Review of Probability Theory

Yangang Cao

April 17, 2019

1 Elmemnts of probability

- **Sample space** Ω: The set of all outcomes of a random experiment.
- **Set of events** $\mathcal{F}$: A set whose elements $A \in \mathcal{F}$ are subsets of Ω
- **Probability measure**: A function P: $\mathcal{F} \to \mathbb{R}$ satisfies the following properties,
  - $P(A) \geqslant 0$, for all $A \in \mathcal{F}$
  - $P(\Omega) = 1$
  - if $A_1, A_2, \ldots$ are disjoint events, then

$$P(\cup_i A_i) = \sum_i P(A_i)$$

These three properties are called the **Axioms of Probalitity**

Properties

- If $A \subseteq B \Longrightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \backslash A) = 1 - P(A)$
- (Law of Total Probability) If $A_1, ...A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$

Conditional probability and independence

The conditional probability of any event A given B is defined as

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

Two events are called independent if and only if

$$P(A \cap B) = P(A)P(B) \textit{ or } P(A|B) = P(A)$$

2 Random variables

- Discrete random variable:

$$P(X = k) := P(\{\omega : X(\omega) = k\})$$

- Continuous random variable:

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$$

2.1 Cumulative distribution functions (CDF)

A cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as

$$F_X(x) \triangleq P(X \leqslant x)$$

Properties

- $0 \leqslant F_X(x) \leqslant 1$
- $\lim_{x \to -\infty} F_X(x) = 0$
- $\lim_{x \to \infty} F_X(x) = 1$
- $x \leqslant y \Longrightarrow F_X(x) \leqslant F_X(y)$

2.2 Probability mass functions (PMF)

A probability mass functions (PMF) is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x)$$

Properties

- $0 \leqslant p_X(x) \leqslant 1$
- $\sum_{x \in Val(X)} p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = P(X \in A)$

We use the notation $Val(X)$ for the set of possible values that the random variable X may assume.

2.3 Probability density functions (PDF)

For some continuous random variables, we define the Probability density functions (PDF) as the derivative of the CDF such that

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

According to the properties of differentiation, for very small Δx

$$P(x \leqslant X \leqslant x + \Delta x) \approx f_X(x)\Delta x$$

Properties

- $f_X(x) \geqslant 0$
- $\int_{-\infty}^{\infty} f_X(x) = 1$
- $\int_{x \in A} f_X(x) dx = P(X \in A)$

2.4 Expectation

X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitary function, $g(X)$ can be considered a random variable, we define the expectation of $g(X)$ as

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x)$$

X is a continuous random variable with PDF $f_X(x)$, then

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Properties

- $E[a] = a$ for any constant $a \in \mathbb{R}$
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$
- $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$
- For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$

2.5 Variance

The variance of a random variable X is a measure of how concentrated the distribution of X is around its mean

$$Var[X] \triangleq E\left[(X - E(X))^2\right]$$

An alternate expression for the variance can be derived

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

Properties

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$

2.6 Some common random variables

Discrete random variables

- $X \sim \mathit{Bernoulli}(p)(0 \leqslant p \leqslant 1)$: one if a coin with heads probability p comes up heads, zero otherwise

$$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1 - p & \text{if } p = 0 \end{cases}$$

- $X \sim \mathit{Binomial}(n, p)(0 \leqslant p \leqslant 1)$: the number of heads in n independent flips of a coin with heads probability p

$$p(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1 - p)^{n-x}$$

- $X \sim Geometric(p)(p > 0)$: the number of flips of a coin with heads probability p until the first heads

$$p(x) = p(1-p)^{x-1}$$

- $X \sim Poisson(\lambda)(\lambda > 0)$: a probability distribution over the nonnegative integers used for modeling the frequency of rare events

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Continuous random variables

- $X \sim \mathit{Uniform}(a, b)(a < b)$: equal probability density to every value between a and b on the real line

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim \mathit{Exponential}(\lambda)(\lambda > 0)$: decaying probability density over the nonnegative reals

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim \mathit{Normal}(\mu, \sigma^2)$: also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

3 Two random variables

3.1 Joint and marginal distributions

Suppose that we have two random variables X and Y, A complicated structure known as the joint cumulative distribution function define as

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

The relationship among $F_{XY}(x, y)$, $F_X(x)$ and $F_Y(y)$ are

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y) dy, \quad F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y) dx$$

We call $F_X(x)$ and $F_Y(y)$ the marginal cumulative distribution functions of $F_{XY}(x, y)$

Properties

- $0 \leqslant F_{XY}(x, y) \leqslant 1$
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$

3.2 Joint and marginal probability mass functions

If X and Y are discrete random variables, then the joint probability mass function $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y)$$

and $0 \leqslant P_{XY}(x, y) \leqslant 1$ for all x, y, $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$

We refer to $p_X(x)$ as the marginal probability mass function of X

$$p_X(x) = \sum_y p_{XY}(x, y)$$

and similarly for $p_Y(y)$

3.3 Joint and marginal probability density functions

If X and Y are continuous random variables, then the joint probability density function $f_{XY}(x, y)$ define as

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

Like in the single-dimensional case

$$\iint_{x \in A} f_{XY}(x, y) dxdy = P((X, Y) \in A)$$

Analagous to the discrete case, the marginal probability density function of X is defined as

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

and similarly for $f_Y(y)$

3.4 Conditional distributions

In the discrete case, the conditional probability mass function of X given Y is defined

$$p_{Y|X}(y|x) = \frac{p_{XY}(x, y)}{p_X(x)}$$

assuming that $p_X x \neq 0$

In the continuous case, the conditional probability density of Y given $X = x$ is defined

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

provided $f_X(x) \neq 0$

3.5 Bayes's rule

A useful formula that often arises when trying to derive expression for the conditional probability of one variable given another, is Bayes's rule.

In the case of discrete random variables X and Y

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y' \in Val(Y)} P_{X|Y}(x|y') \, P_Y(y')}$$

In the case of continuous random variables X and Y

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y') \, f_Y(y') \, dy'}$$

3.6 Independence

Two random variables X and Y are independent if
$F_{XY}(x, y) = F_X(x)F_Y(y)$for al values of x and y.
Equivalently

- For discrete random variables, $p_{XY}(x, y) = p_X(x)p_Y(y)$ for all $x \in Val(X)$, $y \in Val(Y)$
- For discrete random variables, $p_{Y|X}(x|y) = p_Y(y)$ whenever $p_X(x) \neq 0$ for all $y \in Val(Y)$
- For continuous random variables, $f_{XY}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$
- For continuous random variables, $f_{Y|X}(y|x) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbb{R}$

3.7 Expectation

Suppose that we have two discrete random variables X, Y and $g : \mathbf{R}^2 \longrightarrow \mathbf{R}$ is a function of these two variables, the expected value of g is defined as

$$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y)$$

For continuous random variables X, Y, the analogous expression is

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dxdy$$

Properties

- $E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]$
- If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$

3.8 Covariance

We can use the concept of expectation to study the relationship of two random variables with each other. The covariance of X and Y is defined as

$$Cov[X, Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

Using an argument similar to that for variance, we can rewrite this as

$$\begin{aligned} Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E|Y| - E[Y]E[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Properties

- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$
- If X and Y are independent, then $Cov[X, Y] = 0$

3.9 Correlation coefficient

The concept of correlation is used to study the **linear** relationship of two random variables with each other. The correlation coefficient of X and Y is defined as

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Properties

- $|\rho_{XY}| \leqslant 1$
- $|\rho_{XY}| = 1 \Leftrightarrow P\{Y = a + bx\} = 1$

4 Multiple random variables

4.1 Basic properties

We can define the joint distribution function of $X_1, X_2, ..., X_n$, the joint probability density function of $X_1, X_2, ..., X_n$, the marginal probability density function of X_1, and the conditional probability density function of X_1 given $X_2, ..., X_n$, as

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = P(X_1 \leqslant x_1, X_2 \leqslant x_2, \ldots, X_n \leqslant x_n)$$

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = \frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{\partial x_1 \ldots \partial x_n}$$

$$f_{X_1}(X_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)\, dx_2 \ldots dx_n$$

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{f_{X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}$$

$$P((x_1, x_2, \ldots x_n) \in A) = \int_{(x_1,x_2,\ldots x_n)\in A} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)\, dx_1 dx_2 \ldots dx_n$$

Chain rule: From the definition of conditional probability for multiple random variables, one can show that

$$
\begin{aligned}
& f(x_1, x_2, \ldots, x_n) = f(x_n | x_1, x_2 \ldots, x_{n-1})\, f(x_1, x_2 \ldots, x_{n-1}) \\
& = f(x_n | x_1, x_2 \ldots, x_{n-1})\, f(x_{n-1} | x_1, x_2 \ldots, x_{n-2})\, f(x_1, x_2 \ldots, x_{n-2}) \\
& = \ldots = f(x_1) \prod_{i=2}^{n} f(x_i | x_1, \ldots, x_{i-1})
\end{aligned}
$$

Independence: For multiple events, $A_1, ..., A_k$, we say that $A_1, ..., A_k$ are mutually independent if for any subset $S \subseteq \{1, 2, \ldots, k\}$, we have

$$
P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i)
$$

Likewise, we say that random variables $X_1, ..., X_n$ are independent if

$$
f(x_1, \ldots, x_n) = f(x_1)\, f(x_2) \cdots f(x_n)
$$

4.2 Random vectors

Suppose that we have n random variables and put then in a vector $X = [X_1 X_2 ... X_n]^T$, we call it random vector, the joint PDF and CDF will apply to random vectors as well.

Expectation: The expected value of an arbitrary function $g : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) \, f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots x_n) \, dx_1 dx_2 \ldots dx_n$$

If g is

$$g(x) = [g_1(x) \quad g_2(x) \quad \ldots \quad g_m(x)]^T$$

then

$$E[g(X)] = [E[g_1(x)] \quad E[g_2(x)] \quad \ldots \quad E[g_m(x)]]^T$$

Covariance matrix: For a given random vector $X : \Omega \to \mathbb{R}^n$, its covariance matrix Σ is the $n \times n$ square matrix whose entries are given by $\Sigma_{ij} = Cov[X_i, X_j]$. We have

$$\begin{aligned}
\Sigma &= \begin{bmatrix} \operatorname{Cov}[X_1, X_1] & \cdots & \operatorname{Cov}[X_1, X_n] \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}[X_n, X_1] & \cdots & \operatorname{cov}[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E\left[X_1^2\right] - E[X_1]\,E[X_1] & \cdots & E[X_1 X_n] - E[X_1]\,E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]\,E[X_1] & \cdots & E\left[X_n^2\right] - E[X_n]\,E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E\left[X_1^2\right] & \cdots & E[X_1 X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] & \cdots & E\left[X_n^2\right] \end{bmatrix} - \begin{bmatrix} E[X_1]\,E[X_1] & \cdots & E[X_1]\,E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]\,E[X_1] & \cdots & E[X_n]\,E[X_n] \end{bmatrix} \\
&= E\left[XX^T\right] - E[X]E[X]^T = \ldots = E\left[(X - E[X])(X - E[X])^T\right]
\end{aligned}$$

4.3 The multivariate Gaussian distribution

A random vector $X \in \mathbb{R}^n$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_{++}$

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$.

In the case $n = 1$, we get the regular definition of a normal distribution with mean parameter μ_1 and variance Σ_{11}

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

5 Law of large numbers and Central limit theorems

5.1 Law of large numbers

Wiener-khinchin law of large numbers : We suppose that random variables $X_1, X_2, ..., X_n$ are independent and identically distributed, and $E[X_k] = \mu (k = 1, 2, ..., n)$, for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left\{ \left| \frac{1}{n}\sum_{k=1}^{n} X_k - \mu \right| < \varepsilon \right\} = 1$$

Bernoulli law of large numbers : We suppose that the incident A occurs f_A times in n times independent replicated experiments, p is the probability of incident A occuring each time, for any $\varepsilon > 0$

$$\lim_{n\to\infty} P\left\{ \left| \frac{f_A}{n} - p \right| < \varepsilon \right\} = 1$$

5.2 Central limit theorems

The central limit theorem of independent and identical distribution: We suppose that random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed, and $E[X_k] = \mu (k = 1, 2, \ldots)$, $Var[X_k] = \sigma^2 > 0 (k = 1, 2, \ldots, n)$, the standard variable of $\sum_{k=1}^{n} X_k$ is

$$Y_n = \frac{\sum_{k=1}^{n} X_k - E\left(\sum_{k=1}^{n} X_k\right)}{\sqrt{Var\left(\sum_{k=1}^{n} X_k\right)}} = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma}$$

the cumulative distribution function $F_n(x)$ to any x satisfies

$$\begin{aligned} \lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} \left\{ \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \leqslant x \right\} \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t = \Phi(x) \end{aligned}$$

Lyapunov theorem: We suppose that random variables $X_1, X_2, ..., X_n$ are independent, and $E[X_k] = \mu(k = 1, 2, ...)$, $Var[X_k] = \sigma^2 > 0$ $(k = 1, 2, ..., n)$, $B_n^2 = \sum_{k=1}^n \sigma_k^2$. When $n \to \infty$, if there is a positive number δ which satisfies

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n E\left\{\left|X_k - \mu_k\right|^{2+\delta}\right\} \to 0$$

then standard variable of $\sum_{k=1}^n X_k$

$$Z_n = \frac{\sum_{k=1}^n X_k - E\left(\sum_{k=1}^n X_k\right)}{\sqrt{D\left(\sum_{k=1}^n X_k\right)}} = \frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n}$$

its cumulative distribution function $F_n(x)$ to any x satisfies

$$\begin{aligned}\lim_{n\to\infty} F_n(x) &= \lim_{n\to\infty} \left\{\frac{\sum_{k=1}^n X_k - \sum_{k=1}^n \mu_k}{B_n} \leqslant x\right\} \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t = \Phi(x)\end{aligned}$$

De Moivre-Laplace theorem: We suppose that random variables $\eta_n(n = 1, 2, ..., n)$ follows binomial distribution which parameters are n and $p(0 < p < 1)$, then any x satisfies

$$\lim_{n\to\infty} P\left\{ \frac{\eta_n - np}{\sqrt{np(1-p)}} \leqslant x \right\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2} \mathrm{d}t = \Phi(x)$$

This theorem indicate that the normal distribution is the limit distribution of the binomial distribution.

6 Moment estimation and Maximum likelihood estimation

Idea of moment estimation

We suppose that X is random variable, θ are parameters to be evaluated, population k-order moment μ_l is

$$\mu_l = E\left(X^l\right) = \int_{-\infty}^{\infty} x^l f(x;\theta_1,\theta_2,\cdots,\theta_k)\,dx \quad \text{for continuous}$$

$$\mu_l = E\left(X^l\right) = \sum_{x\in R_X} x^l p\left(x;\theta_1,\theta_2,\cdots,\theta_k\right) \quad \text{for discrete}$$

sample moment A_l

$$A_l = \frac{1}{n}\sum_{i=1}^{n} X_i^l$$

According to Wiener-khinchin law of large numbers, we have

$$A_l \xrightarrow{P} \mu_l, \quad l = 1, 2, ..., n$$

Method of moment estimation

Generally, μ_l is function of θ_1,θ_2,...,θ_k, we suppose

$$\begin{cases} \mu_1 = \mu_1\left(\theta_1, \theta_2, \cdots, \theta_k\right) \\ \mu_2 = \mu_2\left(\theta_1, \theta_2, \cdots, \theta_k\right) \\ \vdots \\ \mu_k = \mu_k\left(\theta_1, \theta_2, \cdots, \theta_k\right) \end{cases}$$

and solve θ_1,θ_2,...,θ_k

$$\begin{cases} \theta_1 = \theta_1\left(\mu_1, \mu_2, \cdots, \mu_k\right) \\ \theta_2 = \theta_2\left(\mu_1, \mu_2, \cdots, \mu_k\right) \\ \quad\vdots \\ \theta_k = \theta_k\left(\mu_1, \mu_2, \cdots, \mu_k\right) \end{cases}$$

using A_l replace μ_l, we get

$$\hat{\theta}_i = \theta_i\left(A_1, A_2, \cdots, A_l\right), l = 1, 2, ..., k$$

$\hat{\theta}_i$ is called moment estimation of θ_i

Idea of Maximum likelihood estimation

Population X is random variable, θ are parameters to be evaluated and $\theta \in \Theta$. The joint distribution of sample X_1, X_2,...,X_n is

$$\prod_{i=1}^{n} p\left(x_i; \theta\right) \text{ for discrete; } \prod_{i=1}^{n} f(x_i; \theta)\, dx_i \text{ for continuous}$$

and the probability of $\{X_1 = x_1, X_2 = x_2, ..., X_n = x_n\}$ is

$$L(\theta) = L\left(x_1, x_2, \cdots, x_n; \theta\right) = \prod_{i=1}^{n} p\left(x_i; \theta\right) \text{ or } \prod_{i=1}^{n} f(x_i; \theta)$$

$L(\theta)$ is called likelihood function. Naturally, we should find $\hat{\theta}$ satisfy

$$L\left(x_1, x_2, \cdots, x_n; \hat{\theta}\right) = \max_{\theta \in \Theta} L\left(x_1, x_2, \cdots, x_n; \theta\right)$$

$\hat{\theta}$ is called maximum likelihood estimation of θ

Method of Maximum likelihood estimation

Generally, $p(x;\theta)$ and $f(x;\theta)$ are differentiable to θ, so we can solve $\hat{\theta}$ according to

$$\frac{d}{d\theta}L(\theta) = 0$$

Further more, $L(\theta)$ and $lnL(\theta)$ get extreme value at same θ, we often sovle $\hat{\theta}$ according to

$$\frac{d}{d\theta}\ln L(\theta) = 0$$

and following equation is called logarithmic likelihood equation

7 Hypothesis testing about normal distribution

σ^2 **is known, testing about** μ, we use test statistics

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

compare $|z|$ and parameter about rejection region

σ^2 **isn't known, testing about** μ, we use test statistics

$$t = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

compare $|t|$ and parameter about rejection region