Probability Theory

Daniel Mao

Contents

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'w 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 call the **sample space**, denoted by* Ω*.*

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 $X_1, ..., X_n$ be random variables on S. We define the **joint cumulative distribution function** of $X_1, ..., X_n$, denoted by $F(x_1, ..., x_n)$, to be a function given by*

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

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

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

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

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

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

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

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

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

(4) Right Continuity.

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

(5)

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

(6)

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

(7)

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

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).

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

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

■

1.4 Marginal Distributions

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

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

2

Probability Functions

2.1 Probability Function of Events

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

(1) Non-negativity.
$P(A) \geq 0$ *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*

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

Proposition 2.1.1 (Properties of Probability Functions)**.** *Let* Ω *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 Ω 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

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

So

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

■

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*

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

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*

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

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

Expectation

3.1 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 $X = (X_1, ..., X_n)$ be a random vector. We define the* ***expectation*** *of X to be a vector given by*

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

3.2 Properties of the Expectation Operator

Proposition 3.2.1 (Linearity)**.** *Expectation is a linear operator. i.e., Let* $X = (X_1, ..., X_n)$ *be a random vector. Let* $\vec{\lambda} = (\lambda_1, ..., \lambda_n)$ *be a constant. Then*

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

Or,

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

Proposition 3.2.2. *Let* X *be a random vector. Let* $g_1, ..., g_n$ *be real-valued functions on* X*. Let* $\lambda_1, ..., \lambda_n$ *be constants. Then*

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

3.3 Variance and 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*

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

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

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

Proposition 3.3.2.

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

Proposition 3.3.3.

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

Proposition 3.3.4. *Let X and Y be two random variables. Then*

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

Proof.

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

That is,

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

■

Proposition 3.3.5 (Bilinearity of the Covariance Operator)**.** *Let $X = (X_1, ..., X_n)$ 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*

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

Or,

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

3.4 Theory in Higher Dimensions

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

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

Proposition 3.4.1.

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

Proposition 3.4.2. *Covariance matrices are symmetric.*

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

Proposition 3.4.3. *Let X be a random vector. Then $\operatorname{var}(X)$ is positive definite. i.e., $\forall a \in \mathbb{R}^n : a^T \operatorname{var}(X) a \geq 0$.*

3.5 Moment

Definition (Moment)**.** *We define the k^{th} **moment** (about 0) of X for $k \in \mathbb{N}$ to be the number given by*

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

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

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

Remark. *The first moment is the mean (expectation).*

Proposition 3.5.1.

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

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

Proof.

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

■

3.6 Moment Generating Functions

Definition ((Joint) Moment Generating Function)**.** *Let $X_1, ..., X_n$ be random variables. We define the **(joint) moment generating function** of $X_1, ..., X_n$, denoted by M, to be a function given by*

$$M(t_1, ..., t_n) := \mathbb{E}\big[\exp\big\{\sum_{i=1}^{n} t_i X_i\big\}\big],$$

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

Proposition 3.6.1.

$$M(0) = 1.$$

Proposition 3.6.2 (Expansion of the Moment Generating Function)**.** *Let X be a random variable. Let Φ_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.

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

That is,

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

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

Proposition 3.6.3. *Let X be a random variable. Let Φ_X be the moment generating function of X. Given the moment generating function of X, we can extract its n^{th} moment, for $n \in \mathbb{N}$, via*

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

Proposition 3.6.4 (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 $(-\frac{h}{|a|}, \frac{h}{|a|})$.

Proposition 3.6.5 (Linear Combinations)**.** *Let X_i for $i = 1, ..., n$ be independent random variables. Let M_{x_i} be the moment generating function for X_i, for $i = 1, ..., n$. Let $a_i \in \mathbb{R}$ for $i = 1, ..., n$). Define $X := \sum_{i=1}^{n} a_i X_i$. Then the moment generating function M_X for X is*

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(a_i t).$$

Proposition 3.6.6 (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_X be the cumulative distribution function of Y. Then $M_X = M_Y$ if and only if $F_X = F_Y$.*

4

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.*

4.1 Discrete Uniform Distribution

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

4.2 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 $X \sim Bernoulli(p)$.*

Proposition 4.2.1 (Probability Density Function of Bernoulli Distribution).

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

Proposition 4.2.2 (Expectation of Bernoulli Distribution).

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

Example 4.2.1. *Flipping a coin once.*

4.3 Binomial Distribution

Definition (Binomial Distribution). *Let $X_i \sim Bernoulli(p)$ for $i \in \{1, ..., n\}$. 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 $X \sim Binomial(n, p)$. Then X records the number of "success" trails.*

Proposition 4.3.1 (Probability Density Function of Binomial Distribution).

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

Proposition 4.3.2 (Expectation of Binomial Distribution).

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

4.4 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 $X \sim NB(k, p)$.*

$X :=$ # of 0 outcomes before the $r^{\text{th}} outcome of 1 in repeated Bernoulli(p) experiments$
$X \sim NegBin(r, p)$.
$P(X = x) = \binom{x+r-1}{x}(1-p)^x p^{r-1} p$.
$X = \sum_{i=1}^{r} X_i$
$X_i \sim Geo(p)$.

4.5 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 $X \sim Geo(p)$.*

4.6 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 4.6.1 (Probability Function of Hypergeometric Distribution)**.** *For* $x = \max\{0, n-N+r\}, ..., \min\{n, r\}$,

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

4.7 Multinomial Distribution

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

$$(X_1, ..., X_n) \sim Multinomial(n, p_1, ..., p_k).$$

Joint Probability Mass Function

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

Joint Moment Generating Function

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

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

Marginal Distribution

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

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

Conditional Distribution

Proposition 4.7.1.

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

for $i \neq j$.

Proposition 4.7.2.

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

Other Properties

Proposition 4.7.3. *Let $T := X_i + X_j$. Then $T \sim Binomial(n, p_i + p_j)$.*

Proof. Idea: use MGF. ■

Proposition 4.7.4. $\operatorname{cov}(X_i, X_j) = -np_ip_j$.

Proof.

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

■

4.8 Poisson Distribution

Definition (Poisson Distribution)**.** *Let $X \sim Poisson(\lambda)$ 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 λ 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 4.8.1 (Moment Generating Function)**.** *The moment generating function of a Poisson(λ) 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 4.8.2 (Mean and Variance)**.** *The mean and variance of a Poisson(λ) distributed random variable are*

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

Proof.

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

∎

Proposition 4.8.3. *When n is large and p is small, Poisson(np) can be used bo approximate Binomial(n, p).*

Proof.

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

■

4.9 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*

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

5

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 $\mathbb{R}$.

5.1 Continuous Uniform Distribution

5.2 Beta Distribution

5.3 Exponential Distribution

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

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

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

Proposition 5.3.1 (Mean and Variance)**.** *Then mean and variance of a $Exponential(\lambda)$ distributed random variable are*

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

5.4 Erlang Distribution

Proposition 5.4.1 (Probability Density Function)**.** *For $x > 0$,*

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

Proposition 5.4.2. *$Erlang(1, \lambda) = Exponential(\lambda)$.*

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

$$X \sim Gamma(\alpha, \beta)$$

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

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

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

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

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:

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

Variance

$$\mathbb{E}[X^2] = \mathbb{E}[X^p]\Big|_{p=1} = \frac{\beta^2\Gamma(\alpha+2)}{\Gamma(\alpha)} = \beta^2\alpha(\alpha+1).$$

$$Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \beta^2\alpha(\alpha+1) - (\beta\alpha)^2 = \alpha\beta^2.$$

5.6 Normal Distribution

Probability Density Function

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

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

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

Verification of the properties

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

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

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

So

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

5.7 Bivariate Normal Distribution

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

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

5.8 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:

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

5.9 Chi-squared Distribution

Definition

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

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

Proposition 5.9.1. *If $Z \sim G(0, 1)$, then $Z^2 \sim \chi^2(1)$.*

Proposition 5.9.2. *Let $W_1, ..., W_n$ be independent variables such that $W_i \sim \chi^2(k_i)$ for each $i \in \{1, ..., n\}$. Define $S := \sum_{i=1}^{n} W_i$. then*

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

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

$$E(X) = k$$
$$Var(X) = 2k.$$

5.10 t Distribution

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

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

5.11 Properties

Proposition 5.11.1 (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 $Uniform(0, 1)$ distribution.*

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

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

■

6

Conditional Probability Distributions

6.1 Conditional Probability of Events

Definition (Conditional Probability)**.** *Let Ω 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 $P(A \mid B)$, to be the number given by*

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

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

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

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

provided that $P(B) \neq 0$.
Let $\{A_i\}_{i=1}^{i=n}$ be a sequence of events. Then

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

where A_0 is defined to be Ω.

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

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

■

Proposition 6.1.2 (Law of Total Probability)**.** *Let Ω be a sample space. Let P be a probability function defined on the sample space. Let A be an event in Ω. Let $\{B_i\}_{i\in\mathbb{N}}$ be a countable collection of events in Ω. 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.

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

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)**.**

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

Proof.

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

■

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 $f_X(\cdot \mid y)$ and $f_Y(\cdot \mid x)$ 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*

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

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*

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

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

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 $E\big[X \mid Y = y\big]$.*

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 $Var[X \mid Y = y]$, to be the number given by*

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

Proposition 6.3.1 (Substitution Rule)**.**

$$E\big[h(X, Y) \mid Y = y\big] = E\big[h(X, y) \mid Y = y\big].$$

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

$$E\left[E\big[g(X) \mid Y\big]\right] = E[g(X)].$$

Proof.

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

■

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

$$\mathrm{var}[Y] = E\big[\,\mathrm{var}[Y \mid X]\big] + \mathrm{var}\left[E[Y \mid X]\right].$$

7

Joint Probability Distributions

7.1 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*

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

7.2 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 7.2.1. *Let S be a sample space. Let $X_1, ..., X_n$ be random variables on S. Let f be the joint probability mass function of $X_1, ..., X_n$. Let f_i be the marginal probability mass function of X_i, for some $i \in \{1, ..., n\}$. Then*

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

7.3 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* $\mathrm{range}(X) \times \mathrm{range}(Y)$ *to* $[0, 1]$ *given by*

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

7.4 Joint Expectations

Definition (Joint Expectation of Discrete Random Variables)**.** *Let X be discrete random vector. Let f be the joint probability mass function of X. Let A be the joint support of X. 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*

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

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

Definition (Joint Expectation of Continuous Random Variables)**.** *Let X be a d dimensional continuous random vector. Let f be the joint probability density function of X. Let g be a function on X. We define the **joint expectation** of $g(X)$ to be a number given by*

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

if $\int_{\mathbb{R}^d} |g(x)f(x)| dx < +\infty$; and we say that the expectation of X does not exist otherwise.

8

Independence

8.1 Independence of Events

8.1.1 Definitions

Definition (Independent Events)**.** *Let Ω be a sample space. Let P be a probability function defined on the sample space. Let A and B be two events in Ω. 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 $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.*

Proposition 8.1.1. *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 $P(B)P(A \mid B) = P(A \cap B)$, $P(A)P(B) = P(A \mid B)P(B)$.
Since $P(B) \neq 0$ and $P(A)P(B) = P(A \mid B)P(B)$, $P(A \mid B) = P(A)$.
Since $P(A \cap B) = P(A)P(B)$ and $P(A)P(B \mid A) = P(A \cap B)$, $P(A)P(B) = P(B \mid A)P(A)$.
Since $P(A) \neq 0$ and $P(A)P(B) = P(B \mid A)P(A)$, $P(B \mid A) = P(B)$.
For the reverse direction, assume that $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.
Since $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(A \mid B) = P(A)$, $P(A)P(B) = P(A \cap B)$.

■

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 $\mathbb{A}$ are* ***pairwise independent*** *if any pair of events are independent. i.e., $\forall i, j \in \{1, ..., n\}$, 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 $\mathbb{A}$ are* ***mutually independent*** *if any event* *is independent of the intersection of any other events. i.e.,* $\forall I \subseteq \{1, ..., n\}$, *we have* $P\big(\bigcap_{i\in I} A_i\big) = \prod_{i\in I} P(A_i)$.

8.1.2 Properties

Proposition 8.1.2 (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 8.1.3. *A zero-probability event is independent of any any other event.*

Proof. Let Ω be a sample space. Let P be a probability function defined on the sample space. Let A and B be two events in Ω. 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. ■

8.2 Independent Random Variables

8.2.1 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 8.2.1. *The 4 definitions of independence are equivalent.*

8.2.2 Properties

Proposition 8.2.2. *If X and Y are independent random variables and g and h are functions, then $g(X)$ and $h(Y)$ are independent.*

Proposition 8.2.3. *Let X and Y be random variables. Let g be a function on X. Then if X and Y are independent, we have*

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

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

8.2.3 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 $g : A_X \to \mathbb{R}$ and $h : A_Y \to \mathbb{R}$ 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 $f(x, y) = 0 < f_X(x)f_Y(y)$. ∎

8.2.4 Expectations of Independent Random Variables

Proposition 8.2.4. *Let $X_1, ..., X_n$ be independent random variables. Let g_i be a function on X_i, for each $i \in \{1, ..., n\}$. Then*

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

9

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

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