Contents

Chapter 1

Combinatorics, Set Theory, and Probability

1.1 Counting Methods

1. Basic Principle: $n_1 \times n_2 \times \dots$
2. Ordered Sampling without Replacement—Permutation (or Arrangement):
$$ {}_nA_k = \frac{n!}{(n-k)!}. $$
3. Ordered Sampling with Replacement:
$$ n^k. $$
4. Unordered Sampling without Replacement—Combination:
$$ {}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{n}{n-k}. $$
5. Unordered Sampling with Replacement:
$$ \binom{n+k-1}{k}. $$
6. Combination is Permutation without order. Combination is also called n choose k.
7. Multiple Combination:
   (a) $\binom{n}{k_1, k_2, \dots, k_m} = \frac{n!}{k_1!k_2!\dots k_m!}$ where $n = \sum_{i=1}^{m} k_i$.
   (b) For the two cases situation, $n = k_1 + k_2 \Rightarrow \binom{n}{k_1 k_2} = \frac{n!}{k_1!k_2!}$ *if and only if* $\binom{n}{k_1}$ *if and only if* $\binom{n}{k_2}$.

1.2 From Set to Probability

Set Theory	**Probability**
Element	Outcome
Subset	Event
Universal Set	Sample Space (Ω)

1. Here are the common Set Operations: $A \cup B, A \cap B, A^c$.

2. A probability $\mathbb{P}(\cdot)$ is a function that maps events in the sample space to real numbers such that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\text{Event}) \geqslant 0$, and $\mathbb{P}(\Omega) = 1$, where $\emptyset$ is null set has no element (*i.e.*, event has no outcome).
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(AB)$, where $\mathbb{P}(AB) = \mathbb{P}(A \cap B)$.
4. Union Bound: $\mathbb{P}(A \cup B) \leqslant \mathbb{P}(A) + \mathbb{P}(B)$. And $\mathbb{P}\big(\cup_{i=1}^{N} A_i\big) \leqslant \sum_{i=1}^{N} \mathbb{P}(A_i)$ for more than two sets.

Axiom (Axioms of Probability)**.** *The three axioms of probability are:*

1. $0 \leqslant \mathbb{P}(A) \leqslant 1$ *for any event* A.
2. $\mathbb{P}(\Omega) = 1$.
3. *If* A_i *are Mutually Exclusive, then* $\mathbb{P}\big(\cup_{i=1}^{N} A_i\big) = \sum_{i=1}^{N} \mathbb{P}(A_i)$.

1.3 Venn Diagram

1.4 Set Properties and Corresponding Probability Properties

1. Mutually Exclusive: $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cap B) = 0$, which implies $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
2. Collectively Exhaustive: $\cup_{i=1}^{N} A_i = \Omega \Rightarrow \mathbb{P}\big(\cup_{i=1}^{N} A_i\big) = 1$.
3. Partitions (*i.e.*, Mutually Exclusive & Collectively Exhaustive): $\mathbb{P}\big(\cup_{i=1}^{N} A_i\big) = \sum_{i=1}^{N} \mathbb{P}(A_i) = 1$.
4. All outcomes constitute a partition.
5. A and A^c constitute a partition.
6. If B_i are Collectively Exhaustive, then $\mathbb{P}\big(A \cap (\cup_{i=1}^{N} B_i)\big) = \mathbb{P}(A\Omega) = \mathbb{P}(A)$
7. For any B, $\mathbb{P}(A) = \mathbb{P}(A \cap (B \cup B^c)) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$

Chapter 2

Conditional Probability and Independent

2.1 Conditional Probability

Theorem 2.1 (Conditional Probability)**.** *If* $\mathbb{P}(B) > 0$*, then*

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

1. If A_i are Mutually Exclusive: $\mathbb{P}\big(\cup_{i=1}^{N} A_i \mid B\big) = \sum_{i=1}^{N} \mathbb{P}(A_i \mid B)$.
2. If A_i are Collectively Exhaustive: $\mathbb{P}\big(\cup_{i=1}^{N} A_i \mid B\big) = \mathbb{P}(\Omega \mid B) = \mathbb{P}(\Omega B)/\mathbb{P}(B) = 1$.
3. If A_i are partitions: $\sum_{i=1}^{N} \mathbb{P}(A_i \mid B) = \mathbb{P}\big(\cup_{i=1}^{N} A_i \mid B\big) = \mathbb{P}(\Omega \mid B) = \mathbb{P}(\Omega B)/\mathbb{P}(B) = 1$.
4. If B_i are Mutually Exclusive:

$$\mathbb{P}\big(A \mid \cup_{i=1}^{N} B_i\big) = \mathbb{P}\big(A \cap (\cup_{i=1}^{N} B_i)\big)/\mathbb{P}\big(\cup_{i=1}^{N} B_i\big) = \sum_{i=1}^{N} \mathbb{P}(AB_i)/\sum_{i=1}^{N} \mathbb{P}(B_i).$$

5. If B_i are Collectively Exhaustive: $\mathbb{P}\big(A \mid \cup_{i=1}^{N} B_i\big) = \mathbb{P}(A \mid \Omega) = \mathbb{P}(A\Omega)/\mathbb{P}(\Omega) = \mathbb{P}(A)$
6. If B_i are Partitions **(Law of Total Number)**,

$$\begin{aligned}
\mathbb{P}(A \mid B) &= \sum_{i=1}^{N} \mathbb{P}(AB_i)/\sum_{i=1}^{N} \mathbb{P}(B_i) && (B_i \text{ are Mutually Exclusive})\\
&= \sum_{i=1}^{N} \mathbb{P}(AB_i) && (B_i \text{ are Collectively Exhaustive})\\
&= \sum_{i=1}^{N} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i). && (\text{Definition of Conditional Probability})
\end{aligned}$$

Theorem 2.2 ($\mathbb{P}(\cdot \mid F)$ is a Probability)**.** *The conditional probability satisfies the axioms of probability.*

1. $0 \leqslant \mathbb{P}(E \mid F) \leqslant 1$.

2. $\mathbb{P}(\Omega \mid F) = 1$.

3. If E_i *are mutually exclusive, then* $\mathbb{P}\big(\cup_{i=1}^{N} E_i \mid F\big) = \sum_{i=1}^{N} \mathbb{P}(E_i \mid F)$.

Theorem 2.3 (Bayes' Theorem)**.** *If* $\mathbb{P}(A) > 0$ *and* $\mathbb{P}(B) > 0$*, then*

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

2.2 Independent

Definition 2.4 (Independent). A and B are independent if and only if $\mathbb{P}(AB) = \mathbb{P}(A)\mathbb{P}(B)$.

Theorem 2.5. *If A and B are independent, then*

1. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B)$.
2. $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$.
3. *If A and B are independent then A^c and B are independent and so on.*

Proof.

$$\begin{aligned}
\mathbb{P}(B) = \mathbb{P}((A \cup A^c) \cap B) &= \mathbb{P}(AB) + \mathbb{P}(A^cB) && (A \text{ and } A^c \text{ are partitions}) \\
\Rightarrow \mathbb{P}(A^cB) &= \mathbb{P}(B) - \mathbb{P}(AB) \\
&= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) && (A \text{ and } B \text{ are independent}) \\
&= \mathbb{P}(B)\,(1 - \mathbb{P}(A)) \\
&= \mathbb{P}(B)\mathbb{P}(A^c).
\end{aligned}$$

□

Theorem 2.6 (Independent Trails). *The Probability of k_0 failures and k_1 successes in $n = k_0 + k_1$ Independent Trails with success rate p is*

$$\binom{n}{k_0}(1-p)^{k0}p^{k_1} = \binom{n}{k_1}(1-p)^{k0}p^{k_1}.$$

If $n = k_1 + k_2 + \cdots + k_m$ and success rates are $p_1, p_2, \ldots, p_m$, where $\sum_{i=1}^{m} p_i = 1$, the probability of such independent trails is

$$\binom{n}{k_1, k_2, \ldots, k_m} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m}.$$

2.3 Tree Diagram

Chapter 3

Discrete Random Variables

3.1 Probability Mass Function (PMF)

Definition 3.1. The probability mass function (PMF) of a discrete random variable X is

$$P_X(x) = \mathbb{P}(X = x).$$

Theorem 3.2 (Properties of PMF)**.** *For random variable X, the PMF $P_X(x)$ has the following properties:*

1. $P_X(x) \geqslant 0$.
2. $\sum_{x \in S_X} P_X(x) = 1$.

3.2 Histogram

3.3 Families of Discrete Random Variables

1. Bernoulli(p): Single experiment with success rate p (***i.e.*, Flip a coin**), x is the number of successes

$$P_X(x) = \begin{cases} 1-p & x = 0, \\ p & x = 1, \\ 0 & otherwise. \end{cases}$$

2. Binomial(n, p): A sequence of **n** independent Bernoulli(p) experiments, x is the number of successes

$$P_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \dots, n, \\ 0 & otherwise. \end{cases}$$

Note: Binomial($1, p$) *if and only if* Bernoulli(p).

Note: Binomial(n, p) *if and only if* Independent trails.

3. Poisson(α): Binomial(n, p) with $n \to \infty$, and $\alpha = np$, x is the number of successes

$$P_X(x) = \begin{cases} \dfrac{\alpha^x e^{-\alpha}}{x!} & x = 0, 1, \dots, \\ 0 & otherwise. \end{cases}$$

4. Geometric(p): Get the **1st** success at the **xth** independent Bernoulli(p) experiment

$$P_X(x) = \begin{cases} p(1-p)^{x-1} & x = 1, 2, \dots, \\ 0 & otherwise. \end{cases}$$

5. Pascal(k, p): Get the **kth** success at the **xth** independent Bernoulli(p) experiment

$$P_X(x) = \begin{cases} \binom{x-1}{k-1} p^k (1-p)^{x-k} & x = k, k+1, k+2, \dots, \\ 0 & \text{otherwise.} \end{cases}$$

Note: Pascal(k, p) is also called Negative Binomial(k, p).

Note: Pascal$(1, p)$ *if and only if* Geometric(p).

Note: Pascal(k, p) is a sequence of **k** independent Geometric(p) experiments.

6. Discrete Uniform(k, l): outcomes are uniformly distributed on range (k, l) ***i.e.*, Roll a Die**

$$P_X(x) = \begin{cases} 1/(l-k+1) & x = k, k+1, k+2, \dots, l, \\ 0 & \text{otherwise.} \end{cases}$$

3.4 Cumulative Distribution Function (CDF)

Definition 3.3. The cumulative distribution function (CDF) of a discrete random variable X is

$$F_X(x) = P_X[X \leqslant x] = \sum_{k=0}^{x} P_X(k).$$

$$F_X(b) - F_X(a) = \sum_{k=0}^{b} P_X(k) - \sum_{k=0}^{a} P_X(k) = \sum_{k=a+1}^{b} P_X(k) = P_X(a < X \leqslant b).$$

Theorem 3.4 (The CDF of Geometric (p) is worth to remember)**.**

$$\begin{aligned} F_X(x) &= P_X[X \leqslant x] \\ &= 1 - P_X[X > x] \\ &= 1 - \sum_{i=x+1}^{\infty} p(1-p)^{i-1} \\ &= 1 - (1-p)^x \sum_{i=1}^{\infty} p(1-p)^{i-1} \\ &= 1 - (1-p)^x. \end{aligned}$$

3.5 Expected Value

Definition 3.5 (Average)**.** In ordinary language, an **Average** is a single number taken as representative of a list of numbers.

1. Mode: The outcome appears the most often in the sample space

$$P_X(x_{\text{mode}}) \geqslant P_X(x).$$

2. Median: The outcome which separates the higher half from the lower half of a sample space

$$P_X[X \leqslant x_{\text{med}}] \geqslant 1/2, \qquad P_X[X \geqslant x_{\text{med}}] \geqslant 1/2.$$

3. (Arithmetic) mean/Expectation: The sum of all the outcomes divided by the number of outcomes

$$\mu_x = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Definition 3.6 (Expected Value)**.** The expected value of a discrete random variable X with PMF $P_X(x)$ is

$$\mathbb{E}[X] = \sum_{x \in S_X} x P_X(x). \qquad \text{(First Moment of X)}$$

$$\mathbb{E}\left[X^2\right] = \sum_{x \in S_X} x^2 P_X(x). \qquad \text{(Second Moment of X)}$$

Theorem 3.7 (Important Expectations)**.** *Here are some important expectations:*

1. *Bernoulli(p):*
$$\mathbb{E}[X] = 0 \cdot P_X(0) + 1 \cdot P_X(1) = 0(1-p) + 1(p) = p.$$
2. *Binomial(n, p):*
$$\mathbb{E}[X] = np.$$
3. *Poisson(α):*
$$\mathbb{E}[X] = \alpha.$$
4. *Geometric(p):*
$$\mathbb{E}[X] = p \cdot 1 + (1-p) \cdot (1 + \mathbb{E}[X]) \Rightarrow 1/p.$$
5. *Pascal(k, p):*
$$\mathbb{E}[X] = k/p.$$
6. *Discrete Uniform(k, l):*
$$\mathbb{E}[X] = (k+l)/2.$$

Note:

1. From an engineering perspective, **Mean (including Expectations, etc.)** is numerically easier to calculate, either using human brain or computers, than Mode and Median, when the sample space is humongous.
2. In most cases, average, mean and expectation refer to the same concept.

3.6 Derived Random Variable and Variance

Theorem 3.8 (Derived Random Variable)**.** *Given random variable* X*, let* $Y = g(X)$*, then*

1. $P_Y(y) = P[Y = y] = P[Y = g(x)] = P[g^{-1}(Y) = g^{-1}(g(x))] = P[X = x] = P_X(x)$
2. $\mathbb{E}[Y] = \sum y P_Y(y) = \sum g(x) P_X(x)$
3. $\mathbb{E}[X - \mu_x] = \sum_{x \in S_X} (x - \mu_x) P_X(x) = \sum_{x \in S_X} x P_X(x) - \mu_x \sum_{x \in S_X} P_X(x) = \mathbb{E}[X] - \mu_x \cdot 1 = 0$
4. $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b \Rightarrow \mathbb{E}[b] = \mathbb{E}[0 \cdot X + b] = b$

Definition 3.9 (Variance and Standard Deviation)**.** For random variable X, the variance (σ_x^2) is defined as

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

and the standard deviation (σ_x) is defined as

$$\sigma_x = \sqrt{\mathrm{Var}[X]}.$$

Theorem 3.10. *The variance of a random variable X with has the following properties:*

1. $\mathrm{Var}[X] \geqslant 0$
2. $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$

Theorem 3.11 (Important Variance)**.** *Here are some important variances:*

1. *Bernoulli(p):*
$$\mathrm{Var}[X] = p(1 - p).$$
2. *Binomial(n, p):*
$$\mathrm{Var}[X] = np(1 - p).$$
3. *Poisson(α):*
$$\mathrm{Var}[X] = \alpha.$$
4. *Geometric(p):*
$$\mathrm{Var}[X] = (1 - p)/p^2.$$
5. *Pascal(k, p):*
$$\mathrm{Var}[X] = k(1 - p)/p^2.$$
6. *Discrete Uniform(k, l):*
$$\mathrm{Var}[X] = (l - k)(l - k + 2)/12.$$

Chapter 4

Continuous Random Variables

Axiom. *A random variable* X *is continuous if the sample space* S_X *consists of one or more intervals. For* $x \in S_X$, $P_X(x) = 0$.

4.1 Cumulative Distribution Function (CDF)

Definition 4.1. The CDF of continuous random variable X is

$$F_X(x) = \mathbb{P}(X \leqslant x).$$

Theorem 4.2. *For any random variable* X,

1. $F_X(-\infty) = 0$
2. $F_X(\infty) = 1$
3. $\mathbb{P}(x_1 < X \leqslant x_2) = F_X(x_2) - F_X(x_1)$

4.2 Probability Density Function (PDF)

Start with a continuous random variable X with CDF $F_X(x)$. The probability of "X with volume $\triangle$" is defined as:

$$\begin{aligned}\mathbb{P}(x < X \leqslant x + \triangle) &= F_X(x + \triangle) - F_X(x)\\ &= \frac{F_X(x + \triangle) - F_X(x)}{(x + \triangle) - x} \cdot \triangle.\end{aligned}$$

Definition 4.3 (Probability Density Function (PDF))**.**

$$f_X(x) = \lim_{\triangle \to 0} \frac{F_X(x + \triangle) - F_X(x)}{\triangle} = \frac{dF_X(x)}{dx}.$$

Theorem 4.4. *For a continuous random variable* X *with* PDF $f_X(x)$,

1. $f_X(x) \geqslant 0$ *for all* x
2. $F_X(u) = \int_{-\infty}^{u} f_X(x)\,dx$
3. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$

Theorem 4.5.

$$\mathbb{P}(x_1 < X \leqslant x_2) = \int_{x_1}^{x_2} f_X(x)\,dx.$$

4.3 Expected Value

Definition 4.6 (Expected value)**.**

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

Theorem 4.7 (Integral Identity)**.** *For every non-negative random variable* X*,*

$$\mathbb{E}[X] = \int_0^{\infty} 1 - F_X(u)\, du = \int_0^{\infty} \mathbb{P}(X > u)\, du.$$

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{\infty} x f_X(x)\, dx && (X \text{ is non-negative}) \\
&= \int_0^{\infty} \left(\int_0^{x} 1\, du \right) f_X(x)\, dx \\
&= \int_0^{\infty} \left(\int_u^{\infty} f_X(x)\, dx \right) du && (\text{Some Algebra}) \\
&= \int_0^{\infty} 1 - F_X(u)\, du = \int_0^{\infty} \mathbb{P}(X > u)\, du. && (\text{Definition of CDF})
\end{aligned}$$

□

Theorem 4.8 (Derived Random Variable)**.**

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

Theorem 4.9. *For any random variable* X*,*

1. $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$,
2. $\mathbb{E}[X - \mu_x] = 0$,
3. $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mu_x^2$,
4. $\mathrm{Var}[aX + b] = a^2 \,\mathrm{Var}[X]$.

4.4 Families of Continuous Random Variables

1. Continuous Uniform(k, l): A continuous counterpart of Discrete Uniform(k, l)

$$\begin{aligned}
f_X(x) &= \begin{cases} \frac{1}{l-k} & k \leqslant x \leqslant l \\ 0 & otherwise. \end{cases} \\
F_X(x) &= \frac{x-k}{l-k}. \qquad x \in (k, l) \\
\mathbb{E}[X] &= (l + k)/2. \\
\mathrm{Var}[X] &= (l - k)^2/12.
\end{aligned}$$

2. Exponential(λ): Get the **1st** success at the $\mathbf{x}$**th** unit of time. This is a continuous counterpart of Geometry(p) where $p = \lim_{\triangle t \to 0} \lambda \triangle t$ (*i.e.* λ is the probability density of success per unit of time)

$$\begin{aligned} f_X(x) &= \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & otherwise. \end{cases} \\ F_X(x) &= 1 - e^{-\lambda x}. \\ \mathbb{E}[X] &= 1/\lambda. \\ \mathrm{Var}[X] &= 1/\lambda^2. \end{aligned}$$

3. Erlang(k, λ): Get the **kth** success at the $\mathbf{x}$**th** unit of time. This is a continuous counterpart of Pascal(k,p)

$$\begin{aligned} f_X(x) &= \begin{cases} \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} & x \geqslant 0 \\ 0 & otherwise. \end{cases} \\ F_X(x) &= \frac{\gamma(k, \lambda x)}{(k-1)!}. \\ \mathbb{E}[X] &= k/\lambda. \\ \mathrm{Var}[X] &= k/\lambda^2. \end{aligned}$$

4. Gamma(α, β): Erlang(k, λ) with $k = \alpha$ being a positive real number (not limits to integer), $\lambda = \beta$ and $\Gamma(\alpha) = (\alpha - 1)!$.

$$\begin{aligned} f_X(x) &= \begin{cases} \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} & x \geqslant 0 \\ 0 & otherwise. \end{cases} \\ F_X(x) &= \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}. \\ \mathbb{E}[X] &= \alpha/\beta. \\ \mathrm{Var}[X] &= \alpha/\beta^2. \end{aligned}$$

   Note: Both Erlang(k, λ) and Gamma(α, β) are a sequence of independent Exponential(λ) experiments.

4.5 Gauvsian Random Variables

We have seen the continuous counterpart of Discrete Uniform, Geometric and Pascal random variables. It is natural to ask what is the continuous counterpart of Binomial random variable. The answer is Gauvsian random variable. We will show how to derive the Gauvsian random variable from Binomial random variable in section 7.4. But first, let's start from the PDF.

Definition 4.10 (Gauvsian Random Variable)**.** X is a $\mathsf{Gauvsian}(\mu, \sigma)$ random variable if the PDF of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

X is also called $\mathsf{Normal}(\mu, \sigma)$ random variable. We will use $N(\mu, \sigma)$ in the following content.

Theorem 4.11 (The Expectation and Variance of $X \sim N(\mu, \sigma)$)**.**

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Theorem 4.12. *If* X *is* $N(\mu, \sigma)$*,* $Y = aX + b$ *is* $N(a\mu + b, a\sigma)$.

Definition 4.13 (Standard Normal Random Variable)**.** The $\mathsf{N}(\mu, \sigma)$ with $\mu = 0, \sigma = 1$ is called standard normal random variable $\mathsf{Z} \sim \mathsf{N}(0, 1)$. The PDF is,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}).$$

And the CDF is

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

Theorem 4.14. *If* X *is* $\mathsf{N}(\mu, \sigma)$*, the* CDF *of* X *is*

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

The probability that X *is in the interval* (a, b) *is*

$$\mathbb{P}(a < X \leqslant b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Theorem 4.15. $\Phi(-z) = 1 - \Phi(z)$*.*

Chapter 5

Joint Random Variables

5.1 Joint CDF

Definition 5.1 (Joint CDF)**.** The joint CDF of random variables X and Y is

$$F_{X,Y}(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y).$$

The joint CDF is a **complete** probability model for any pair of random variables X and Y.

Theorem 5.2. *For any pair of random variables, X and Y, the following properties hold:*

1. $0 \leqslant F_{X,Y}(x, y) \leqslant 1$,
2. $F_{X,Y}(\infty, \infty) = 1$,
3. $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = 0$,
4. $F_{X,Y}(x, y)$ *is non-decreasing in x and y.*

Definition 5.3 (Marginal CDF)**.**

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

5.2 Joint PMF

Definition 5.4 (Joint PMF)**.** The joint PMF of random variables X and Y is

$$P_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

The joint PMF is a **complete** probability model for any pair of discrete random variables X and Y.

Theorem 5.5. *For discrete random variables X and Y and any set B in the X, Y plane, the probability of the event is*

$$\mathbb{P}(\{\, B \,\}) = \sum_{(x,y)\in B} P_{X,Y}(x, y).$$

Apparently, the joint PMF is non-negative and sums to one.

$$\sum_{x\in S_X} \sum_{y\in S_Y} P_{X,Y}(x, y) = 1.$$

Definition 5.6 (Marginal PMF)**.** For discrete random variables X and Y with joint PMF $P_{X,Y}(x, y)$,

$$P_X(x) = \sum_{y\in S_Y} P_{X,Y}(x, y), \qquad P_Y(y) = \sum_{x\in S_X} P_{X,Y}(x, y).$$

For discrete random variables, the marginal PMF $P_X(x)$ and $P_Y(y)$ are probability models for the individual random variables X and Y, but they only provide an **incomplete** probability model for the pair of random variables X and Y.

5.3 Joint PDF

Definition 5.7 (Joint PDF)**.** The joint CDF of continuous random variables X and Y is a function $f_{X,Y}(x,y)$ with the property

$$F_{X,Y}(u,v) = \int_{-\infty}^{u} \int_{-\infty}^{v} f_{X,Y}(x,y)\,dx\,dy.$$

Apparently, we can then derive the joint PDF as follows,

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

The joint PDF is a **complete** probability model for any pair of continuous random variables X and Y.

Theorem 5.8. *The probability that the continuous random variables* (X,Y) *are in any set* A

$$\mathbb{P}(\{\,A\,\}) = \iint_A f_{X,Y}(x,y)\,dx\,dy.$$

The joint PDF is non-negative and integrates to one.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$

Definition 5.9 (Marginal PDF)**.** For continuous random variables X and Y with joint PDF $f_{X,Y}(x,y)$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

For continuous random variables, the marginal PDFs $f_X(x)$ and $f_Y(y)$ are probability models for the individual random variables X and Y, but they only provide an **incomplete** probability model for the pair of random variables X and Y.

5.4 Expected Value

Theorem 5.10 (Expected Value of a Function of Two Random Variables)**.** *The function of two random variables (*i.e., $g(X,Y)$*) is also a random variable, then the expected value of* $g(X,Y)$ *is*

$$\mathbb{E}[g(X,Y)] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x,y) P_{X,Y}(x,y); \qquad \text{(Discrete)}$$

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy. \qquad \text{(Continuous)}$$

Theorem 5.11. *The expectation of a linear combination of several functions can be easily derived as follows,*

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

Corollary 5.12 (Sum of two random variables)**.** *For random variables* X *and* Y*,*

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

Note: The corollary 5.12 states that the expectation of sum of random variables does not need the joint probability model of all random variables. However, this is not true for the variance.

Corollary 5.13 (The variance of the sum of two random variables)**.**

$$\mathrm{Var}[aX+bY] = a^2\,\mathrm{Var}[X] + b^2\,\mathrm{Var}[Y] + 2ab\mathbb{E}[(X-\mu_X)(Y-\mu_Y)].$$

5.5 Covariance, Correlation and Independent

Definition 5.14 (Covariance)**.** The covariance of two random variables X and Y is

$$\sigma_{xy} = \mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Definition 5.15 (Correlation Coefficient)**.** The correlation coefficient of two random variables X and Y is

$$\rho_{xy} = \mathrm{Corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Note: Correlation coefficient is a dimensionless quantity. There are other definitions of correlation coefficient, but this is the most common one.

Theorem 5.16. *If X and Y are random variables such that $Y = aX + b$,*

$$\rho_{X,Y} = \begin{cases} -1, & a < 0 \\ 0, & a = 0 \\ 1, & a > 0 \end{cases}$$

which also implies

$$-1 \leqslant \rho_{xy} \leqslant 1.$$

Definition 5.17 (Correlation)**.** The correlation of random variables X and Y is

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

This is a different parameter from the correlation coefficient (and widely used in engineer major).

Definition 5.18 (Uncorrelatedness)**.** Random variables X and Y are uncorrelated if and only if

$$\mathrm{Cov}[X, Y] = 0.$$

Definition 5.19 (Independence)**.** Random variables X and Y are independent if and only if

$$P_{X,Y}(x, y) = P_X(x)P_Y(y); \qquad \text{(Discrete)}$$
$$f_{X,Y}(x, y) = f_X(x)f_Y(y). \qquad \text{(Continuous)}$$

It's easy to show that if X and Y are independent, then

$$F_{X,Y}(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y) = \mathbb{P}(X \leqslant x)\mathbb{P}(Y \leqslant y) = F_X(x)F_Y(y).$$

Theorem 5.20. *If Random variables X and Y are independent*

1. $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, *so that* $\mathrm{Cov}[X, Y] = 0$,
2. $\mathrm{Var}[aX + bY] = a^2\,\mathrm{Var}[X] + b^2\,\mathrm{Var}[Y]$.

Definition 5.21 (Orthogonality)**.** If $r_{X,Y} = \mathbb{E}[XY] = 0$, then X and Y are orthogonal.

Theorem 5.22. *The relationship between uncorrelatedness, independence and orthogonality*

1. *Uncorrelatedness means changing the value of one random variable does not affect the **mean** of the other random variable.*
2. *Independence means changing the value of one random variable does not affect the probability distribution (*i.e.*, mean, variance and other moments as well) of the other random variable.*
3. *Uncorrelated is linear independent. But independent includes both linear and nonlinear independent. Thus, uncorrelatedness does not imply independence.* e.g., $X \sim \mathsf{Unif}[-1, 1]$ *and* $Y = X^2$.

4. *Orthogonality is a different concept from uncorrelatedness and independence. From the perspective of Linear Algebra, orthogonality is a concept regarding angle between random variables, while independence and uncorrelatedness are concepts regarding the length (or projection) of random variables.*

5. *Independence and Uncorrelatedness are preferred for easier calculation of the variance of linear combination of random variables.*

6. *Joint Gauvsian Random Variables has a preferred property: uncorrelatedness* if and only if *independence. This is the fundamental of a lot of (I would say more than 85%) stochastic modeling, machine learning, etc.*

5.6 Bivariate Gauvsian Random Variables

Definition 5.23 (Bivariate Gauvsian Random Variables)**.** Random variables X and Y are bivariate Gauvsian if and only if their joint PDF is given by the following equation,

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right\}.$$

Where $\mu_x = \mathbb{E}[X]$, $\mu_y = \mathbb{E}[Y]$, $\sigma_x^2 = \mathrm{Var}[X]$, $\sigma_y^2 = \mathrm{Var}[Y]$, and $\rho = \mathrm{Corr}[X,Y]$.

Theorem 5.24. *If X and Y are bivariate Gauvsian, then X is the $N(\mu_X, \sigma_X)$ and Y is the $N(\mu_Y, \sigma_Y)$.*

Theorem 5.25. *Bivariate Gauvsian Random Variables are uncorrelated if and only if they are independent.*

Theorem 5.26. *If X and Y are bivariate Gauvsian, and W_1 and W_2 are given by linear independent equations,*

$$W_1 = a_1X + b_1Y, \qquad W_2 = a_2X + b_2Y,$$

then W_1 and W_2 are bivariate Gauvsian random variables such that

$$\begin{aligned}
\mathbb{E}[W_i] &= a_i\mu_X + b_i\mu_Y, \\
\mathrm{Var}[W_i] &= a_i^2\sigma_X^2 + b_i^2\sigma_Y^2 + 2a_ib_i\sigma_X\sigma_Y\rho, \\
\mathrm{Cov}[W_1, W_2] &= a_1a_2\sigma_X^2 + b_1b_2\sigma_Y^2 + (a_1b_2 + a_2b_1)\sigma_X\sigma_Y\rho.
\end{aligned}$$

Chapter 6

Conditional Random Variables

6.1 Conditioning by an Event

6.1.1 Conditioning One Random Variable by an Event

Definition 6.1 (Conditional CDF)**.** Given the event B with $\mathbb{P}(B) > 0$, the conditional CDF of X given B is

$$F_{X|B}(x) = \mathbb{P}(X \leqslant x \mid B).$$

Definition 6.2 (Conditional PMF)**.** Given the event B with $\mathbb{P}(B) > 0$, the conditional PMF of X given B is

$$P_{X|B}(x) = \mathbb{P}(X = x \mid B) = \begin{cases} \frac{P_X(x)}{\mathbb{P}(B)} & x \in B, \\ 0 & \textit{otherwise}. \end{cases}$$

Definition 6.3 (Conditional PDF)**.** Given the event B with $\mathbb{P}(B) > 0$, the conditional PDF of X given B is

$$f_{X|B}(x) = \frac{dF_{X|B}(x)}{dx} = \begin{cases} \frac{f_X(x)}{\mathbb{P}(B)} & x \in B, \\ 0 & \textit{otherwise}. \end{cases}$$

Theorem 6.4. *Discrete* X*:*

1. *For any* $x \in B$*,* $P_{X|B}(x) \geqslant 0$*,*
2. $\sum_{x \in B} P_{X|B}(x) = 1$*,*
3. *The conditional probability that* X *is in the set* C *is* $\mathbb{P}(C \mid B) = \sum_{x \in C} P_{X|B}(x)$*.*

Continuous X*:*

1. *For any* $x \in B$*,* $f_{X|B}(x) \geqslant 0$*,*
2. $\int_{x \in B} f_{X|B}(x)\,dx = 1$*,*
3. *The conditional probability that* X *is in the set* C *is* $\mathbb{P}(C \mid B) = \int_{x \in C} f_{X|B}(x)\,dx$*.*

6.1.2 Conditional Expected Value by an Event

Definition 6.5 (Conditional Expected Value)**.** The conditional expected value of random variable X and $Y = g(X)$ given condition B is

Discrete:

$$\mathbb{E}[X \mid B] = \sum_{x \in B} x P_{X|B}(x),$$

$$\mathbb{E}[Y \mid B] = \sum_{x \in B} g(x) P_{X|B}(x),$$

Continuous:

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

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

Theorem 6.6. *The conditional variance of random variable* X *given condition* B *is*

$$\mathrm{Var}[X \mid B] = \mathbb{E}\left[X^2 \mid B\right] - \mathbb{E}[X \mid B]^2.$$

To get used to the conditional probability, think "$X \mid B$" as a random variable instead of an operation.

6.1.3 Conditioning Joint Random Variables by an Event

Definition 6.7 (Conditional Joint CDF)**.** For random variables X and Y and an event with $\mathbb{P}(B) > 0$, the conditional joint CDF of X and Y given B is

$$F_{X,Y|B}(x, y) = \mathbb{P}(X \leqslant x, Y \leqslant y \mid B).$$

Definition 6.8 (Conditional Joint PMF)**.** For discrete random variables X and Y and an event with $\mathbb{P}(B) > 0$, the conditional joint PMF of X and Y given B is

$$P_{X,Y|B}(x, y) = \mathbb{P}(X = x, Y = y \mid B) = \begin{cases} \frac{P_{X,Y}(x,y)}{\mathbb{P}(B)}. & (x, y) \in B, \\ 0. & otherwise. \end{cases}$$

Definition 6.9 (Conditional Joint PDF)**.** For continuous random variables X and Y and an event with $\mathbb{P}(B) > 0$, the conditional joint PDF of X and Y given B is

$$f_{X,Y|B}(x, y) = \frac{\partial^2 F_{X,Y|B}(x, y)}{\partial x \partial y} = \begin{cases} \frac{f_{X,Y}(x,y)}{\mathbb{P}(B)}. & (x, y) \in B, \\ 0. & otherwise. \end{cases}$$

Definition 6.10 (Conditional Marginal CDF, PMF/PDF)**.** For random variables X and Y and an event with $\mathbb{P}(B) > 0$, the conditional marginal CDF, PMF and PDF of X given B is

$$\begin{aligned} F_{X|B}(x) &= F_{X,Y|B}(x, \infty), \\ P_{X|B}(x) &= \sum_{y \in S_Y} P_{X,Y|B}(x, y), \\ f_{X|B}(x) &= \int_{-\infty}^{\infty} f_{X,Y|B}(x, y)\, dy. \end{aligned}$$

Definition 6.11 (Conditional Joint Expected Value)**.** For random variables X and Y and an event with $\mathbb{P}(B) > 0$, the conditional expected value of $W = g(X, Y)$ given B is

Discrete: $$\mathbb{E}[W \mid B] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) P_{X,Y|B}(x, y),$$

Continuous: $$\mathbb{E}[W \mid B] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y|B}(x, y)\, dx\, dy.$$

Theorem 6.12. *The conditional variance of* $W = g(X, Y)$ *given* B *is*

$$\mathrm{Var}[W \mid B] = \mathbb{E}\left[W^2 \mid B\right] - \mathbb{E}[W \mid B]^2.$$

6.2 Conditioning by a Random Variable

Definition 6.13. For any event $Y = y$ such that $P_Y(y) > 0$, the conditional PMF of X given $Y = y$ is

$$P_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}.$$

Definition 6.14. For any event $Y = y$ such that $f_Y(y) > 0$, the conditional PDF of X given $Y = y$ is

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

Theorem 6.15. *If* X *and* Y *are independent, then*

$$P_{X|Y}(x \mid y) = P_X(x), \qquad P_{Y|X}(y \mid x) = P_Y(y).$$

6.2.1 Expectation of Conditioning by a Random Variable with Fixed Value

Definition 6.16. The conditional expected value of X given $Y = y \in S_Y$ is

$$\text{Discrete:} \qquad \mathbb{E}[X \mid Y = y] = \sum_{x \in S_X} x P_{X|Y}(x \mid y),$$

$$\text{Continuous:} \qquad \mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx.$$

Theorem 6.17. *If random variables X and Y are independent, then*

$$\mathbb{E}[X \mid Y = y] = \mathbb{E}[X], \qquad \mathbb{E}[Y \mid X = x] = \mathbb{E}[Y].$$

Note: The result of $\mathbb{E}[X \mid Y = y]$ is a function of y and the result of $\mathbb{E}[Y \mid X = x]$ is a function of x. When X and Y are independent, the result of $\mathbb{E}[X \mid Y = y]$ is a constant and the result of $\mathbb{E}[Y \mid X = x]$ is a constant. Because when X and Y are independent, changing one variable does not affect the probability distribution of the other.

Theorem 6.18. *The conditional variance of X given $Y = y \in S_Y$ is*

$$\mathrm{Var}[X \mid Y = y] = \mathbb{E}\left[X^2 \mid Y = y\right] - \mathbb{E}[X \mid Y = y]^2.$$

6.2.2 Expectation of Conditioning by a Random Variable

Theorem 6.19. *If Y is unspecified, then "$X \mid Y$" is a function of both X and Y, whose expectation is determined by joint distribution $P_{X,Y}(x, y)$, (equivalently, you can think Y as a partitions).*

$$\textit{Discrete:} \qquad \mathbb{E}[X \mid Y] = \sum_{x \in S_X} \sum_{y \in S_Y} x P_{X|Y}(x \mid y) P_y(y) = \sum_{x \in S_X} \sum_{y \in S_Y} x P_{X,Y}(x, y),$$

$$\textit{Continuous:} \qquad \mathbb{E}[X \mid Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y)\, dx\, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x, y)\, dx\, dy.$$

Theorem 6.20. *Following theorem 6.19, the variance of "$X \mid Y$" is*

$$\begin{aligned} \mathrm{Var}[X \mid Y] &= \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^2 \mid Y\right] \\ &= \mathbb{E}\left[X^2 \mid Y\right] - \mathbb{E}[X \mid Y]^2. \end{aligned}$$

Theorem 6.21 (Law of Total Expectation)**.** *If Y is unspecified, then $\mathbb{E}[X \mid Y]$ is the function of Y, which is also a random variable. We certainly are interested in* ***the expectation of*** $\mathbb{E}[X \mid Y]$,

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

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \int_{-\infty}^{\infty} \mathbb{E}[X \mid Y = y] f_Y(y)\, dy && (\mathbb{E}[X \mid Y] \text{ is a function of } Y) \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx \right) f_Y(y)\, dy && (\text{definition 6.16}) \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\, dy\, dx && (\text{Some Algebra}) \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, dx && (\text{definition 6.14}) \\
&= \int_{-\infty}^{\infty} x f_X(x)\, dx && (\text{definition 5.9}) \\
&= \mathbb{E}[X]. && (\text{definition of } \mathbb{E}[X])
\end{aligned}$$

□

Theorem 6.22. *Following theorem 6.21, the variance of* $\mathbb{E}[X \mid Y]$ *is*

$$\mathrm{Var}[\mathbb{E}[X \mid Y]] = \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right] - \mathbb{E}[X]^2.$$

Theorem 6.23. *We also are interested the expectation of* $\mathrm{Var}[X \mid Y]$ *for the same reason of theorem 6.21,*

$$\mathbb{E}[\mathrm{Var}[X \mid Y]] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[\mathbb{E}[X \mid Y]^2\right].$$

Theorem 6.24 (Law of Total Variance)**.** *The combination of theorems 6.22 and 6.23 leads to a very useful formula,*

$$\mathrm{Var}[X] = \mathbb{E}[\mathrm{Var}[X \mid Y]] + \mathrm{Var}[\mathbb{E}[X \mid Y]].$$

Chapter 7

Derived Probability Models

7.1 Functions of Discrete Random Variables

Theorem 7.1. *For discrete random variables* X *and* Y*, the derived random variable* $W = g(X, Y)$ *has* PMF

$$P_W(w) = \sum_{\{(x,y) \mid g(x,y)=w\}} P_{X,Y}(x, y).$$

7.2 Functions of Continuous Random Variables

Theorem 7.2. *For continuous random variables* X *and* Y*, the* PDF *of derived random variable* $W = g(X, Y)$ *can be derived as*

1. *Find the* CDF $F_W(w) = P_W(W \leqslant w)$.
2. *Find the* PDF $f_W(w) = dF_W(w)/dw$.

7.2.1 Functions of One Continuous Random Variable

Theorem 7.3. *If* $W = aX$ *where* $a > 0$ *is a constant, then*

$$F_W(w) = F_X(w/a) \qquad f_W(w) = \frac{1}{a} f_X(\frac{w}{a}).$$

1. *If* X *is* $\mathsf{Unif}(b, c)$*, then* W *is* $\mathsf{Unif}(ab, ac)$.
2. *If* X *is* $\mathsf{Exp}(\lambda)$*, then* W *is* $\mathsf{Exp}(\lambda/a)$.
3. *If* X *is* $\mathsf{Erlang}(k, \lambda)$*, then* W *is* $\mathsf{Erlang}(k, \lambda/a)$.
4. *If* X *is* $\mathsf{N}(\mu, \sigma)$*, then* W *is* $\mathsf{N}(a\mu, a\sigma)$.

Theorem 7.4. *If* $W = X + b$*, where* b *is a constant,*

$$F_W(w) = F_X(w - b) \qquad f_W(w) = f_X(w - b).$$

Theorem 7.5. *Let* U *be a* $\mathsf{Unif}(0, 1)$ *random variable and let* $F(x)$ *denote a* CDF *with an inverse* $F^{-1}(u)$ *defined for* $0 < u < 1$. *Then* $X = F^{-1}(U)$ *is a random variable with* CDF $F_X(x) = F(x)$.

7.2.2 Functions of Joint Continuous Random Variables

Following the same idea as in theorem 7.2, we can easily derive the PDF of derived random variable $W = g(X, Y)$ when the function is linear. It is more complex for other functions.

Theorem 7.6. *For continuous random variables* X *and* Y*, the* CDF *of* $W = g(X, Y)$ *is*

$$F_W(w) = \mathbb{P}(W \leqslant w) = \iint_{\{(x,y)|g(x,y)\leqslant w\}} f_{X,Y}(x, y)\,dx\,dy.$$

Corollary 7.7. *For continuous random variables* X *and* Y*, the* CDF *of* $W = \max(X, Y)$ *is*

$$F_W(w) = F_{X,Y}(w, w) = \int_{-\infty}^{w}\int_{-\infty}^{w} f_{X,Y}(x, y)\,dx\,dy.$$

Hint: $\{\max(X, Y) \leqslant w\} = \{\{X \leqslant w\} \cap \{Y \leqslant w\}\}$.

Theorem 7.8 (The sum of two random variables)**.** *The* PDF *of* $W = X + Y$ *is*

$$f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(x, w - x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(w - y, y)\,dy.$$

Corollary 7.9. *When* X *and* Y *are independent, the* PDF *of* $W = X + Y$ *is*

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)f_Y(w - x)\,dx = \int_{-\infty}^{\infty} f_X(w - y)f_Y(y)\,dy.$$

7.3 Sum (*i.e.*, Linear Combinations) of Random Variables

7.3.1 Basic Properties

Definition 7.10. Random Variables of the form

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

are called **sums of random variables**.

Theorem 7.11 (Expected Values of Sums)**.** *A generalized version of corollary 5.12*

$$\mathbb{E}[W_n] = \mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

Theorem 7.12 (Variance of Sums)**.** *A generalized version of corollary 5.13*

$$\mathrm{Var}[W_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}[X_i, X_j].$$

Corollary 7.13. *When* $X_1, X_2, \ldots, X_n$ *are uncorrelated, the variance of the sum is*

$$\mathrm{Var}[W_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

7.3.2 Methods of Generating Functions

It is quite intuitive to consider an experimental random variable as the linear combination of many well-defined random variables. However, it is not easy to calculate the moments related information (*e.g.* mean, variance, *etc.*) in such a case. The generating function method provides a way to solve this problem.

Definition 7.14 (Probability Generating Function)**.** If X is a **non-negative discrete** random variable, then the probability generating function of X is defined as

$$G_X(z) = \mathbb{E}\left[z^X\right] = \sum_{x=0}^{\infty} z^x P_X(x).$$

You can think of it as the *Z-Transform* of $P_X(x)$.

Definition 7.15 (Moment Generating Function)**.** For a random variable X, if the **moment generating function** is existed, it is defined as

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \begin{cases} \sum_{x \in S_X} e^{tx} P_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx & \text{if } X \text{ is continuous} \end{cases} \end{aligned}$$

You can think of it as the *Laplace Transform* of $P_X(x)/f_X(x)$. Note that $M_X(t) = G_X(e^t)$.

Random Variables	MGF
Bernoulli(p)	$1 - p + pe^t$
Binomial(n,p)	$(1 - p + pe^t)^n$
Geometric(p)	$\frac{pe^t}{1-(1-p)e^t}$
Pascal(k,p)	$\left(\frac{pe^t}{1-(1-p)e^t}\right)^k$
Poisson(α)	$e^{\alpha(e^t-1)}$
Discrete Univorm(k,l)	$\frac{e^{kt}-e^{(l+1)t}}{e^t-1}$
Constant(c)	e^{ct}
Continuous Univorm(k,l)	$\frac{e^{kt}-e^{lt}}{t(l-k)}$
Exponential(λ)	$\frac{\lambda}{\lambda-t}$
Erlv(n,λ)	$\left(\frac{\lambda}{\lambda-t}\right)^n$
Gauvsian(μ,σ)	$e^{\mu t+\frac{1}{2}\sigma^2 t^2}$

Table 7.1: Common MGF

Theorem 7.16. *A random variable* X *with* MGF $M_X(t)$ *has* n*th moment*

$$\mathbb{E}[X^n] = M_X^{(n)}(0) = \left.\frac{d^n M_X(t)}{dt^n}\right|_{t=0}.$$

Theorem 7.17. *The* MGF *of* $Y = aX + b$ *is* $M_Y(t) = e^{bt} M_X(at)$.

Proof.

$$M_Y(t) = \mathbb{E}\left[e^{tY}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[e^{t(aX+b)}\right] \\
&= e^{bt}\mathbb{E}\left[e^{(at)X}\right] \\
&= e^{bt}M_X(at).
\end{aligned}
$$

□

Corollary 7.18 (Central Moment). *The* MGF *of* $Y = X - \mu$ *is,*

$$M_Y(t) = e^{-\mu t}M_X(t).$$

Corollary 7.19 (standardized Moment). *The* MGF *of* $Y = (X - \mu)/\sigma$ *is,*

$$M_Y(t) = e^{-\mu t/\sigma}M_X(t/\sigma).$$

Remark. Let's clarify a few concepts.

1. Centering: $Y = X - \mu$. Align the distribution to the origin.
2. Normalization: $Y = X/\sigma$. Eliminate the variance. This make the distribution independently of any linear change of scale. It is important when dealing with joint distributions.
3. Standardization: $Y = (X - \mu)/\sigma$. Normalized Centering.

Note: If the MGF exists, it uniquely determines the probability distribution. It is powerful because the calculation of moments becomes derivative instead of integral. MGF fails to calculate the probability.

Theorem 7.20 (MGF of the Sum of Independent Random Variables). *If* $X_1, X_2, \ldots, X_n$ *are independent random variables with* MGF*s* $M_{X_1}(t), M_{X_2}(t), \ldots, M_{X_n}(t)$*, then the* MGF *of* $W_n = X_1 + X_2 + \cdots + X_n$ *is*

$$M_{W_n}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t).$$

If $X_1, X_2, \ldots, X_n$ *are also identically distributed, then*

$$M_{W_n}(t) = [M_X(t)]^n.$$

Definition 7.21 (Random Sums of i.i.d. Random Variables). Let N be a random variable with PMF $P_N(n)$, and let $X_1, X_2, \ldots$ be a sequence of independent random variables. Then the random variable $W = X_1 + X_2 + \cdots + X_N$ is called a **random sum of i.i.d. random variables**.

Theorem 7.22 (MGF of Random Sums of i.i.d. Random Variables). *Let* $\{X_i\}$ *be a sequence of i.i.d. random variables with* MGF $M_X(t)$*, and let* N *be a non-negative integer-valued random variable independent of* $\{X_i\}$ *with* PMF $P_N(n)$*. Then the* MGF *of* $W = X_1 + X_2 + \cdots + X_N$ *is*

$$M_W(t) = M_N(\ln M_X(t)).$$

Theorem 7.23 (Expectation of Random Sums of i.i.d. Random Variables).

$$\mathbb{E}[W] = \mathbb{E}[N]\mathbb{E}[X].$$

Theorem 7.24 (Variance of Random Sums of i.i.d. Random Variables).

$$\mathrm{Var}[W] = \mathrm{Var}[X]\mathbb{E}[N] + \mathbb{E}[X]^2\,\mathrm{Var}[N].$$

Definition 7.25 (Cumulant Generating Function). The **cumulant** of a random variable X is defined as the natural logarithm of the MGF of X:

$$K_X(t) = \ln M_X(t).$$

The n-th-order cumulant K_n is

$$K_n(X) = K^{(n)}(0).$$

Note: In some cases theoretical treatments of problems in terms of cumulants are simpler than those using moments. In particular, when two or more random variables are statistically independent, the n-th-order cumulant of their sum is equal to the sum of their n-th-order cumulants, which is preferred over the product of their n-th-order moments.

Note: The third and higher-order cumulants of a **Gauvsian** distribution are zero, and it is the only distribution with this property.

Theorem 7.26 (The First Three Cumulants). *For random variable* X*,*

1. $K_1(X) = \mathbb{E}[X]$ *which is the mean.*
2. $K_2(X) = \mathrm{Var}[X]$ *which is the variance or the second central moment.*
3. $K_3(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^3\right]$ *is the third central moment.*

Order	Moment			Cumulant	
	Raw	Central	Standarized	Raw	Normalized
1	Mean	0	0	Mean	-
2	-	Variance	1	Variance	1
3	-	-	Skewness	-	Skewness

Table 7.2: Relation between moments and cumulants on the first three orders:

Definition 7.27 (Characteristic Function). For a random variable X, the **characteristic function** is defined as

$$\begin{aligned}\phi_X(t) &= \mathbb{E}\left[e^{itX}\right] \\ &= \begin{cases} \sum_{x\in S_X} e^{itx} P_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} e^{itx} f_X(x)\,dx & \text{if } X \text{ is continuous} \end{cases}\end{aligned}$$

You can think of it as the *Fourier Transform* of $P_X(x)/f_X(x)$. Note that $\phi_X(t) = M_X(it)$. The characteristic function is always defined.

7.4 Central Limit Theorem

In this part of the lecture, we will see a sequence of essentially random or unpredictable events can sometimes be expected to settle down into a behavior that is essentially unchanging when items far enough into the sequence are studied. This is the fundamental theory that why probability (*i.e.* a measure of uncertainty) can be used to model the deterministic world.

Theorem 7.28 (Central Limit Theorem). *Given* $X_1, X_2, \ldots$ *a sequence of i.i.d. random variables with expected value* μ_X *and variance* σ_X^2*, the* CDF *of* $Z_n = (\sum_{i=1}^n X_i - n\mu_X)/\sqrt{n\sigma_X^2}$ *has the property*

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z).$$

*That is the **CDF** of "sum of standardized i.i.d. random variables (not necessary Gauvsian)" converges to the standard Gauvsian random variable as* $n \to \infty$*. Alternatively, we can express it as*

$$\lim_{n\to\infty} \frac{n(\bar{X} - \mu_X)}{\sqrt{n}\sigma_X} \sim N(0,1).$$

Proof.

$$\mathbb{E}\left[\exp\left\{it \cdot \frac{\sum_{i=1}^n X_i - n\mu_X}{\sqrt{n\sigma_X^2}}\right\}\right] = \mathbb{E}\left[\exp\left\{\sum_{i=1}^n it \cdot \frac{X_i - \mu_X}{\sqrt{n\sigma_X^2}}\right\}\right]$$

$$
\begin{aligned}
&= \left\{ \mathbb{E}\left[\exp\left\{ it \cdot \frac{X_i - \mu_X}{\sqrt{n\sigma_X^2}} \right\} \right] \right\}^n \\
&= \left\{ 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{2n} \right) \right\}^n \\
&\to \exp\left\{ -\frac{1}{2} t^2 \right\} \sim \mathsf{N}(0,1).
\end{aligned}
$$

□

Corollary 7.29. *With theorem 7.28, we can express the* $W_n = X_1 + X_2 + \cdots + X_n$ *with i.i.d.* X_i *as*

$$W_n = \sqrt{n\sigma_X^2} Z_n + n\mu_X.$$

The CDF *of* W_n *is*

$$F_{W_n}(w) = \mathbb{P}\left(\sqrt{n\sigma_X^2} Z_n + n\mu_X \leqslant w \right) = \mathbb{P}\left(Z_n \leqslant \frac{w - n\mu_X}{\sqrt{n\sigma_X^2}} \right) = \Phi\left(\frac{w - n\mu_X}{\sqrt{n\sigma_X^2}} \right).$$

Theorem 7.30 (De Moivre-Laplace Theorem)**.** *For a binomial random variable* $X \sim \mathsf{Bin}(n, p)$*,*

$$\mathbb{P}(x_1 \leqslant X \leqslant x_2) \approx \mathbb{P}(x_1 - 0.5 \leqslant X \leqslant x_2 + 0.5) \approx \Phi\left(\frac{x_2 + 0.5 - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{x_1 - 0.5 - np}{\sqrt{np(1-p)}} \right).$$

Chapter 8

Introduction of Information Theory

Definition 8.1 (bit)**.** The **bit** is the unit of information. It is representing a binary choice between two alternatives (such as true/false, on/off, or 0/1 in binary code). Therefore, if an event has a probability of 1/2, it carries 1 bit of information, because observing the event eliminates one out of two equally likely possibilities. If the probability is 1/4, it carries 2 bits of information, and so on.

Definition 8.2 (self-information)**.** The **self-information** of an event X with PMF $P_X(x)$ is defined as

$$I(X) = -\log_2 P_X(x).$$

Consequently, the self-information of 1 bit (*i.e.*, $p(x) = 1/2$) is 1. The 2-based logarithmic is the most common choice in the information theory, but the natural logarithm is also used in some cases.

Definition 8.3 (Shannon-entropy)**.** The **Shannon-entropy** of a discrete random variable X with PMF $P_X(x)$ is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x) = \sum_{x\in\mathcal{X}} P_X(x)I(x) = \mathbb{E}[I(X)].$$

The $\mathcal{X}$ is the sample space of X, which is the common notation in the information theory.

Definition 8.4 (Joint entropy)**.** The **joint entropy** of two discrete random variables X and Y with PMF $P_{X,Y}(x, y)$ is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x, y)\log_2 P_{X,Y}(x, y) = \mathbb{E}[I(X, Y)].$$

Definition 8.5 (Conditional entropy)**.** The **conditional entropy** of two discrete random variables X and Y with PMF $P_{XY}(x, y)$ is defined as

$$\begin{aligned}
H(X|Y) &= \sum_{y\in\mathcal{Y}} P_Y(y)H(X|Y = y)\\
&= -\sum_{y\in\mathcal{Y}} P_Y(y)\sum_{x\in\mathcal{X}} P_{X|Y}(x|y)\log_2 P_{X|Y}(x|y)\\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{X,Y}(x, y)\log_2 P_{X|Y}(x|y)\\
&= \mathbb{E}[I(X \mid Y)].
\end{aligned}$$

Theorem 8.6 (Chain Rule of Entropy)**.**

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

The entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the amount of information required on the average to describe the random variable. The relative entropy is a measure of the distance between two distributions or the inefficiency of assuming that the distribution is q when the true distribution is p.

Definition 8.7 (Relative entropy (Kullback-Leibler divergence))**.** The **relative entropy** of two PMF $P_X(x)$ and $Q_X(x)$ is defined as

$$D(P_X \| Q_X) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{Q_X(x)}.$$

Definition 8.8 (Mutual information)**.** The **mutual information** of two discrete random variables X and Y is the KL-divergence between their joint distribution and their products (marginal) distributions.

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log_2 \frac{P_{X,Y}(x,y)}{P_X(x) P_Y(y)}$$
$$D(P_{X,Y} \| P_X P_Y).$$

Theorem 8.9 (Mutual Information and Entropy)**.**

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
I(X;Y) &= H(Y) - H(Y|X) \\
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
I(X;Y) &= I(Y;X) \\
I(X;X) &= H(X)
\end{aligned}$$

So mutual information is the reduction in the uncertainty of X due to the knowledge of Y. If X and Y are independent, then $I(X;Y) = 0$. This leads to the interpretation of independence as observing Y does not reduce the uncertainty in X if they are independent.

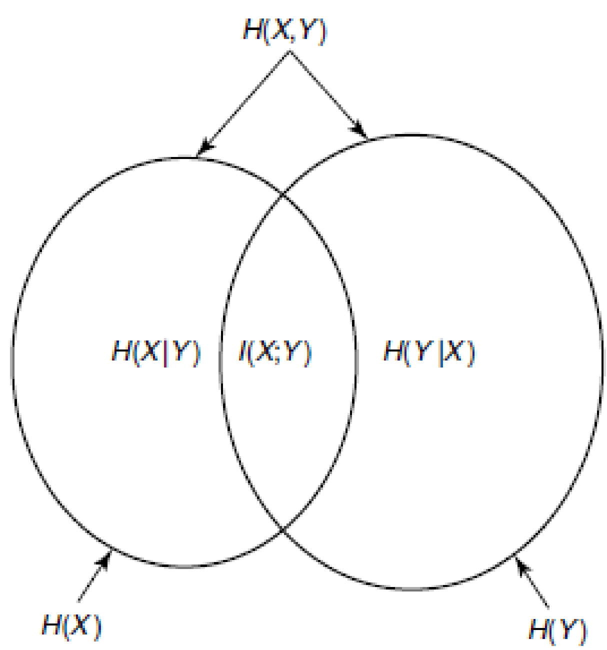

Figure 8.1: Relationship between entropy and mutual information.