1. CONVENTIONS

- $X \sim f(x, \xi)$: a random variables X with the PDF $f(x, \xi)$.

2. GENERAL PROBABILITY THEORY

2.1. Random Variables

Definition 2.1. A **probability space** $(\Omega, \mathfrak{R}, \mathbb{P})$ is a triple of a set Ω and σ-additive measure $\mathbb{P}$ with domain $\mathfrak{R}$, a σ-algebra defined on Ω, satisfying $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\varnothing) = 0$.

A event can be represented as an element in Ω. A type of events can be abstracted as a subset $A \in \Omega$. The measure $\mathbb{P}(A)$ is called the probability of event A happens. If $\mathbb{P}(A) = 1$, we say that A will occurs almost surely.

Definition 2.2. A **random variable** X on $(\Omega, \mathfrak{R}, \mathbb{P})$ is a $\mathfrak{R}$-measurable function $X : \Omega \to \mathbb{R}$. The $\mathfrak{R}$-measurable here means:

$$X^{-1}(B) = \{\omega : \omega \in \Omega, X(\omega) \in B\} \in \mathfrak{R}, \tag{1}$$

where B is any Borel subset of $\mathbb{R}$.

A random variables is a function encoding events into real numbers for mathematical modeling purpose. Furthermore, as a random variable is a mapping from Ω to $\mathbb{R}$, it will natrually induce the following practical conceptation.

Definition 2.3. The **distribution measure** μ_X of X is a pushforward measure induced by X as $\mu_X(B) = X_*\mathbb{P} = \mathbb{P}\{X^{-1}(B)\}$, where B is any Borel subset of $\mathbb{R}$.

The **Radon-Nikodym**'s theorem implies that there exists a non-negative function $f(x)$ bridged the distribution measure μ_X and the natural linear measure of $\mathbb{R}$ as

$$\mu_X(B) = \int_B f(x)dx, \quad \forall B \in \mathbb{R}, \tag{2}$$

where the B is a Borel subset of $\mathbb{R}$. This function $f(X)$ is called the **probability density function** (PDF). A **cumulative distribution function**(CDF) $F(x)$ is defined as $F(x) = \mathbb{P}\{X \leq x\}$. Assum g is a measurable function and $g(x)f(x)$ is integrable, then

$$\int_{\mathbb{R}} g(x)d\mu_X = \int_{X^{-1}(\mathbb{R})} (g \circ X)(\omega)d\mathbb{P} = \int_{\mathbb{R}} g(x)f(x)dx. \tag{3}$$

To simplify, we assume that $X^{-1}(\mathbb{R}) = \Omega$ for any random variable defined on $(\Omega, \mathfrak{R}, \mathbb{P})$. Furthermore, suppose $G(x)$ is a function of X, and $f(x)$ is PDF of X, then

$$\int_\Omega G[X(\omega)]d\mathbb{P} = \int_{\mathbb{R}} G(x)f(x)dx, \tag{4}$$

if $G(x)f(x)$ is integrable over $\mathbb{R}$ respect to the natural linear measure.

Definition 2.4. If a random vairable X is integrable, the **expectation** of X, denoted as $\mathbb{E}(X)$ is defined as

$$\mathbb{E}(X) = \int_\Omega X(\omega)d\mathbb{P}. \tag{5}$$

Based on the Eq.4, the expectation $\mathbb{E}(X)$ can be calculated from

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx.$$

Definition 2.5. A σ-algebra generated by a random variable X, denoted as $\sigma(X)$ is the collection of subsets $X^{-1}(B)$ where B is any Borel subset of $\mathbb{R}$. Since X is required to be $\mathfrak{R}$-measurable by definition, it follows $\sigma(X) \subseteq \mathfrak{R}$. Furthermore, suppose two σ-algebra $\mathfrak{G}, \mathfrak{H} \subseteq \mathfrak{R}$, we called them are **independent** with each other if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \forall A \in \mathfrak{G}, \forall B \in \mathfrak{H}. \tag{6}$$

We say two random varialbes X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent, denoted as $X \perp\!\!\!\perp Y$.

Definition 2.6. A **moment generating function** $M_X(t)$, $t \in \mathbb{R}$ for a random varialbe X is definted as $M_X(t) = \mathbb{E}e^{tX}$.

Theorem 2.1. The following properties of moment generating functions are straight forward:

1. $\mathbb{E}(X^n) = M_X^{(n)}(0)$, nth derivative of $M_X(t)$.
2. If $M_X(t) = M_Y(t)$, then $X = Y$.

2.2. Joint Probabilities and Independence

Definition 2.7. Given two random variables X, Y, the pair (X, Y) forms a mapping $X \times Y : \Omega \to \mathbb{R}^2$, the **joint probability measure** $\mu_{X,Y}$ is defined as a pushforward measure

$$\mu_{X,Y}(A \times B) = \mathbb{P}[(X \times Y)^{-1}(A \times B)], \quad \forall A \times B \in \mathfrak{B}(\mathbb{R}^2), \tag{7}$$

where $\mathfrak{B}(\mathbb{R}^2)$ represent all the Borel subsets of $\mathbb{R}^2$ and

$$(X \times Y)^{-1}(A \times B) = X^{-1}(A) \cap Y^{-1}(B), \tag{8}$$

Theorem 2.2. Suppose X, Y, then the following conditions are equivalent

1. $X \perp\!\!\!\perp Y$;
2. For the joint measure $\mu_{X,Y}(A \times B) = \mu_X(A)\mu_Y(B)$, $\forall A \times B \in \mathfrak{B}(\mathbb{R}^2)$

3. For the PDF $f_X(x)$, $f_Y(y)$ and $f_{X,Y}(x,y)$ or CDF $F_X(x)$, etc:

$$\begin{aligned} f_{X,Y}(a,b) &= f_X(a) f_Y(b), \quad \forall \text{a.e.} (a,b) \in \mathbb{R}^2, \\ F_{X,Y}(a,b) &= F_X(a) F_Y(b), \quad \forall (a,b) \in \mathbb{R}^2; \end{aligned} \tag{9}$$

4. For the joint moment generating function:

$$\mathbb{E} e^{uX+vY} = \mathbb{E} e^{uX} \mathbb{E} e^{vY}; \tag{10}$$

Proof. Assuming the condition satisfied, the 2nd condition comes immediately from the Eq. 6. Consequently, 3rd one holds as

$$F_{X,Y}(a,b) = \mu_{X,Y}([-\infty, a] \times [-\infty, b]). \tag{11}$$

The 2nd condition also implies that Fubini's theorem valid for any $h(x,y)$ integrable function and we have

$$\begin{aligned} \mathbb{E} h(x,y) &= \int_{\mathbb{R}^2} h(x,y) d\mu_{X,Y} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y) f_X(x) dx f_Y(y) dy. \end{aligned} \tag{12}$$

This leads to 4th condition holds. □

2.3. Information and Conditioning

For a given probability space $(\Omega, \mathfrak{R}, \mathbb{P})$, Ω suppose to contained all the possible events occur, and the σ-algebra $\mathfrak{R}$ represents all the possible set to be distinguished, or measured by probability $\mathbb{P}$. The information about the event is ability to label the event with more details. This means that the more information we have, the smaller subset of Ω can be measured. Based on this idea, a σ-algebra $\mathfrak{G} \subseteq \mathfrak{R}$ stays for the limit we can measure under certain information condition.

Definition 2.8. Let $\mathfrak{G}$ be a sub-σ-algebra of $\mathfrak{R}$ in $(\Omega, \mathfrak{R}, \mathbb{P})$ and X is a non-negative or integrable random variable. The **conditoinal expectation** of X given condition $\mathfrak{G}$, denoted as $\mathbb{E}(X|\mathfrak{G})$, is any random varialbe satisfies

1. Measurability: $\mathbb{E}(X|\mathfrak{G})$ is $\mathfrak{G}$-measurable;
2. Partial average:

$$\int_A \mathbb{E}(X|\mathfrak{G})(\omega) d\mathbb{P}(\omega) = \int_A X(\omega) d\mathbb{P}(\omega), \quad \forall A \in \mathfrak{G}. \tag{13}$$

If $\mathfrak{G} = \sigma(W)$, a σ-algebra generated by random variable W, then we denoted $\mathbb{E}(X|W) := \mathbb{E}(X|\sigma(W))$.

The requriements in the definitions preserved the existance and the uniqueness of the $\mathbb{E}(X|\mathfrak{G})$. The Eq. 13 defined a new measure, denoted as $\mu_X|_{\mathfrak{G}}$ on $(\Omega, \mathfrak{G}, \mathbb{P}|_{\mathfrak{G}})$ where $\mathbb{P}|_{\mathfrak{G}}$ is a restrict of $\mathbb{P}$ to $\mathfrak{G}$. Based on the Radon-Nikodym theorem, it implies the existance of $\mathbb{E}(X|\mathfrak{G})$ which equal to the Radon-Nikodym derivative $d\mu_X|_{\mathfrak{G}}/d\mathbb{P}|_{\mathfrak{G}}$. The uniqueness can be varified as follow: Assuming Y, Z are two variables satisfying the Eq. 13, and A is a set that $Y(a) \leq Z(a), \forall a \in A$, then the integral of $Z - Y$ should be non-negative, however,

$$\int_A \{Z(a) - Y(a)\} \, d\mathbb{P} = 0, \quad \forall A \in \mathfrak{G},$$

which implies that $Z = Y$ by mean (in other term, it is called $Z = Y$ almost surely).

Theorem 2.3. Let $(\Omega, \mathfrak{R}, \mathbb{P})$ be a probability space, X be a integrable random variable, and $\mathfrak{G}, \mathfrak{H}$ be sub-σ-algebra.

1. Linearilty: Given integrable random variables X, Y and $a, b \in \mathbb{R}$, then

$$\mathbb{E}(aX + bY|\mathfrak{G}) = a\mathbb{E}(X|\mathfrak{G}) + b\mathbb{E}(Y|\mathfrak{G}). \tag{14}$$

2. If X, Y are integrable, XY is integrable as well, and X is $\mathfrak{G}$-measurable, then

$$\mathbb{E}(XY|\mathfrak{G}) = X\mathbb{E}(Y|\mathfrak{G}). \tag{15}$$

3. Suppose $\mathfrak{H}$ is a σ-algebra that $\mathfrak{H} \subseteq \mathfrak{G}$, then

$$\mathbb{E}\left[(X|\mathfrak{G})|\mathfrak{H}\right] = \mathbb{E}(X|\mathfrak{H}). \tag{16}$$

4. If $\sigma(X) \perp\!\!\!\perp \mathfrak{G}$, then

$$\mathbb{E}(X|\mathfrak{G}) = \mathbb{E}X \tag{17}$$

5. Jensen's inequality: If $\varphi(x)$ is a convex function, then

$$\mathbb{E}[\varphi(X)|\mathfrak{G}] \geq \varphi\left[\mathbb{E}(X|\mathfrak{G})\right]. \tag{18}$$

3. DISTRIBUTIONS

3.1. Poisson Distribution

Definition 3.1 (Poisson assumption)**.** Assume a integer valued random variable K with a PDF $g(k,h)$ where an parameter h satisfying the following assumption when $h \to 0$:

1. $g(1,h) = \lambda h + o(h)$;
2. $\sum_{k=2}^{\infty} g(k,h) = o(h)$;
3. $g(0,h) g(0,w) = g(0, h+w)$;
4. $g(x, w+h) = g(x,w) g(0,h) + g(x-1, w) g(1,h)$.

Then $g(x,w)$ is a **Poisson distribution**:

$$g(x,w) = \frac{1}{x!} (\lambda w)^x e^{-\lambda w}, \quad x = 1, 2, 3, \ldots. \tag{19}$$

Proof. The $o(h)$ means that $\lim_{h\to 0} o(h)/h = 0$.

$$\begin{aligned} g(0, w+h) &= g(0,w)[1-\lambda h - o(h)], \\ \frac{dg(0,w)}{dw} &= \lambda g(0,w), \\ g(0,w) &= ce^{-\lambda w}. \end{aligned}$$

Repeat the similar procedure to Eq.4, we get the formula:

$$\partial_w g(x,w) = -\lambda g(x,w) + \lambda g(x-1,w).$$

Using this equation, the conclusion can be approved by induction. □

Suppose $g(x,h)$ is the probability of x changes in a interval with a width h, if it satisfies the Poisson's assumptions, it means that the event that changing of x depends only on the width of the interval and this probability can be approximated linearly when h is small enough. One example of many applications satisfying the Poisson's assumptions is that the atomic decay with time. In this case, $g(x,h)$ represents the number of decays x happened inside of time interval h.

3.2. Normal distributions

Definition 3.2. A **normal distribution** with mean μ and variance σ^2, denoted as $N(\mu,\sigma^2)$ is

$$N(\mu,\sigma^2)(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{20}$$

If a random variable $X \sim N(\mu,\sigma^2)$, this variable is called **Gaussian**. And we call X as **standard normal** variable if $X \sim N(0,1)$. A random varialbe Y is called n-dimensional Gassian random variable if $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, where $X_i \sim N(\mu_i, \sigma_i^2)$ and $X_i \perp\!\!\!\perp X_j, \forall i \neq j$.

Theorem 3.1. Assume $X_i \sim N(\mu_i,\sigma_i^2)$, $X_i \perp\!\!\!\perp X_j, \forall i \neq j$, $a_i \in \mathbb{R}$, and $Y = \sum_{i=1}^n a_i X_i$, then:

1. The moment generating function of X is $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$
2. The PDF of Y is
$$Y \sim N\left(\sum_{i=1}^n a_i\mu_i, \sum_{i=1}^n a_i^2\sigma_i^2\right). \tag{21}$$
3. If $Z_i = (X_i - \mu_i)/\sigma_i$, then $Z_i \sim N(0,1)$;
4. $Z_i^2 \sim \chi^2(1)$, and if $Z = \sum_{i=1}^n Z_i^2$, then $Z \sim \chi^2(n)$.

Proof. Just briefly draw the line of the proof:

1. It follows from the straight forward calculation of $\mathbb{E}(e^{Xt})$.
2. Consider $n=2$ case, since X_i are independent, the moment generating function is
$$\begin{aligned} M_{a_1X_1+a_2X_2}(t) &= M_{a_1X_1}(t)M_{a_2X_2}(t) \\ &= \exp\left[a_1\mu_1 + a_2\mu_2 + \frac{(a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2}{2}\right], \end{aligned}$$
which is the same as the variable $Z \sim N(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1 + a_2^2\sigma_2)$. On the other hand, if we consider the PDF $f(cx)dx$ with substituting the x by $y = x/c$, the calculation leads to $g(y)dy$ where $g(y) = N(c\mu, c^2\sigma^2)$. It implies that $M_{a_iX_i}(t) = \exp(a_i\mu_i + a_i^2\sigma_i^2/2)$.
3. Here we need to show that $X - c \sim N(\mu - c, \sigma^2)$ if $X \sim N(\mu,\sigma^2)$. In fact, shifting the integral center by a finite number won't affect the integral as the integral range is $[-\infty, +\infty]$.

□

3.3. χ^2istributions

4. MAXIMUM LIKELIHOOD METHODS

4.1. Maximum Likelihood Estimation

Definition 4.1. Consider the case that $X \sim f(x;\theta)$ where θ is a parameter, the n samples $x_i, i = 1, \dots, n$ on X at a fix parameter value, say $\theta = \theta_0$. The **likelihood function** $L(\theta;\boldsymbol{x})$ is defined as

$$L(\theta;\boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta). \tag{22}$$

A **log likelihood function** $l(\theta;\boldsymbol{x})$ is

$$l(\theta;\boldsymbol{x}) = \log L(\theta;\boldsymbol{x}) = \sum_{i=1}^n \log f(x_i;\theta). \tag{23}$$

Definition 4.2. (Regularity Conditions) Given a PDF $f(x;\theta)$ with the set Θ as the domain of θ, the regularity conditions for this PDF are

1. Distinctive: $f(x;\theta) \neq f(x;\theta')$ if $\theta \neq \theta'$;
2. The support of $f(x;\theta)$ independnt on θ;
3. The θ_0 is a interior point of Θ.

INDEX