb{E}[X] = M'(0) = \frac{d\left[\left(\frac{1}{1 - \beta t}\right)^\alpha\right]}{dt} \Big|_{t=0} = (\alpha \beta (1 - \beta t)^{-\alpha-1}) \Big|_{t=0} = \alpha \beta.$$

**Variance**

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[X^p] \Big|_{p=1} = \frac{\beta^2 \Gamma(\alpha+2)}{\Gamma(\alpha)} = \beta^2 \alpha(\alpha+1). \\ \text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \beta^2 \alpha(\alpha+1) - (\beta\alpha)^2 = \alpha\beta^2.\end{aligned}$$

**9.6 Normal Distribution****Probability Density Function**

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

for  $\mu \in \mathbb{R}, \sigma^2 > 0$ .

$$X \sim \text{Normal}(\mu, \sigma^2)$$

**Verification of the properties**

$$\begin{aligned}& \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \\&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)\right] \sigma \frac{1}{\sqrt{2}} \left(\frac{(x-\mu)^2}{2\sigma^2}\right)^{\frac{1}{2}-1} d\left[\frac{(x-\mu)^2}{2\sigma^2}\right] \\&= \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi}} e^{-y} y^{\frac{1}{2}-1} dy \\&= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{\frac{1}{2}-1} e^{-y} dy \\&= \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) \\&= \frac{1}{\sqrt{\pi}} \sqrt{\pi} \\&= 1.\end{aligned}$$

**Moment Generating Function** Say  $X \sim N(\mu, \sigma^2)$ . So  $X = \sigma Z + \mu$  for some  $Z \sim N(0, 1)$ . Then

$$\begin{aligned}M_Z(t) &= \mathbb{E}[e^{tZ}] \\&= \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\&= e^{t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-t)^2}{2}\right\} dx \\&= e^{t^2/2} \cdot 1 \\&= e^{t^2/2}.\end{aligned}$$

So

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \frac{\sigma^2 t^2}{2}}.$$

## 9.7 Bivariate Normal Distribution

Let  $\mathbf{X} = (X_1, \dots, X_n)$  be a random vector. Let  $\boldsymbol{\mu}$  be a vector of expectations. Let  $\Sigma$  be a matrix of covariates.

$$X \sim MVN(\boldsymbol{\mu}, \Sigma).$$

## 9.8 Weibull Distribution

Probability Density Function:

$$f(x) = \begin{cases} \frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(\frac{x}{\theta})^\beta}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

for  $\alpha, \beta > 0$ .

$$X \sim Weibull(\theta, \beta)$$

Verification of the properties:

$$\begin{aligned} & \int_{-\infty}^{+\infty} f(x) dx \\ &= \int_0^\infty \frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(\frac{x}{\theta})^\beta} dx \\ &= \int_0^\infty \frac{\beta}{\theta^\beta} \theta^{\beta-1} \left[\left(\frac{x}{\theta}\right)^\beta\right]^{\frac{\beta-1}{\beta}} e^{-(\frac{x}{\theta})^\beta} \frac{\theta}{\beta} \left[\left(\frac{x}{\theta}\right)^\beta\right]^{\frac{1}{\beta}-1} d\left[\left(\frac{x}{\theta}\right)^\beta\right] \\ &= \int_0^\infty e^{-(\frac{x}{\theta})^\beta} d\left[\left(\frac{x}{\theta}\right)^\beta\right] \\ &= \int_0^\infty e^{-y} dy \\ &= 1. \end{aligned}$$

## 9.9 Chi-squared Distribution

**Definition**

$$\chi_{(k)}^2 = \sum_{i=1}^k Z_i^2$$

where  $Z_1, \dots, Z_k \stackrel{iid}{\sim} N(0, 1)$ .

**Proposition 9.9.1.** *If  $Z \sim G(0, 1)$ , then  $Z^2 \sim \chi^2(1)$ .*

**Proposition 9.9.2.** *Let  $W_1, \dots, W_n$  be independent variables such that  $W_i \sim \chi^2(k_i)$  for each  $i \in \{1, \dots, n\}$ . Define  $S := \sum_{i=1}^n W_i$ . then*

$$S \sim \chi^2\left(\sum_{i=1}^n k_i\right).$$

#### Probability Density Function

$$f(x, k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}.$$

#### Moment Generating Function

$$M_{\chi^2_{(k)}}(t) = (1 - 2t)^{-k/2}.$$

#### Mean and Variance

Let  $X \sim \chi^2(k)$ . Then

$$\begin{aligned} E(X) &= k \\ \text{Var}(X) &= 2k. \end{aligned}$$

## 9.10 t Distribution

#### Definition

Let  $X \sim N(0, 1)$  and  $Y \sim \chi^2_{(n)}$  be independent. Then

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_{(n)}.$$

## 9.11 Properties

**Proposition 9.11.1** (Probability Integral Transformation). *Let  $X$  be a continuous random variable. Let  $F$  be the cumulative distribution function of  $X$ . Let  $Y$  be a random variable given by  $Y = F(X)$ . Then  $Y$  has a  $\text{Uniform}(0, 1)$  distribution.*

*Proof.* For  $y \in (0, 1)$ ,

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(F(X) \leq y) \\ &= P(X \leq F^{-1}(y)) \\ &= F(F^{-1}(y)) \\ &= y. \end{aligned}$$

■

10

## Unclassified

**Theorem 3.** *Let  $X$  and  $Y$  be continuous random variables. Let  $f$  be a joint probability density function of  $X$  and  $Y$ . Let  $S$  be an injective transformation given by*

$$S(x, y) = (u, v) = (h_1(x, y), h_2(x, y)).$$

*Let  $T$  denote the inverse transformation of  $S$ .*

$$T(u, v) = (x, y) = (w_1(u, v), w_2(u, v)).$$

*Let  $g$  denote the joint probability density function of  $U$  and  $V$ . Then*

$$g(u, v) = f(w_1(u, v), w_2(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|.$$