CONTENTS

1. CONVENTIONS

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

2. GENERAL PROBABILITY THEORY

2.1. Arrangement and Combination

The classical arrangement and combination problems and be classified into the following cases:

1. Given a set containing n distinguishable elements, the number of possible distinguishable results after m sampling over this set is given by the binomial coefficient $\binom{n}{m}$:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}. \tag{1}$$

   This coefficient satisfies the following equations

$$\begin{aligned} &\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}, \\ &\sum_{k=0}^{n} \binom{n}{k} = 2^n. \end{aligned} \tag{2}$$

2. Following the case 1

2.2. 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:

$$X^{-1}(B) = \{\omega : \omega \in \Omega, X(\omega) \in B\} \in \mathfrak{R}, \tag{3}$$

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

$$\mu_X(B) = \int_B f(x)dx, \quad \forall B \in \mathbb{R}, \tag{4}$$

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

$$\int_{\mathbb{R}} g(x)d\mu_X = \int_{X^{-1}(\mathbb{R})} (g \circ X)(\omega)d\mathbb{P} = \int_{\mathbb{R}} g(x)f(x)dx. \tag{5}$$

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

$$\int_\Omega G[X(\omega)]d\mathbb{P} = \int_{\mathbb{R}} G(x)f(x)dx, \tag{6}$$

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

$$\mathbb{E}(X) = \int_\Omega X(\omega)d\mathbb{P}. \tag{7}$$

Based on the Eq.6, the expectation $\mathbb{E}(X)$ can be calculated from

$$\mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx. \tag{8}$$

Furthermore, a variance $\mathrm{Var}X$ is defined as

$$\mathrm{Var}X = \int_\Omega (X - \mathbb{E}X)^2 d\mathbb{P}, \tag{9}$$

in an extreme case, if $\mathrm{Var}X = 0$ implies that X convergence to $\mathbb{E}X$ by mean (another term we say X converges to $\mathbb{E}X$ almost surely).

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

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B), \quad \forall A \in \mathfrak{G}, \forall B \in \mathfrak{H}. \tag{10}$$

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 $X = Y$.

2.3. 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

$$\mu_{X,Y}(A \times B) = \mathbb{P}[(X \times Y)^{-1}(A \times B)], \quad \forall A \times B \in \mathfrak{B}(\mathbb{R}^2), \tag{11}$$

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

$$(X \times Y)^{-1}(A \times B) = X^{-1}(A) \cap Y^{-1}(B), \tag{12}$$

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:
$$\begin{aligned} f_{X,Y}(a, b) &= f_X(a)f_Y(b), \quad \forall \text{a.e.}(a, b) \in \mathbb{R}^2, \\ F_{X,Y}(a, b) &= F_X(a)F_Y(b), \quad \forall (a, b) \in \mathbb{R}^2; \end{aligned} \tag{13}$$
4. For the joint moment generating function:
$$\mathbb{E}e^{uX+vY} = \mathbb{E}e^{uX}\mathbb{E}e^{vY}; \tag{14}$$

Proof. Assuming the condition satisfied, the 2nd condition comes immediately from the Eq. 10. Consequently, 3rd one holds as

$$F_{X,Y}(a, b) = \mu_{X,Y}([-\infty, a] \times [-\infty, b]). \tag{15}$$

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

$$\begin{aligned} \mathbb{E}h(x, y) &= \int_{\mathbb{R}^2} h(x, y)d\mu_{X,Y} \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} h(x, y)f_X(x)dx f_Y(y)dy. \end{aligned} \tag{16}$$

This leads to 4th condition holds. □

Theorem 2.3. Suppose $X \sim f_X(x)$ and $Y \sim f_Y(y)$ are two independent variables, then the PDF of $Z = XY$ is given by the formula

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z/y)f_Y(y)\frac{dy}{|y|}. \tag{17}$$

Proof. To get the PDF of the Z, we first construct the CDF of Z as

$$\begin{aligned} \mathbb{P}(Z \le z) =& \mathbb{P}(XY \le z) \\ =& \mathbb{P}(XY \le z, Y < 0) + \mathbb{P}(XY \le z, Y \ge 0) \\ =& \mathbb{P}(X \ge z/Y, Y < 0) + \mathbb{P}(X \le z/Y, Y \ge 0) \\ =& \left(\int_{-\infty}^{0}\int_{z/y}^{\infty} + \int_0^{\infty}\int_{-\infty}^{z/y}\right)\Big[f_X(x)f_Y(y)dx\Big]dy. \end{aligned} \tag{18}$$

The PDF is obtained from differentiate the formula above with respect to z:

$$\int_0^{\infty} f_X(z/y)f_Y(y)\frac{dy}{y} - \int_{-\infty}^{0} f_X(z/y)f_Y(y)\frac{dy}{y}. \tag{19}$$

□

2.4. 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:

$$\int_A \mathbb{E}(X|\mathfrak{G})(\omega)d\mathbb{P}(\omega) = \int_A X(\omega)d\mathbb{P}(\omega), \quad \forall A \in \mathfrak{G}. \tag{20}$$

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. 20 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. 20, 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 2.4. 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

$$\mathbb{E}(aX + bY|\mathfrak{G}) = a\mathbb{E}(X|\mathfrak{G}) + b\mathbb{E}(Y|\mathfrak{G}). \tag{21}$$

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

$$\mathbb{E}(XY|\mathfrak{G}) = X\mathbb{E}(Y|\mathfrak{G}). \tag{22}$$

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

$$\mathbb{E}\left[(X|\mathfrak{G})|\mathfrak{H}\right] = \mathbb{E}(X|\mathfrak{H}). \tag{23}$$

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

$$\mathbb{E}(X|\mathfrak{G}) = \mathbb{E}X \tag{24}$$

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

$$\mathbb{E}[\varphi(X)|\mathfrak{G}] \geq \varphi\left[\mathbb{E}(X|\mathfrak{G})\right]. \tag{25}$$

Proof. The linearity comes from the linearity properties of Lebesgue integral. For the 2nd point, it is enough to notice that

$$\int_A \mathbb{I}_B \mathbb{E}(Y|\mathfrak{G})(\omega)d\mathbb{P} = \int_{A\cap B} Y(\omega)d\mathbb{P} = \int_A \mathbb{I}_B Y(\omega)d\mathbb{P},$$

where $B \subseteq A \in \mathfrak{G}$. Any integrable X can be approximated by the summation of $\mathbb{I}_B, \forall B \in \mathfrak{G}$ monotonically. It follows that the integral convergs which conclude this theorem. For the 3rd point, it is obvious by expanding two sides for the equation by the definition Eq.20.

To prove the 4th one, suppose $X = \mathbb{I}_B$ where $B \in \sigma(X)$, then

$$\int_A X(\omega)d\mathbb{P} = \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = \int_A \mathbb{E}X d\mathbb{P},$$

which conclude this theorem. The last point comes from the properties of a convex function $\varphi(x)$ that

$$\varphi(tx_1 + (1-t)x_2) \leq t\varphi(x_1) + (1-t)\varphi(x_2), \tag{26}$$

and this definition can be extend to

$$\varphi\left(\sum_{i=1}^n t_i x_i\right) \leq \sum_{i=1}^n t_i \varphi(x_i), \tag{27}$$

since, by induction, we assume the equation valid for $n-1$ case, and for n term case, it can be expended as

$$\begin{aligned}\varphi\left(\sum_{i=1}^n t_i x_i\right) &\leq t_n\varphi(x_n) + \frac{1}{1-t_n}\varphi\left(\sum_{i=1}^{n-1}\frac{t_i}{1-t_n}x_i\right), \\ &\leq \sum_{i=1}^n t_i\varphi(x_i),\end{aligned}$$

as the $\sum_{i=1}^{n-1} t_i/(1-t_n) = 1$. On the other hand, a expectation of X can be approximated by a sequence of simple function like $X_n = \sum x_n \mathbb{P}(A_{x_n})$ and $\sum \mathbb{P}(A_{x_n}) = 1$ by definition, where $A_{x_n} = X^{-1}[x_n, x_n + 1/n]$. These two points lead to this theorem. □

3. DISTRIBUTIONS

3.1. Binomial Distribution

3.2. Hypergeometric Distribution

Definition 3.1. Consider m times sampling from a set with n elements, satisfying the following conditions:

- The sample space can be classified into 2 mutually exclusive groups (no overlapping), and for convenience, we assume the first group is the group of interest.
- No replacement during sampling,

then the random variable X denoting the number of getting k samples from the interest group is obey the hypergeometric distribution Hypergeometric(n,h,m):

$$\Pr(X=k)=\frac{\binom{h}{k}\binom{n-h}{m-k}}{\binom{n}{m}}, \tag{28}$$

where the h is the sample size of the interest group in the sample space.

3.3. Poisson Distribution

Definition 3.2 (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**:

$$g(x,w)=\frac{1}{x!}(\mu)^x e^{-\mu},\quad \mu=\lambda w \text{ and } x=0,1,2,\dots \tag{29}$$

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.

Theorem 3.1. Suppose $X\sim \text{Poisson}(\mu)$, then

1. $\mathbb{E}(X)=\mu$ and $\text{Var}(X)=\mu^2$;
2. The moment generating function
$$M_X(t)=\exp[\mu(e^t-1)]; \tag{30}$$
3. If $Y=\sum_i X_i$ where $X_i\sim\text{Poisson}(\mu_i)$ and $x_i \perp\!\!\!\perp x_j, \forall i\neq j$, then
$$Y\sim \text{Poisson}\left(\sum_i \mu_i\right). \tag{31}$$

Proof. We first prove the second point and the first one follows from the theorem 2.1. In fact

$$\mathbb{E}(e^{xt})=\sum_{x=0}^{\infty} e^{xt}\frac{\mu^x e^{-\mu t}}{x!}=\sum_{x=0}^{\infty}\frac{(\mu e^t)^x e^{-\mu}}{x!}=e^{\mu(e^t-1)}.$$

□

3.4. Normal distributions

Definition 3.3. A **normal distribution** with mean μ and variance σ^2, denoted as $N(\mu,\sigma^2)$ is

$$N(\mu,\sigma^2)(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{32}$$

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 Gaussian 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 3.2. 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
$$Y\sim N\left(\sum_{i=1}^n a_i\mu_i,\sum_{i=1}^n a_i^2\sigma_i^2\right). \tag{33}$$
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 $Y=\sum_{i=1}^n Z_i^2$, then $Y\sim\chi^2(n)$.

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

$$M_{a_1X_1+a_2X_2}(t) = M_{a_1X_1}(t)M_{a_2X_2}(t) = \exp\left[a_1\mu_1 + a_2\mu_2 + \frac{(a_1^2\sigma_1^2 + a_2^2\sigma_2^2)t^2}{2}\right],$$

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 $[-\infty, +\infty]$.

4. We start from calculating $\mathbb{E}(e^{Z^2t})$ for $Z \sim N(0, 1)$, then

$$\mathbb{E}(e^{Z^2t}) = \int_{-\infty}^{\infty} e^{z^2t}e^{-z^2/2}\frac{dz}{\sqrt{2\pi}}, = \int_0^{\infty} e^{xt}\frac{x^{-\frac{1}{2}}e^{-x/2}}{\sqrt{2\pi}}dx,$$

which is the mgf of $\chi^2(1) = \Gamma(1/2, 2)$. For the general case, notice that the volume $V_n(R)$ of a n-dimensional ball is

$$V_n(R) = \frac{\pi^{n/2}R^n}{\Gamma(n/2+1)}, \tag{34}$$

so that the mgf of $Y = \sum_i Z_i^2$ is

$$\begin{aligned}\mathbb{E}Y &= \prod_{i=1}^n \left\{\int \frac{e^{-z_i^2/2+z_i^2t}dz_i}{\sqrt{2\pi}}\right\} \\ &= \int_0^{\infty} \frac{1}{(2\pi)^{n/2}}e^{-r^2/2+r^2t}\frac{\partial V_n(r)}{\partial r}dr, \\ &= \int_0^{\infty} \frac{\rho^{n/2-1}e^{-\rho/2}}{\Gamma(n/2)2^{n/2}}e^{\rho t}d\rho,\end{aligned} \tag{35}$$

where I substituted R by $\rho = R^2$ and the final expression is the mgf of $\Gamma(n/2, 2) = \chi^2(n)$.

□

3.5. Γ and χ^2distributions

Definition 3.4. A Gamma distribution $\Gamma(\alpha, \beta), \alpha, \beta > 0$ is defined as

$$\begin{aligned}\Gamma(\alpha, \beta)(x) &= \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}\Theta(x), \\ \Gamma(\alpha) &= \int_0^{\infty} \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}dx,\end{aligned} \tag{36}$$

where $\Theta(x)$ is a step function $\Theta(x) = 1, \forall x \geq 0$ and $\Theta(x) = 0, \forall x < 0$.

Theorem 3.3. Given a variable $X \sim \Gamma(\alpha, \beta)$, we have

1. The mgf $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \beta^{-1}$;
2. $\mathbb{E}X = \alpha\beta$ and $\mathrm{Var}X = \alpha\beta^2$;
3. Suppose $X_i \sim \Gamma(\alpha_i, \beta)$ and $Y = \sum_i X_i$, then

$$Y \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right). \tag{37}$$

Proof. To validate the 1st equation

$$\begin{aligned}\mathbb{E}e^{Xt} &= \int_0^{\infty} \frac{x^{\alpha-1}e^{(-x/\beta+xt)}}{\Gamma(\alpha)\beta^{\alpha}}dx, \\ &= \int_0^{\infty} \frac{x^{\alpha-1}e^{-x/\beta'}}{\Gamma(\alpha)\beta'^{\alpha}}dx, \\ &= (1 + \beta' t)^{\alpha} \\ &= (1 - t\beta)^{-\alpha}, \quad t\beta < 1,\end{aligned} \tag{38}$$

where $1/\beta' = t + 1/\beta > 0$. Then the expectation and variance follows immediately. The 3rd point is obvious from checking mgf $M_Y(t) = \prod_i M_{X_i}(t)$. □

Definition 3.5. If $X \sim \Gamma(n/2, 2), n \in \mathbb{Z}^+$ (positive integer), we say it is a χ^2-distribution denoted as $X \sim \chi^2(n)$. $\mathbb{E}X = n$, $\mathrm{Var}(X) = 2n$ from the theorem 3.3.

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, a n-triple $\boldsymbol{x} = (x_1, \ldots, x_n)$ represents n samples over X at a fix parameter value $\theta = \theta_0$ (turth). The **likelihood function** $L(\theta; \boldsymbol{x})$ is defined as

$$L(\theta; \boldsymbol{x}) = \prod_{i=1}^n f(x_i; \theta). \tag{39}$$

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

$$l(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x}) = \sum_{i=1}^n \log f(x_i; \theta). \tag{40}$$

An estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ is called the **Maximum Likelihood Estimator** (MLE) of θ if

$$\hat{\theta} = \mathrm{Argmax}\, L(\theta; \boldsymbol{x}), \tag{41}$$

means that $L(\theta; \boldsymbol{x})$ achieves its maximum at $\hat{\theta}$. It implies that MLE satisfies the equation $\partial l(\theta)/\partial\theta = 0$.

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.

Theorem 4.1. Suppose $X \sim f(x,\theta)$ and true parameter θ_0 with samples $\boldsymbol{x}$ satisfy the regularity condition R1 and R2, then

$$\lim_{n\to\infty} P_{\theta_0}[L(\theta_0;\boldsymbol{x}) > L(\theta;\boldsymbol{x})] = 1, \quad \forall\theta \neq \theta_0. \tag{42}$$

Furthermore, if they satisfy all R1 to R3, then the solution $\hat{\theta}$ of the likelihood equation $\partial l(\theta)/\partial\theta = 0$ or $\partial L(\theta)/\partial\theta = 0$ converges to the truth θ_0 in probability.

5. STOCHASTIC PROCESSES

Definition 5.1. A sequence of σ-algebra $\mathfrak{F}(t)$, ordered by a parameter t, is called a **Filtration** if $\mathfrak{F}(s) \subseteq \mathfrak{F}(t), \forall s \leq t$. Given a sequence of random varialbes $X(t)$ indexed by $t \in [0,T]$ is called an **Adapted Stochastic Process** if $X(t)$ is $\mathfrak{F}(t)$-measurable $\forall t \in [0,T]$. A stochastic process $X(t)$ is called **Martingale** if

$$\mathbb{E}[X(t)|\mathfrak{F}(s)] = X(s), \quad \forall s \leq t \in [0,T]. \tag{43}$$

Furthermore, let $f(x), g(x)$ are both Borel-measurable functions, we call $X(t)$ as a **Markov Process** if

$$\mathbb{E}[f(X(t))|\mathfrak{F}(s)] = g(X(s)) \quad \forall s \leq t \in [0,T]. \tag{44}$$

5.1. Brownian Motion

Definition 5.2. Given a space $(\Omega, \mathfrak{R}, \mathbb{P})$ and a given increments $0 = t_0 < t_1 < \cdots < t_n$, a continuous function $W(t), t \in [0,\infty)$, with $W(0) = 0$ is called a **Brownian motion** if it satisfies any one of the follow three condition:

1. $D(t_i) \perp\!\!\!\perp D(t_j), \forall i \neq j$ and $D(t_i) \sim N(0, t_i - t_{i-1})$ where $D(t_i) = W(t_i) - W(t_{i-1})$;
2. $\mathbb{E}\boldsymbol{W} = \boldsymbol{0}$ and $\mathrm{Var}\boldsymbol{W} = \boldsymbol{\Sigma}$, where $\boldsymbol{W}$ is a jointly normal distribution $\boldsymbol{W} = (W_1, \ldots, W_n)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbb{E}[W^2(t_1)], & \mathbb{E}[W(t_1)W(t_2)], & \ldots & \mathbb{E}[W(t_1)W(t_n)] \\ \mathbb{E}[W(t_2)W(t_1)], & \mathbb{E}[W^2(t_2)], & \ldots & \mathbb{E}[W(t_2)W(t_n)] \\ \vdots & \vdots & \ldots & \vdots \\ \mathbb{E}[W(t_n)W(t_1)] & \mathbb{E}[W(t_n)W(t_2)], & \ldots & \mathbb{E}[W^2(t_n)] \end{bmatrix}.$$

$$= \begin{bmatrix} t_1, & t_1, & \ldots & t_1 \\ t_1, & t_2, & \ldots & t_2 \\ \vdots & \vdots & \ldots & \vdots \\ t_1, & t_2, & \ldots & t_n \end{bmatrix} \tag{45}$$

3. The mgf of $\boldsymbol{W}$ is

$$\mathbb{E}\left(\sum_{i=1}^n u_i W_i\right) = \exp\left\{\sum_{k=0}^{n-1}\left(\sum_{j=k+1}^n u_j\right)^2 (t_{k+1} - t_k)\right\}. \tag{46}$$

Proof. We first show that for $s \leq t$:

$$\begin{aligned} \mathbb{E}[W(s)W(t)] &= \mathbb{E}[W(s)(W(t) - W(s))] + \mathbb{E}[W^2(s)] \\ &= \mathbb{E}[W^2(s)] = \mathbb{E}[(W(s) - W(0))^2] \\ &= \mathrm{Var}W^2(s) = s, \end{aligned} \tag{47}$$

where the assumption $D(t) \perp\!\!\!\perp W(s), \forall t \geq s$ used. Furthermore, $\mathbb{E}[W(s)W(t)] = \mathbb{E}[W(t)W(s)] = s, \forall t \geq s$. This proved the Eq.45. The same trick can be used repeatly to derive the mgf above. □

Theorem 5.1. Suppose the sequence of $0 = t_0 \leq t_1 \leq \ldots t_n$ and $W(t_i), t_i \in [0,T]$ is a Brownian motion, $D(t_i) = W(t_i) - W(t_{i-1}), i = 1, 2, \ldots$ and define

$$\int_0^T [dW(t)]^2 = \lim_{\max(dt)\to 0} \sum_{i=1}^n [D(t_i)]^2, \tag{48}$$

where $\max(dt)$ is the maximum of $t_i - t_{i-1}$, for all $i = 1, 2, \ldots$. Then

$$\int_0^T [dW(t)]^2 = T \quad \text{almost surely.} \tag{49}$$

Proof. It is easy to verify the following equation by using the definition $\mathrm{Var}D(t_i) = t_i - t_{i-1}$:

$$\mathbb{E}\sum_{i=0}^N [D(t_i)]^2 \Big|_{t_0=0}^{t_N=T} = T, \tag{50}$$

then the proof left is to show $\mathrm{Var}\left(\int_0^T [dW(t)]^2\right) = 0$. Note that

$$\begin{aligned} &\mathrm{Var}\int_0^T [dW(t)]^2 = \mathbb{E}\left(\int_0^T [dW(t)]^2 - T\right)^2, \\ &= \lim_{\max(dt)\to 0} \sum_{\{t_i\}} \left\{\mathbb{E}[D(t_i)]^4 - 2dt_i \mathbb{E}[D(t_i)]^2 + (dt_i)^2\right\}, \\ &= \lim_{\max(dt)\to 0} \sum_{\{t_i\}} 2(dt_i)^2 = 0. \end{aligned}$$

The last step comes from the fact that $\int_0^T (dt)^2 \leq dT \int_0^T dt \to 0$. The second last step comes from the $\mathbb{E}[D(t_i)]^4 = M_D^{(4)}(0) = 3(t_i - t_{i-1}$, where $M_D(t)$ is the mgf of $D(t_i)$. Again it leads to 0 as requiring $\max(dt) \to 0$. □

The theorem above can be summarized as

$$\begin{aligned} dW(t)dW(t) &= dt + o(dt), \\ dW(t)dt &\sim o(dt), \\ dtdt &\sim o(dt). \end{aligned} \tag{51}$$

It implies that only the first quantities are need to be considered in an integral, the rest can be disregarded.

5.2. Brownian Bridge

Definition 5.3. A **Gaussian process** $X(t), t \geq 0$, is a stochstic process that has the property that, for arbitrary times $0 < t_1 < \cdots < t_n$, the random variables $X(t_i), i = 1, \ldots, n$ are jointly normally distributed.

Definition 5.4. A **Brownian bridge** is a Brownian motion $X(t)$ subject to the condition that $W(0) = X(T) = 0$. The expectation of the bridge at time t is defined as

$$B(t) = \mathbb{E}(X(t)|X(T) = 0), \quad t \in [0, T]. \tag{52}$$

Easy to see, Brownian motion is a Gaussian process. Given a Brownian motion $W(t), t > 0$, then

$$X(t) = W(t) - \frac{t}{T}W(T), \tag{53}$$

is a Brownian bridge. From this construction, we have

$$\mathbb{E}X(t) = \mathbb{E}\left[W(t) - \frac{t}{T}W(T)\right] = 0 \tag{54}$$

Further more, the Brownian bridge $X^{a\to b}(t)$ with a generalized condition $X(0) = a, X(T) = b$ can be constructed as

$$X^{a\to b}(t) = a + (b-a)\frac{t}{T} + X(t), \tag{55}$$

where $X(t)$ is a standard Brownian bridge ($X(0) = X(T) = 0$). In this case, $\mathbb{E}X^{a\to b}(t) = a + (b-a)t/T$.

6. STOCHASTIC CALCULUS

6.1. Itô Integral

1. Itô Integral of Simple Functions

Definition 6.1. Given a probability Brownian motion $W(t)$ with a filter $\mathcal{F}(t)$, and $f(t)$ which is a $\mathcal{F}(t)$-measuable simple function. The **Itô integral** $I(T)$ for the simple function $f(t)$, the values are discrete in the interval $(t_0 = 0, \ldots, t_n = T)$, is defined as

$$I(T) = \int_0^T f(t)dW(t) = \sum_{i=0}^{m-1} f(t_i)[W(t_{i+1}) - W(t_i)]. \tag{56}$$

Theorem 6.1. Itô integral have the following properties

1. **Linearity**: For $c \in \mathbb{R}$ and another $\mathcal{F}(t)$-measurable simple function $g(t)$, $c\int_0^T f(t)dt = \int_0^T cf(t)dt$ and $\int_0^T [f(t) + g(t)]dt = \int_0^T f(t)dt + \int_0^T g(t)dt$;
2. **Adapted**: $I(t)$ is continuous and $\mathcal{F}(t)$-measurable;
3. $I(t)$ is a **martingale**;
4. **Itô isometry**: $\mathbb{E}I^2(t) = \mathbb{E}\int_0^t f^2(u)du$;
5. **Quadratic variation**: $[I, I](t) = \int_0^t f^2(u)du$.

Proof. **Itô isometry:** Expanding the integral, we get

$$\begin{aligned} &\mathbb{E}\left\{\int_0^T f(t)dW(t)\right\}^2 \\ =&\mathbb{E}\left\{\sum_i f(t_i)^2 dW^2(t_i)\right\} + \sum_{i>j} f(t_i)f(t_j)\mathbb{E}[D(t_i)]\mathbb{E}[D(t_j)] \\ =&\mathbb{E}\left\{\sum_i f(t_i)^2 dt_i\right\} = \int_0^T f(t)^2 dt. \end{aligned}$$

The second term above vanishes due to the property 1 in Def. 5.2 and the $dW^2 = dt$ used in the first term. **Conitnuity**: For a given t, we can always find a small enough dt such that

$$I(t + dt) - I(t) = f(t)[dW(t + dt) - dW(t)]$$

and the continuity is the consequence from the fact that $\lim_{dt\to 0} dW(t) \to 0$. □

Theorem 6.2. Let $f(x, t)$ be a function for which the partial derivatives f_t, f_x and f_{xx} are defined and continuous, and let $W(t)$ be a Brownian motion. Then, for every $T \geq 0$:

$$f(T, W(T)) = f(0, W(0)) + \int_0^T f_t(t, W(t))dt + \int_0^T f_x(t, W(t))dW(t) + \frac{1}{2}\int_0^T f_{xx}(t, W(t))dt. \tag{57}$$

Proof. By the Taylor expansion of f,

$$df = \partial_t f dt + \partial_x f dx + \frac{1}{2}\partial_{xx} f (dx)^2 + o(dt, dx), \quad (58)$$

and substitute the x by a function $W(T)$, this variation is valid as well:

$$\begin{aligned} df(t,W) =& f_t(t,W)dt + f_x(t,W)dW \\ &+ \frac{1}{2} f_{xx}(t,W)(dW)^2 + o[dt, (dW)^2], \end{aligned} \quad (59)$$

and we proved $(dW)^2 \to dt$ almost surely, and hence, we proved the Equation 57 hold almost surely. □

The Itô-Doeblin formula can also be used to calcualte the Iô integral.

7. EMPIRICAL STATISTICS

Definition 7.1. A **empirical cumulative distribution function** (ECDF) $F_n(x)$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x}, \quad (60)$$

where $X_i \sim f(x), i = 1, \ldots, n$, are independent, and $\mathbb{I}_{X_i \leq x}$ is a Bernoulli random variable with parameter that $p = F(x)$, where $F(x)$ is the CDF of $f(x)$.

Theorem 7.1. Suppose $F(x)$ is the CDF of random variable $X \sim f(x)$, given the ECDF $F_n(X)$ for $F(x)$, then

- ECDF $F_n(x)$ is a unbiased estimator of $F(x)$.
- $D_n(x) = F_n(x) - F(x)$ is a Brownian bridge.

Proof. The ECDF is a unbiased estimation of $F(x)$ since $\sum_{i=1}^{n} \mathbb{I}_{X \leq x} \sim b(n, p)$, a binomial distribution so that $\mathbb{E}[nF_n(x)] = pn = F(x)n$. Defining a normalized ECDF $D_n(x) = F_n(x) - F(x)$, $\mathbb{E}D_n(x) = 0$. For every CDF $F(x)$, it usually exists a number $x_{\max}$ such that $F(x < x_{\max}) = 1$, then the ECDF $F_n(x)$ of $F(x)$ will have the same property that $F_n(x < x_{\max}) = 1$ which implies that $D_n(x = 0) = D_n(x = x_{\max}) = 0$, which shows that $D_n(x)$ is a Brownian bridge. □

8. STATISTICAL INFERENCE

8.1. Fisher's exact test

Given a sample which can be presented in a 2×2 contingency table, grouped the samples into 2 mutually exclusive classes, say either A/B or α/β:

	A	B	row total
α	a	b	a+b
β	c	d	c+d
Column total	a+c	b+d	n=a+b+c+d

With the hyperthesis that the sampling classification into α or β is independent to the label A or B, the probability of getting this sample result satisfies the hypergeometric distribution and hence given by

$$\Pr = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! \cdot (c+d)! \cdot (a+c)! \cdot (b+d)!}{a! \cdot b! \cdot c! \cdot d! \cdot n!}. \quad (61)$$

To reject this hyperthesis, once should not only calculate the probability of getting this table, but also need to calculate the significance of this samples. Usually, one-tailed or two-tailed tests are needed. For instance, one need to calculate all probability distributions and sum over all the probabilities that less than the probability of getting this contengency table to see where is this table located in a CDF.

8.2. Optimal Test of Hypertheses

A **statistical hypothesis**, is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. For every hypothesis, there's a **null hypothesis**, denoted as H_0, exists corresponding to the statement of reject the hypothesis it associated to.

Definition 8.1. Given a random variable $X \sim f(x; \theta)$, where θ is a unknow parameter defining on a probablity space $(\Omega, \mathfrak{R}, \mathbb{P})$, a **parametric hypothesis test** is a null hypothesis H_0 with an alternative hypothesis H_1

$$H_0 : \theta \in \omega_0 \quad \text{vesus} \quad H_1 : \theta \in \omega_1. \quad (62)$$

A testing for these hypotheses is based on a sample $X_1, \ldots, X_n$ which is iid. to $X \sim f(x; \theta)$. The actual samples are subset of the support of $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then the test rule is

$$\begin{aligned} \boldsymbol{X} \in C &\to \text{Reject } H_1 \\ \boldsymbol{X} \notin C &\to \text{Reject } H_0, \end{aligned} \quad (63)$$

where C is a subset of support of $\boldsymbol{X}$ is called **critical region**.

To evaluate the performance of a hypothesis test, we define the following quantities:

Definition 8.2. A **size** or **significance level** of the test is the probability:

$$\alpha = \max_{\theta \in \omega_0} \mathbb{P}(\boldsymbol{X} \in C). \quad (64)$$

A **power function** of the hypothesis test is

$$\gamma_C(\theta) = \mathbb{P}_{\theta \in \omega_1}(\boldsymbol{X} \in C) \quad (65)$$

Definition 8.3. Suppose a random variable $X \sim f(x; \theta)$, where θ is a unknow parameter defining on a probablity space $(\Omega, \mathfrak{R}, \mathbb{P})$, $C \subset \Omega$ and a null hypothesis $H_0 : \theta \in \omega_0$ with alternative hpothesis $H_1 : \theta \in \omega_1$. Then we say that C is a best **critical region** of size α for testing the hypothesis H_0 against H_1 if

- $\mathbb{P}_{\theta\in\omega_0}[\boldsymbol{X}\in C]=\alpha$
- $\forall A\in\Omega:\mathbb{P}_{\theta\in\omega_0}[\boldsymbol{X}\in A]=\alpha$, then

$$\mathbb{P}_{\theta\in\omega_1}[\boldsymbol{X}\in C]\geq\mathbb{P}_{\theta\in\omega_1}[\boldsymbol{X}\in A] \tag{66}$$

The second condition implies that the C is the region with size α, where optimized the power for testing H_0 against H_1. Furthermore, a test is called **unbiased** if the power of the critical region C is larger than its size

$$\mathbb{P}_{\theta\in\omega_1}[\boldsymbol{X}\in C]\geq\mathbb{P}_{\theta\in\omega_0}[\boldsymbol{X}\in C] \tag{67}$$

INDEX