Probability Theory

Daniel Mao

Contents

Chapter 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).

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

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

■

Chapter 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 Mass Functions

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

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

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.*

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

2.3 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 2.3.1. *The probability density of a singleton set is 0.*

Proposition 2.3.2. $\forall x \in \mathbb{R}, f(x) \geq 0$.

Proposition 2.3.3.

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

Chapter 3

Joint Probability Distributions

> *Given random variables $X, Y, ...$, that are defined on a probability space, the joint probability distribution for $X, Y, ...$ is a probability distribution that gives the probability that each of $X, Y, ...$ falls in any particular range or discrete set of values specified for that variable. In the case of only two random variables, this is called a bivariate distribution, but the concept generalizes to any number of random variables, giving a multivariate distribution.*
>
> *The joint probability distribution can be expressed either in terms of a joint cumulative distribution function or in terms of a joint probability density function (in the case of continuous variables) or joint probability mass function (in the case of discrete variables). These in turn can be used to find two other types of distributions: the marginal distribution giving the probabilities for any one of the variables with no reference to any specific ranges of values for the other variables, and the conditional probability distribution giving the probabilities for any subset of the variables conditional on particular values of the remaining variables.*
>
> — Wikipedia, Joint probability distribution

3.1 Joint Cumulative Distribution Functions

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

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

3.2 Joint Probability Mass Functions

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

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

Proposition 3.2.1. *Let S be a sample space. Let $X_1, ..., X_n$ 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*

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

3.3 Joint Probability Density Functions

Definition (Joint Probability Density Functions)**.** *Let X and Y be continuous random variables. Let F be the joint cumulative distribution function of X and Y. We define the **joint probability density function** f of X and Y to be a function from* $\text{range}(X) \times \text{range}(Y)$ *to* $[0,1]$ *given by*

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

3.4 Marginal Distributions

Definition (Marginal Cumulative Distribution Function)**.** *Let S be a sample space. Let $X_1, ..., X_n$ 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*

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

Chapter 4

Expectation

4.1 Definition

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

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

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

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

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

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

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

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

4.2 Properties of the Expectation Operator

Proposition 4.2.1 (Expectation is a linear operator)**.** *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 4.2.2. *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*

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

4.3 Variance

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

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

or equivalently,

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

Proposition 4.3.1.

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

Proposition 4.3.2.

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

4.4 Moment

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

Proposition 4.4.1.

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

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

Proof.

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

■

4.5 Moment Generating Function

Proposition 4.5.1.

$$M(0) = 1.$$

Proposition 4.5.2 (Expansion of the Moment Generating Function)**.** *Let X be a random variable. Let 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.

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

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 4.5.3. *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 4.5.4 (Linear Transformations)**.** *Let X be a random variable. Let M_X be the moment generating function for X on $(-h, h)$ for some $h > 0$. Let $\alpha, \beta \in \mathbb{R}$ and $\alpha \neq 0$. Then the moment generating function $M_{\alpha X+\beta}$ for the random variable $\alpha X + \beta$ is*

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

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

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

Chapter 5

Joint Expectation

5.1 Joint Expectation

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

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

if $\sum_{\boldsymbol{x} \in A} |g(\boldsymbol{x}) f(\boldsymbol{x})| < +\infty$; and we say that $\mathbb{E}[g(\boldsymbol{X})]$ does not exist otherwise.

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

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

if $\int_A |g(\boldsymbol{x}) f(\boldsymbol{x})| d\boldsymbol{x} < +\infty$; and we say that $\mathbb{E}[g(\boldsymbol{X})]$ does not exist otherwise.

5.2 Covariance

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

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

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

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

Proposition 5.2.2. *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 5.2.3 (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,

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

5.3 Joint Moment

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

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

5.4 Joint Moment Generating Function

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 Φ, to be a function from $\mathbb{R}^n$ to $\mathbb{R}$ given by*

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

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 Φ is the set of all tuples $(t_1, ..., t_n)$ such that the RHS is defined.

5.5 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*

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

Proposition 5.5.1.

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

Proposition 5.5.2. *Covariance matrices are symmetric.*

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

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

Chapter 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 Probability Mass Function

Definition (Conditional Probability Mass Function)**.** *Let X and Y be two* discrete *random variables. Let f denote the joint probability mass function of X and Y. Let f_Y be the marginal probability mass function of Y. We define the* ***conditional probability mass function*** *of X given $Y = y_0$, denoted by $f_{X|Y}(\cdot \mid y_0)$, to be a function given by*

$$f_{X|Y}(x \mid y_0) := \frac{f(x, y_0)}{f_Y(y_0)},$$

provided that $f_Y(y_0) \neq 0$.

Definition (Conditional Probability Mass Function)**.** *Let $\mathcal{K}$ be a finite index set. Let $\mathcal{I}$ and $\mathcal{J}$ be a partition of $\mathcal{K}$. Let $(X_k)_{k\in\mathcal{K}}$ be discrete random variables. Let f denote the joint probability mass function of $(X_k)_{k\in\mathcal{K}}$. Let $f_{\mathcal{I}}$ denote the joint probability mass function of $(X_i)_{i\in\mathcal{I}}$. Let $f_{\mathcal{J}}$ denote the joint probability mass function of $(X_j)_{j\in\mathcal{J}}$. We define the* ***conditional probability mass function*** *of $(X_i)_{i\in\mathcal{I}}$ given $(X_j)_{j\in\mathcal{J}} = (x_j)_{j\in\mathcal{J}}$, denoted by $f_{\mathcal{I}|\mathcal{J}}(\cdot \mid (x_j)_{j\in\mathcal{J}})$, to be a function from $\mathbb{R}^{\mathcal{I}}$ to $\mathbb{R}$ given by*

$$f_{\mathcal{I}|\mathcal{J}}((x_i)_{i\in\mathcal{I}} \mid (x_j)_{j\in\mathcal{J}}) := \frac{f((x_k)_{k\in\mathcal{K}})}{f_{\mathcal{J}}((x_j)_{j\in\mathcal{J}})}.$$

Example 6.2.1. *Let $X_1 \sim Binomial(n_1, p)$ and $X_2 \sim Binomial(n_2, p)$. Suppose that X_1 and X_2 are independent. Then*

$$(X_1 \mid X_1 + X_2 = m) \sim HyperGeo(n_1 + n_2, n_1, m).$$

Proof.

$$\begin{aligned}
\mathbb{P}(X_1 = x \mid X_1 + X_2 = m) &= \frac{\mathbb{P}(X_1 = x \text{ and } X_1 + X_2 = m)}{\mathbb{P}(X_1 + X_2 = m)} \\
&= \frac{\mathbb{P}(X_1 = x \text{ and } X_2 = m - x)}{\mathbb{P}(X_1 + X_2 = m)} \\
&= \frac{\mathbb{P}(X_1 = x)\mathbb{P}(X_2 = m - x)}{\mathbb{P}(X_1 + X_2 = m)} \\
&= \frac{\binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{m-x} p^{m-x} (1-p)^{n_2 - m + x}}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}} \\
&= \frac{\binom{n_1}{x}\binom{n_2}{m-x}}{\binom{n_1+n_2}{m}},
\end{aligned}$$

provided that $x \in [\max\{0, m - n_2\}, \min\{n, m\}]$. So

$$(X_1 \mid X_1 + X_2 = m) \sim HyperGeo(n_1 + n_2, n_1, m).$$

■

Example 6.2.2. *Let $X_i \sim Poisson(\lambda_i)$ for $i \in \{1, ..., n\}$. Suppose that $X_1, ..., X_n$ are independent. Then*

$$(X_j \mid \sum_{i=1}^n X_i = m) \sim Binomial(m, \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}).$$

Proof.

$$\mathbb{P}(X_j = x \mid \sum_{i=1}^n X_i = m) = \frac{\mathbb{P}(X_j = x \text{ and } \sum_{i=1}^n X_i = m)}{\mathbb{P}(\sum_{i=1}^n X_i = m)}$$

$$\begin{aligned}
&= \frac{\mathbb{P}(X_j = x \text{ and } \sum_{i\neq j} X_i = m - x)}{\mathbb{P}(\sum_{i=1}^n X_i = m)} \\
&= \frac{\mathbb{P}(X_j = x)\mathbb{P}(\sum_{i\neq j} X_i = m - x)}{\mathbb{P}(\sum_{i=1}^n X_i = m)} \\
&= \frac{\frac{e^{-\lambda_j}\lambda_j^x}{x!}\frac{e^{-\sum_{i\neq j}\lambda_j}(\sum_{i\neq j}\lambda_i)^{m-x}}{(m-x)!}}{\frac{e^{-\sum_{i=1}^n\lambda_i}(\sum_{i=1}^n\lambda_i)^m}{m!}} \\
&= \binom{m}{x}\frac{\lambda_j^x(\lambda_Y - \lambda_j)^{m-x}}{\lambda_Y^m} \\
&= \binom{m}{x}(\frac{\lambda_j}{\lambda_Y})^x(1 - \frac{\lambda_j}{\lambda_Y})^{m-x},
\end{aligned}$$

provided that $x \in [0, m]$. So

$$(X_j \mid \sum_{i=1}^n X_i = m) \sim Binomial(m, \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}).$$

■

6.3 Conditional Probability Density Function

Definition (Conditional Probability Density Function)**.** *Let X and Y be* <u>continuous</u> *random variables. Let f denote the joint probability density function of X and Y. Let f_Y denote the marginal probability density function of Y. We define the* ***conditional probability density function*** *of X given $Y = y_0$, denoted by $f_{X|Y}(\cdot \mid y_0)$, to be a function given by*

$$f_{X|Y}(x \mid y_0) := \frac{f(x, y_0)}{f_Y(y_0)},$$

provided that $f_Y(y_0) \neq 0$.

6.4 Mixed Conditional Probability Distribution

6.5 Conditional Expectations

Definition (Conditional Expectation)**.** *Let X and Y be random variables. Let g be a function on X. Let y_0 be an arbitrary element in* range(Y)*. Let $f_{X|Y}(\cdot \mid y_0)$ be the conditional probability function of X given $Y = y_0$. Let A be the support set of $f_{X|Y}(\cdot \mid y_0)$. We define the* ***conditional expectation*** *of $g(X)$ given $Y = y_0$ to be a number given by*

$$E\big[g(X) \mid Y = y_0\big] = \begin{cases} \sum_{x\in A} g(x) f_{X|Y}(x \mid y_0), & \textit{if } X \textit{ is discrete} \\ \int_{x\in A} g(x) f_{X|Y}(x \mid y_0)dx, & \textit{if } X \textit{ is continuous.} \end{cases}$$

if $\sum_{x\in A} \big|g(x) f_{X|Y}(x \mid y_0)\big| \neq +\infty$ or $\int_{x\in A} \big|g(x) f_{X|Y}(x \mid y_0)\big|dx \neq +\infty$.

Proposition 6.5.1 (Linearity of the Conditional Expectation Operator)**.** *Let $\mathcal{I}$ be a finite index set. Let $(X_i)_{i\in\mathcal{I}}$ be random variables. Let $(a_i)_{i\in\mathcal{I}}$ be real numbers. Let Y be a random variable. Then*

$$\mathbb{E}[\sum_{i\in\mathcal{I}} a_i X_i \mid Y = y] = \sum_{i\in\mathcal{I}} a_i \mathbb{E}[X_i \mid Y = y].$$

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 $Y = y_0$ to be the number $E\big[X \mid Y = y_0\big]$.*

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 $Y = y_0$, denoted by $\operatorname{var}[X \mid Y = y_0]$, to be the number given by*

$$\operatorname{var}[X \mid Y = y_0] := \mathbb{E}\big[(X - \mathbb{E}[X \mid Y = y_0])^2 \mid Y = y_0\big].$$

6.6 Properties of Conditional Expectations

Proposition 6.6.1. *Let X and Y be two random variables. Then*

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

Proof.

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

■

Proposition 6.6.2 (Substitution Rule)**.**

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

Theorem 1 (Law of Total Expectation)**.** *Let X and Y be random variables. Let g be a function on X.*

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

Proof.

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

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

$\blacksquare$

Proposition 6.6.3 (Law of Total Variance)**.**

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

Proof.

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

That is,

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

$\blacksquare$

6.7 Examples & Exercises

Example 6.7.1 (Random Sum)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be an independent and identically distributed sequence of random variables. Suppose that the X_i's have common mean μ and common standard deviation σ. Let N be a discrete non-negative integer-valued random variable. Suppose N is independent of all the X_i's. Define the* ***random sum*** *T as $T := \sum_{i=1}^{N} X_i$. Then the mean and variance of T are:*

$$\mathbb{E}[T] = \mu\mathbb{E}[N] \quad \text{and} \quad \operatorname{var}[T] = \sigma^2\mathbb{E}[N] + \mu^2 \operatorname{var}[N].$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{E}[\mathbb{E}[T \mid N]], \text{ by the Law of Total Expectation} \\
&= \mathbb{E}[\mathbb{E}[\sum_{i=1}^{N} X_i \mid N]], \text{ by definition of } T \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i \mid N]], \text{ by linearity of the expectation operator} \\
&= \mathbb{E}[\sum_{i=1}^{N} \mathbb{E}[X_i]], \text{ since } X_i \text{ and } N \text{ are independent} \\
&= \mathbb{E}[\sum_{i=1}^{N} \mu], \text{ by definition of } \mu \\
&= \mathbb{E}[N\mu] \\
&= \mu\mathbb{E}[N], \text{ by linearity of the expectation operator.}
\end{aligned}$$

That is,

$$\mathbb{E}[T] = \mu\mathbb{E}[N].$$

We also have

$$\begin{aligned}
\operatorname{var}[T] &= \mathbb{E}[\operatorname{var}[T \mid N]] + \operatorname{var}[\mathbb{E}[T \mid N]], \text{ by the Law of Total Variance} \\
&= \mathbb{E}[\operatorname{var}[T \mid N]] + \operatorname{var}[N\mu] \\
&= \mathbb{E}[\operatorname{var}[T \mid N]] + \mu^2 \operatorname{var}[N] \\
&= \mathbb{E}[\operatorname{var}[\sum_{i=1}^{N} X_i \mid N]] + \mu^2 \operatorname{var}[N], \text{ by definition of } T \\
&= \mathbb{E}[\sum_{i=1}^{N} \operatorname{var}[X_i \mid N]] + \mu^2 \operatorname{var}[N] \\
&= \mathbb{E}[\sum_{i=1}^{N} \operatorname{var}[X_i]] + \mu^2 \operatorname{var}[N], \text{ since } X_i \text{ and } N \text{ are independent} \\
&= \mathbb{E}[\sum_{i=1}^{N} \sigma^2] + \mu^2 \operatorname{var}[N], \text{ by definition of } \sigma \\
&= \mathbb{E}[N\sigma^2] + \mu^2 \operatorname{var}[N] \\
&= \sigma^2\mathbb{E}[N] + \mu^2 \operatorname{var}[N], \text{ by linearity of the expectation operator.}
\end{aligned}$$

That is,

$$\operatorname{var}[T] = \sigma^2\mathbb{E}[N] + \mu^2 \operatorname{var}[N].$$

■

Exercise 6.7.1. *Let X and Y be independent and identically distributed continuous random variables. Then $P(X < Y) = \frac{1}{2}$.*

Proof.

$$\begin{aligned}
P(X < Y) &= \int_{-\infty}^{+\infty} P(X < Y \mid Y = y) f(y) dy && \text{by the Law of Total Probability} \\
&= \int_{-\infty}^{+\infty} P(X < y \mid Y = y) f(y) dy && \text{by the Substitution Rule} \\
&= \int_{-\infty}^{+\infty} P(X < y) f(y) dy && \text{by independence} \\
&= \int_{-\infty}^{+\infty} P(X \leq y) f(y) dy \\
&= \int_{-\infty}^{+\infty} F(y) f(y) dy \\
&= \int_{-\infty}^{+\infty} F(y) F'(y) dy \\
&= \left[\frac{1}{2}(F(y))^2\right]\Big|_{-\infty}^{+\infty} \\
&= \left[\frac{1}{2}1^2\right] - \left[\frac{1}{2}0^2\right] \\
&= \frac{1}{2}.
\end{aligned}$$

That is,

$$P(X < Y) = \frac{1}{2}.$$

■

Exercise 6.7.2. *Let $X \sim Exp(\lambda_1)$ and $Y \sim Exp(\lambda_2)$. Suppose X and Y are independent. Then*

$$P(X < Y) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Proof.

$$\begin{aligned}
P(X < Y) &= \int_0^{+\infty} F_X(y) f_Y(y) dy \\
&= \int_0^{+\infty} (1 - e^{-\lambda_1 y})(\lambda_2 e^{-\lambda_2 y}) dy \\
&= \int_0^{+\infty} \lambda_2 e^{-\lambda_2 y} dy - \int_0^{+\infty} \lambda_2 e^{-(\lambda_1 + \lambda_2) y} dy \\
&= 1 - \int_0^{+\infty} \lambda_2 e^{-(\lambda_1 + \lambda_2) y} dy
\end{aligned}$$

$$\begin{aligned}
&= 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} \int_0^{+\infty} (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2) y} dy \\
&= 1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot 1 \\
&= \frac{\lambda_1}{\lambda_1 + \lambda_2}.
\end{aligned}$$

That is,

$$P(X < Y) = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

■

Chapter 7

Independence

7.1 Independent Events

7.1.1 Definitions

Definition (Independent Events - 1)**.** *Let Ω be a sample space. Let* $\Pr$ *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* $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

Definition (Independent Events - 2)**.** *Let Ω be a sample space. Let* $\Pr$ *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* $\Pr(A \mid B) = \Pr(A)$, *provided that* $\Pr(B) \neq 0$, *or* $\Pr(B \mid A) = \Pr(B)$, *provided that* $\Pr(A) \neq 0$.

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

Proof. Assume without loss of generality that $\Pr(B) \neq 0$. Then

$$\begin{aligned}
& & \Pr(A \cap B) &= \Pr(A)\Pr(B) \\
&\iff & \frac{\Pr(A \cap B)}{\Pr(B)}\Pr(B) &= \Pr(A)\Pr(B) \\
&\iff & \frac{\Pr(A \cap B)}{\Pr(B)} &= \Pr(A) \\
&\iff & \Pr(A \mid B) &= \Pr(A).
\end{aligned}$$

That is,

$$\Pr(A \cap B) = \Pr(A)\Pr(B) \iff \Pr(A \mid B) = \Pr(A).$$

■

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

Definition (Mutually Independent)**.** *Let $\mathcal{A} = \{A_i\}_{i=1}^{n}$ be a finite collection of events where $n \in \mathbb{N}$. We say that the events in $\mathcal{A}$ are **mutually independent** if any event is independent of the intersection of any other events. i.e., $\forall I \subseteq \{1, ..., n\}$, we have $P\big(\bigcap_{i\in I} A_i\big) = \prod_{i\in I} P(A_i)$.*

7.1.2 Properties

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

Proof.

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

■

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

Proof. Let 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. ■

Proposition 7.1.4 (Complements)**.** *If A and B are independent events, then A and B^c are independent.*

Proof.

$$\begin{aligned}
\Pr(A \cap B^c) &= \Pr(A \setminus A \cap B) \\
&= \Pr(A) - \Pr(A \cap B), \text{ by countable additivity of } \Pr \\
&= \Pr(A) - \Pr(A)\Pr(B), \text{ since } A \text{ and } B \text{ are independent} \\
&= \Pr(A)(1 - \Pr(B)) \\
&= \Pr(A)\Pr(B^c).
\end{aligned}$$

That is,

$$\Pr(A \cap B^c) = \Pr(A)\Pr(B^c).$$

So A and B^c are independent. ■

7.2 Independent Random Variables

7.2.1 Definitions

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

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

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

$$f = f_X f_Y.$$

i.e., if

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

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

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

$$F = F_X F_Y.$$

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

$$M = M_X M_Y.$$

Definition (Independence - 5)**.** *Let X and Y be two random variables. Let f_X be the marginal probability function of X. Let f_Y be the marginal probability function of Y. Let $f_X(\cdot \mid y)$ be the conditional probability function of X. Let $f_Y(\cdot \mid x)$ be the conditional probability function of Y. We say that X and Y are **independent** if*

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

Proposition 7.2.1. *The 5 definitions of independence are equivalent.*

7.2.2 Properties

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

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

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

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

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

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

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

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

Then the moment generating function Φ_X of X is

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

Proof.

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

That is,

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

■

7.2.3 Factorization

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

Remark. *Note that this is not the same as an alternative definition of independence. The functions g and h may not be the marginal probability functions.*

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 $f(x, y) = 0 < f_X(x)f_Y(y)$. ■

Chapter 8

Discrete Random Variables

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

8.1 Discrete Uniform Distribution

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

8.2 Bernoulli Distribution

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

Proposition 8.2.1 (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 8.2.2 (Expectation of Bernoulli Distribution)**.**

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

Example 8.2.1. *Flipping a coin once.*

8.3 Binomial Distribution

Definition (Binomial Distribution)**.** *Let $X_i \sim Bernoulli(p)$ 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 8.3.1 (Probability Density Function of Binomial Distribution)**.**

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

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

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

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

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

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

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

■

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

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

Proof Approach 1.

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

■

Proof Approach 2.

$$\mathbb{E}[X] = \Phi'_X(t)|_{t=0}$$

$$
\begin{aligned}
&= \frac{d}{dt}((pe^t) + (1-p))^n|_{t=0} \\
&= n(pe^t + 1 - p)^{n-1} pe^t|_{t=0} \\
&= np.
\end{aligned}
$$

■

Proposition 8.3.4 (Variance of Binomial DIstribution). *Let $X \sim Binomial(n, p)$. Then*

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

Proof Approach 2.

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

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

■

8.4 Negative Binomial Distribution

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

$X :=$ # of 0 outcomes before the $r^{\text{th}}outcomeof1inrepeatedBernoulli(p)experiments$
$X \sim NegBin(r, p).$
$P(X = x) = \binom{x+r-1}{x}(1-p)^x p^{r-1} p.$
$X = \sum_{i=1}^{r} X_i$
$X_i \sim Geo(p).$

8.5 Geometric Distribution

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

Proposition 8.5.1 (Mean of Geometric Distribution)**.** *Let $X \sim Geo(p)$. Then*

$$\mathbb{E}[X] = \frac{1}{p}.$$

Proposition 8.5.2 (Variance of Geometric Distribution)**.** *Let $X \sim Geo(p)$. Then*

$$\text{var}[X] = \frac{1-p}{p^2}.$$

8.6 Hypergeometric Distribution

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

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

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

8.7 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 8.7.1 (Moment Generating Function)**.** *The moment generating function of a $Poisson(\lambda)$ distributed random variable is*

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

Proof.

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

$$\begin{aligned}&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x e^{tx}}{x!}\\&= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}\\&= e^{\lambda(e^t-1)},\end{aligned}$$

for any $t \in \mathbb{R}$. ■

Proposition 8.7.2 (Mean and Variance)**.** *The mean and variance of a Poisson(λ) distributed random variable are*

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

Proof.

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

■

Proposition 8.7.3 (Sum of Independent Poisson Random Variables)**.** *Let $X_i \sim Poisson(\lambda_i)$ for $i \in \{1, ..., n\}$. Suppose that $X_1, ..., X_n$ are independent. Then*

$$\sum_{i=1}^{n} X_i \sim Poisson(\sum_{i=1}^{n}\lambda_i).$$

Proposition 8.7.4. *When n is large and p is small, Poisson(np) can be used bo approximate Binomial(n, p).*

Proof.

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

■

8.8 Shifted Poisson Distribution

Definition (Shifted Poisson Distribution)**.** *Let $X \sim POI(\lambda)$. Let $Y \mid (X = x) \sim BIN(x, p)$ for some $p \in [0, 1]$. We define a* ***shifted Poisson distribution****, y units to the right, to be the distribution of the random variable $X \mid (Y = y)$.*

Proposition 8.8.1 (Probability Mass Function of Shifted Poisson Distribution)**.** *The probability mass function $p_{X|Y}(x \mid y)$ of a shifted Poisson distribution, y units to the right, is given by*

$$p(x \mid y) = \frac{(\lambda(1-p))^{x-y} e^{-\lambda(1-p)}}{(x-y)!}, \text{ for } x = y, y+1, \ldots$$

Proof. We want to calculate the conditional PMF of $X \mid (Y = y)$, to be denoted by $p_{X|Y}(x \mid y)$. Then

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

First of all, note that

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

which implies that

$$\begin{aligned} P(X = x, Y = y) &= P(Y = y \mid X = x) P(X = x) \\ &= \frac{e^{-\lambda}\lambda^x}{x!}\binom{x}{y} p^y (1-p)(x-y). \end{aligned}$$

$$\begin{aligned} P(Y = y) &= \sum_x P(X = x, Y = y) = \sum_{x=y}^{\infty} P(X = x, Y = y) \\ &= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\binom{x}{y} p^y (1-p)^{x-y} \\ &= \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!}\frac{x!}{(x-y)!y!} p^y (1-p)^{x-y} \\ &= \sum_{x=y}^{\infty} e^{-\lambda}\lambda^x \frac{1}{(x-y)!y!} p^y (1-p)^{x-y} \\ &= \frac{e^{-\lambda}p^y}{y!}\sum_{x=y}^{\infty} \lambda^x \frac{(1-p)^{x-y}}{(x-y)!} \\ &= \frac{e^{-\lambda}p^y\lambda^y}{y!}\sum_{x=y}^{\infty} \lambda^{x-y} \frac{(1-p)^{x-y}}{(x-y)!} \\ &= \frac{e^{-\lambda}p^y\lambda^y}{y!}\sum_{x=y}^{\infty} \frac{(\lambda(1-p))^{x-y}}{(x-y)!} \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{-\lambda}p^y\lambda^y}{y!}e^{\lambda(1-p)} \\
&= \frac{e^{-p\lambda}(p\lambda)^y}{y!}.
\end{aligned}
$$

That is,

$$P(Y=y) = \frac{e^{-p\lambda}(p\lambda)^y}{y!}.$$

So in fact, $Y \sim \text{POI}(p\lambda)$. Therefore,

$$
\begin{aligned}
p_{X|Y}(x \mid y) &= \frac{e^{-\lambda}\lambda^x}{x!}\frac{x!}{y!(x-y)!}p^y(1-p)^{x-y} \Big/ \frac{e^{-p\lambda}(p\lambda)^y}{y!} \\
&= e^{-\lambda}\lambda^x \frac{1}{y!(x-y)!}p^y(1-p)^{x-y} \Big/ \frac{e^{-p\lambda}(p\lambda)^y}{y!} \\
&= e^{-\lambda}\lambda^x \frac{1}{(x-y)!}p^y(1-p)^{x-y} \Big/ e^{-p\lambda}(p\lambda)^y \\
&= e^{-\lambda}\lambda^x \frac{1}{(x-y)!}(1-p)^{x-y} \Big/ e^{-p\lambda}\lambda^y \\
&= e^{-\lambda}\lambda^{x-y} \frac{1}{(x-y)!}(1-p)^{x-y} \Big/ e^{-p\lambda} \\
&= \frac{e^{-\lambda(1-p)}(\lambda(1-p))^{x-y}}{(x-y)!}.
\end{aligned}
$$

That is,

$$p_{X|Y}(x \mid y) = \frac{e^{-\lambda(1-p)}(\lambda(1-p))^{x-y}}{(x-y)!}.$$

■

Remark. *The above conditional PMF is recognized as that of a shifted Poisson distribution (y units to the right). Specifically, we have that $(X \mid Y = y) \sim W + y$ where $W \sim$ POI$(\lambda(1-p))$.*

8.9 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.

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

Proposition 8.9.1 (Joint Probability Mass Function)**.**

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

Proposition 8.9.2 (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.*

Proposition 8.9.3 (Marginal Distribution)**.** • $X_i \sim Binomial(n, p_i)$.

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

Proposition 8.9.4 (Conditional Distribution)**.** •

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

for $i \neq j$*.*

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

Proposition 8.9.5. *Let* $T := X_i + X_j$*. Then* $T \sim Binomial(n, p_i + p_j)$*.*

Proof. Idea: use MGF. ■

Proposition 8.9.6. $\mathrm{cov}(X_i, X_j) = -np_i p_j$*.*

Proof.

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

$$
\begin{aligned}
&= \frac{1}{2}\big[-2np_ip_j\big] \\
&= -np_ip_j.
\end{aligned}
$$

■

8.10 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*

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

Chapter 9

Continuous Random Variables

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

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

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

9.1 Continuous Uniform Distribution

9.2 Beta Distribution

9.3 Exponential Distribution

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

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

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

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

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

9.4 Erlang Distribution

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

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

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

9.5 Gamma Distribution

Definition (Gamma Distribution)**.**

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

Proposition 9.5.1 (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 $\alpha, \beta \geq 0$.

Verification of the properties

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

Proposition 9.5.2 (Moment of Gamma Distribution)**.** *Let $X \sim Gamma(\alpha, \beta)$. The the p^{th} moment $\mathbb{E}[X^p]$ of X is*

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

Proof.

$$\mathbb{E}[X^p]$$

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

■

Proposition 9.5.3 (Moment Generating Function of Gamma Distribution)**.**

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

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

Proof.

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

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

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

if $t < \frac{1}{\beta}$. ■

Proposition 9.5.4 (Mean of Gamma Distribution)**.** *Let $X \sim Gamma(\alpha, \beta)$. The the mean $\mathbb{E}[X]$ of X is*

$$\mathbb{E}[X] = \alpha\beta.$$

Proof. 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:

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

■

Proposition 9.5.5 (Variance of Gamma Distribution)**.** *Let $X \sim Gamma(\alpha, \beta)$. The the variance* var$[X]$ *of X is*

$$\text{var}[X] = \alpha\beta^2.$$

Proof.

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

$$\begin{aligned}\text{var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\\ &= \beta^2\alpha(\alpha+1) - (\beta\alpha)^2\\ &= \alpha\beta^2.\end{aligned}$$

■

9.6 Normal Distribution

Probability Density Function

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

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

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

Verification of the properties

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

$$
\begin{aligned}
&= \frac{1}{\sqrt{\pi}} \int_0^{\infty} y^{\frac{1}{2}-1} e^{-y} dy \\
&= \frac{1}{\sqrt{\pi}} \Gamma(\frac{1}{2}) \\
&= \frac{1}{\sqrt{\pi}} \sqrt{\pi} \\
&= 1.
\end{aligned}
$$

Proposition 9.6.1 (Moment Generating Function of Normal Distribution)**.** *Let $X \sim N(\mu, \sigma^2)$. Then*

$$M_Z(t) = e^{t^2/2}.$$

Proof. 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

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

■

9.7 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.

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

9.8 Weibull Distribution

Probability Density Function:

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

for $\alpha, \beta > 0$.

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

Verification of the properties:

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

9.9 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 9.9.1. *If $Z \sim G(0,1)$, then $Z^2 \sim \chi^2(1)$.*

Proposition 9.9.2. *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

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

9.10 t Distribution

Definition

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

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

9.11 Properties

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

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

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

■

Chapter 10

Markov's Inequality

10.1 Statement

Theorem 3 (Markov's Inequality)**.** *Let X be a non-negative random variable. Let $a > 0$. Then*

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

Theorem 4. *Let X be a random variable. Let $a \geq 0$. Let φ be a strictly positive monotonically increasing function on $\mathbb{R}_+$. Then*

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

Corollary. *Let X be a random variable. Let $a \geq 0$. Then for any $n \in \mathbb{N}$,*

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

10.2 Hoeffding's Inequality

Lemma 1 (Hoeffding's Lemma)**.** *Let X be a real random variable such that* $\Pr(X \in [a,b]) = 1$*. Then*

$$\mathbb{E}[\exp(s(X - \mathbb{E}[X]))] \leq \exp(-\frac{1}{8}s^2(b-a)^2).$$

Theorem 5 (Hoeffding's Inequality)**.** *Let $X_1, ..., X_n$ be independent random variables in* $[0,1]$*. Let $\bar{X}$ denote the empirical mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ of the X_i's. Then*

$$\Pr(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2).$$

Chapter 11

Unclassified

Theorem 6. *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

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