概率论知识

复习一下概率论的知识。

随机变量及其分布

离散型随机变量

有些随机变量，它全部可能取到的值是有限多个或可列无限多个，这种随机变量称为**离散型随机变量**。

（一）(0—1) 分布

设随机变量 X 只可能取 0 与 1 两个值，它的分布律是

$$P\{X=k\}=p^k(1-p)^{1-k}, k=0,1 \quad (0<p<1)$$

（二）伯努利试验、二项分布

设试验 E 只有两个可能结果，A 及 $\bar{A}$，则称 E 为**伯努利试验**。设 $P(A)=p(0<p<1)$，此时 $P(\bar{A})=1-p$，将 E 独立重复地进行 n 次，称这重复的独立试验为 n 重伯努利试验。

以 X 表示 n 重伯努利试验中事件 A 发生的次数，X 是一个随机变量，因此事件 A 在指定的 k $(0\le k\le n)$ 次试验中发生，在其他 $n-k$ 次试验中不发生的概率为 $p^k(1-p)^{n-k}$，这种指定的方式共有 $\binom{n}{k}$ 种，它们是两两互不相容的，记 $q=1-p$，则在 n 次试验中 A 发生 k 次的概率为:

$$P\{X=k\}=\binom{n}{k}p^kq^{n-k},\ k=0,1,2,\dots,n$$

称随机变量 X 服从参数为 n,p 的二项分布，记为 $X\sim b(n,p)$

（三）泊松分布

设随机变量 X 所有可能取的值为 $0,1,2,\dots$，而取各个值的概率为

$$P\{X=k\}=\frac{\lambda^k e^{-\lambda}}{k!},\ k=0,1,2,\dots$$

其中 $\lambda>0$ 是常数，则称 X 服从参数为 λ 的**泊松分布**，记为 $X\sim\pi(\lambda)$

泊松定理 设 $\lambda>0$ 是一个常数，n 是任意正整数，设 $np_n=\lambda$，则对于任一固定的非负整数 k，有

$$\lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k}=\frac{\lambda^k e^{-\lambda}}{k!}$$

随机变量的分布函数

设 X 是一个随机变量，x 是任意实数，函数

$$F(x) = P\{X \leq x\}, \ -\infty < x < \infty$$

称为 X 的分布函数。对于任意实数 $x_1, x_2(x_1 < x_2)$，有

$$P\{x_1 < X \leq x_2\} = P\{X \leq x_2\} - P\{X \leq x_1\} = F(x_2) - F(x_1)$$

因此，若已知 X 的分布函数，就可知道 X 落在任一区间 $(x_1, x_2]$ 上的概率。分布函数完整描述了随机变量的统计规律性。

连续型随机变量及其概率密度

如果对于随机变量 X 的分布函数 $F(x)$，存在非负可积函数 $f(x)$，使对于任意实数 x 有

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

则称 X 为连续型随机变量，$f(x)$ 称为 X 的**概率密度函数**，简称**概率密度**。

（一）均匀分布

若连续型随机变量 X 具有概率密度

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{a < x < b,} \\ 0, & \text{其他,} \end{cases}$$

则称 X (a, b) $X \sim U(a, b)$

（二）指数分布

若连续型随机变量 X 的概率密度为

$$f(x) = \begin{cases} \dfrac{1}{\theta} e^{-x/\theta}, & \text{x > 0,} \\ 0, & \text{其他,} \end{cases}$$

其中 $\theta > 0$ 为常数，则称 X 服从参数为 θ 的指数分布。

服从指数分布的随机变量 X 具有一下性质，对于任意 $s, t > 0$，有

$$P\{X > s + t \mid X > s\} = P\{X > t\}$$

这种性质叫做**无记忆性**。

（三）正态分布

若连续型随机变量 X 的概率密度为

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

其中，$\mu,\sigma(\sigma>0)$ 为常数，则称 X 服从参数为 μ,σ 的正态分布或高斯分布，记为 $X\sim N(\mu,\sigma^2)$。特别地，当 $\mu=0,\sigma=1$ 时称随机变量 X 服从**标准正态分布**。

引理 若 $X\sim N(\mu,\sigma^2)$，则 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$。

多维随机变量及其分布

二维随机变量

设 E 是一个随机试验，样本空间是 $S=\{e\}$，设 $X=X(e)$ 和 $Y=Y(e)$ 是定义在 S 上的随机变量，由它们构成的一个向量 (X,Y) 叫做**二维随机变量**。

定义 设 (X,Y) 是二维随机变量，对于任意实数 x,y，二元函数：

$$F(x,y)=P\{(X\le x)\cap(Y\le y)\}=P\{X\le x,Y\le y\}$$

称为二维随机变量 (X,Y) 的分布函数，或称随机变量 X 和 Y 的联合分布函数。

如果二维随机变量 (X,Y) 全部可能取到的值是有限对或可列无限多对，则称 (X,Y) 是**离散型**的随机变量。对于二维随机变量 (X,Y) 的分布函数 $F(x,y)$，如果存在非负可积函数 $f(x,y)$ 使对于任意 x,y 有

$$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x}f(u,v)dudv$$

则称 (X,Y) 是连续型的二维随机变量，函数 $f(x,y)$ 称为二维随机变量 (X,Y) 的**概率密度**，或称为随机变量 X 和 Y 的联合概率密度。

边缘分布

二维随机变量 (X,Y) 作为一个整体，具有分布函数 $F(x,y)$，而 X 和 Y 都是随机变量，各自也有分布函数，记为 $F_X(x)$，$F_Y(y)$，依次称为二维随机变量 (X,Y) 关于 X 和 关于 Y 的**边缘分布函数**。

$$F_X(x)=P\{X\le x\}=P\{X\le x,Y<\infty\}=F(x,\infty)$$

同样有，$F_Y(y)=F(\infty,y)$

对于离散型随机变量，可得

$$F_X(x)=F(x,\infty)=\sum_{x_i\le x}\sum_{j=1}^{\infty}p_{ij}$$

所以，X 的分布律为

$$P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij},\ i=1,2,\dots$$

同样，Y 的分布律为

$$P\{Y=y_i\}=\sum_{i=1}^{\infty}p_{ij},\ j=1,2,\dots$$

上式分别是 (X,Y) 关于 X 和关于 Y 的**边缘分布律**。

对于连续型随机变量 (X,Y)，设它的概率密度为 $f(x,y)$，由于

$$F_X(x)=F(x,\infty)=\int_{-\infty}^{x}\left[\int_{-\infty}^{\infty}f(x,y)dy\right]dx$$

可知，X 是一个连续型随机变量，且其概率密度为

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

同样，Y 也是一个连续型随机变量，其概率密度为

$$f_Y(y)=\int_{-\infty}^{\infty}f(x,y)dx$$

分别称 $f_X(x),f_Y(y)$ 为 (X,Y) 关于 X 和关于 Y 的边缘概率密度。

要点：已知关于 X 和关于 Y 的边缘分布，一般来说是不能确定随机变量 X 和 Y 的联合分布的。

条件分布

设 (X,Y) 是二维离散型随机变量，其分布律为

$$P\{X=x_i,Y=y_j\}=p_{ij},\ i,j=1,2,\dots$$

(X,Y) 关于 X 和关于 Y 的边缘分布律分别为：

$$P\{X=x_i\}=p_{i\cdot}=\sum_{j=1}^{\infty}p_{ij},\ i=1,2,\dots,$$

$$P\{Y=y_i\}=p_{\cdot j}=\sum_{i=1}^{\infty}p_{ij},\ j=1,2,\dots,$$

设 $p_{\cdot j}>0$，考虑在事件 $\{Y=y_j\}$ 已发生的条件下事件 $\{X=x_i\}$ 发生的概率，由条件概率公式，可得

$$P\{X=x_i|Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}},\ i=1,2,\ldots,$$

上述条件概率具有分布律的性质：

1. $P\{X=x_i|Y=y_j\}\geq 0;$
2. $\sum_{i=1}^{\infty}P\{X=x_i|Y=y_j\}=1$

于是，引入以下的定义：

定义 设 (X,Y) 是二维离散型随机变量，对于固定的 j，若 $P\{Y=y_j\}>0$，则称

$$P\{X=x_i|Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}},\ i=1,2,\cdots$$

为在 $Y=y_j$ 条件下随机变量 X 的条件分布律。

定义 设二维随机变量 (X,Y) 的概率密度为 $f(x,y)$，(X,Y) 关于 Y 的边缘概率密度为 $f_Y(y)$。若对于固定的 y，$f_Y(y)>0$，则称 $\frac{f(x,y)}{f_Y(y)}$ 为在 $Y=y$ 条件下 X 的条件概率密度，记为

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

类似可定义

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

相互独立的随机变量

定义 设 $F(x,y)$ 及 $F_X(x),F_Y(y)$ 分别是二维随机变量 (X,Y) 的分布函数及边缘分布函数。若对于所有 x,y 有

$$P\{X\leq x,Y\leq y\}=P\{X\leq x\}P\{Y\leq y\}$$

即 $F(x,y)=F_X(x)F_Y(y)$，则称随机变量 X 和 Y 是相互独立的。

设 (X,Y) 是连续型随机变量，$f(x,y)$，$f_X(x)$，$f_Y(y)$ 分别为 (X,Y) 的概率密度和边缘概率密度，则 X 和 Y 相互独立的条件等价为：等式

$$f(x,y)=f_X(x)f_Y(y)$$

在平面上几乎处处成立。

当 (X,Y) 是离散型随机变量时，X 和 Y 相互独立的条件等价于：对于 (X,Y) 的所有可能取的值 (x_i,y_j)，有

$$P\{X=x_i,Y=y_j\}=P\{X=x_i\}P\{Y=y_j\}$$

随机变量的数字特征

主要包括数学期望、方差、相关系数和矩。

数学期望

定义 设离散型随机变量 X 的分布律为 $P\{X=x_k\}=p_k$，$k=1,2,\ldots$，若级数 $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛，则称级数 $\sum_{k=1}^{\infty} x_k p_k$ 的和为随机变量 X 的数学期望，记为 $E(X)$，即

$$E(x)=\sum_{k=1}^{\infty} x_k p_k$$

设连续型随机变量 X 的概率密度为 $f(x)$，若积分

$$\int_{-\infty}^{\infty} xf(x)dx$$

绝对收敛，则称积分 $\int_{-\infty}^{\infty} xf(x)dx$ 的值为随机变量 X 的数学期望，记为 $E(x)$，即

$$E(x)=\int_{-\infty}^{\infty} xf(x)dx$$

数学期望简称**期望**，又称为**均值**。

数学期望的几个性质：

1. 设 C 是常数，则有 $E(C)=C$。
2. 设 X 是一个随机变量，C 是常数，则有 $E(CX)=CE(X)$。
3. 设 X, Y 是两个随机变量，则有 $E(X+Y)=E(X)+E(Y)$。
4. 设 X, Y 是相互独立的随机变量，则有 $E(XY)=E(X)E(Y)$。

方差

定义 设 X 是一个随机变量，若 $E\{[X-E(X)]^2\}$ 存在，则称 $E\{[X-E(X)]^2\}$ 为 X 的**方差**，记为 $D(X)$ 或 $Var(X)$，即

$$D(X)=Var(X)=E\{[X-E(X)]^2\}$$

在应用上还引入量 $\sqrt{D(X)}$，记为 $\sigma(X)$，记为**标准差**或**均方差**。

有定义知，方差实际上是随机变量 X 的函数 $g(X)=[X-E(X)]^2$ 的数学期望，所以对离散型随机变量有

$$D(X)=\sum_{k=1}^{\infty}[x_k-E(X)]^2 p_k$$

其中 $P\{X=x_k\}=p_k, k=1,2,\ldots$ 是 X 的分布律。

对于连续型随机变量，有

$$D(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f(x) dx$$

其中 $f(x)$ 是 X 的概率密度。

方差可按公式计算：$D(X) = E(X^2) - [E(X)]^2$ 方差的几个性质：

1. 设 C 是常数，则 $D(C) = 0$。
2. 设 X 是随机变量，C 是常数，则有 $D(CX) = C^2 D(X)$，$D(X + C) = D(X)$。
3. 设 X, Y 是两个随机变量，则有 $D(X + Y) = D(X) + D(Y) - 2E\{(X - E(X))(Y - E(Y))\}$，特别，若 X, Y 相互独立，则有 $D(X + Y) = D(X) + D(Y)$。
4. $D(X) = 0$ 的充要条件是 X 以概率 1 取常数 $E(X)$，即 $P\{X = E(X)\} = 1$

证明 3. $D(X + Y) = E\{[(X + Y) - E(X + Y)]^2\}$

$= E\{[(X - E(X)) + (Y - E(Y))]^2\}$

$= E\{(X - E(x))^2\} + E\{(Y - E(Y))^2\} + 2E\{[X - E(X)][Y - E(Y)]\}$

$= D(X) + D(Y) + 2E\{[X - E(X)][Y - E(Y)]\}$

上式右端第三项：$2E\{[X - E(X)][Y - E(Y)]\}$

$= 2E\{XY - XE(Y) - YE(X) + E(X)E(Y)\}$

$= 2\{E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)\}$

$= 2\{E(XY) - E(X)E(Y)\}$

若 X, Y 相互独立，则上式右端为 0，于是 $D(X + Y) = D(X) + D(Y)$。

切比雪夫(Chebyshev)不等式：设随机变量 X 具有数学期望 $E(X) = \mu$，方差 $D(X) = \sigma^2$，则对于任意整数 ϵ，不等式

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$$

成立。

协方差

定义 量 $E\{[X - E(X)][Y - E(Y)]\}$ 称为随机变量 X 与 Y 的协方差，记为 $Cov(X, Y)$，即

$$Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

而 $\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$ 称为随机变量 X 与 Y 的**相关系数**。

由定义，可知

$$Cov(X, Y) = Cov(Y, X),\ Cov(X, X) = D(X)$$

对于任意两个随机变量 X 和 Y，下列等式成立：

$$D(X+Y)=D(X)+D(Y)+2Cov(X,Y)$$

将 $Cov(X,Y)$ 的定义式展开，得

$$Cov(X,Y)=E(XY)-E(X)E(Y)$$

常利用这一式子计算协方差。

协方差具有下述性质：

1. $Cov(aX,bY)=abCov(X,Y), a,b$ 为常数。
2. $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$

ρ_{XY} 的两条重要性质：

1. $|\rho_{XY}|\le 1$,
2. $|\rho_{XY}|=1$ 的充要条件是，存在常数 a,b 使 $P\{Y=a+bX\}=1$

当 $\rho_{XY}=0$ 时，称 X 和 Y 不相关。

矩、协方差矩阵

设 (X,Y) 是二维随机变量

若 $E(X_k),\ k=1,2,\ldots$ 存在，称它为 X 的 k **阶原点矩**，简称 k **阶矩**。

若 $E\{[X-E(X)]^k\},\ k=2,3,\ldots$ 存在，称它为 X 的 k **阶中心距**。

若 $E(X^kY^l),\ k,l=1,2,\ldots$ 存在，称它为 X 和 Y 的 $k+l$ **阶混合矩**。

若 $E\{[X-E(X)]^k[Y-E(Y)]^l\},\ k,l=1,2,\ldots$ 存在，称它为 X 和 Y 的 $k+l$ **阶混合中心矩**

显然，X 的数学期望 $E(X)$ 是 X 的一阶原点矩，方差 $D(X)$ 是 X 的二阶中心矩，协方差 $Cov(X,Y)$ 是 X 和 Y 的二阶混合中心矩。

二维随机变量 (X_1,X_2) 有四个二阶中心矩（设它们都存在），分别记为

$$c_{11}=E\{[X_1-E(X_1)]^2\},$$

$$c_{12}=E\{[X_1-E(X_1)][X_2-E(X_2)]\},\text{（二阶混合中心矩）}$$

$$c_{21}=E\{[X_2-E(X_2)][X_1-E(X_1)]\},$$

$$c_{22}=E\{[X_2-E(X_2)]^2\},$$

将它们排成矩阵的形式：

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

这个矩阵称为随机变量 (X_1, X_2) 的协方差矩阵，由于 $c_{12} = c_{21}$，上述矩阵是个对称矩阵。可推广到 n 维随机变量 $(X_1, X_2, \ldots, X_n)$，若二阶混合中心矩
$c_{ij} = Cov(X_i, X_j) = E\{[X_i - E(X_i)][X_j - E(X_j)]\}$，$i, j = 1, 2, \ldots, n$ 都存在，则称矩阵

$$\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

为 n 阶随机变量 $(X_1, X_2, \ldots, X_n)$ 的协方差矩阵（同样为对称矩阵）。

大数定律及中心极限定理

大数定律是叙述随机变量序列的前一些项的算术平均值在某种条件下收敛到这些项的均值的算术平均值；中心极限定理则是确定在什么条件下，大量随机变量之和的分布逼近于正态分布。

大数定律

弱大数定理（辛钦大数定理）设 $X_1, X_2, \cdots$ 是相互独立，服从同一分布的随机变量序列，且具有数学期望 $E(X_k) = \mu \ (k = 1, 2, \cdots)$。作前 n 个变量的算术平均 $\frac{1}{n}\sum_{k=1}^{n} X_k$，则对于任意 $\epsilon > 0$，有

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

辛钦大数定理又可叙述为：

弱大数定理（辛钦大数定理）设随机变量 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立，服从同一分布且具有数学期望 $E(X_k) = \mu \ (k = 1, 2, \cdots)$，则序列 $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$ 依概率收敛于 μ，即 $\bar{X} \xrightarrow{P} \mu$。

伯努利大数定理 设 f_A 是 n 次独立重复试验中事件 A 发生的次数，p 是事件 A 在每次试验中发生的概率，则对于任意正数 $\epsilon > 0$，有

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

或

$$\lim_{n\to\infty} \left\{\left|\frac{f_A}{n} - p\right| \geq \epsilon\right\} = 0$$

中心极限定理（略）

参数估计

点估计

设总体 X 的分布函数的形式已知，但它的一个或多个参数未知，借助于总体 X 的一个样本来估计总体未知参数的值的问题称为参数的点估计问题。

点估计问题的一般提法如下：设总体 X 的分布函数 $F(x;\theta)$ 的形式已知，θ 是待估参数。$X_1,X_2,\cdots,X_n$ 是 X 的一个样本，$x_1,x_2,\cdots,x_n$ 是相应的一个样本值。点估计问题就是要构造一个适当的统计量 $\widehat{\theta}(X_1,X_2,\cdots,X_n)$，用它的观察值 $\widehat{\theta}(x_1,x_2,\cdots,x_n)$ 作为未知参数 θ 的近似值。称 $\widehat{\theta}(X_1,X_2,\cdots,X_n)$ 为 θ 的估计量，称 $\widehat{\theta}(x_1,x_2,\cdots,x_n)$ 为 θ 的估计值。

有两种常用的构造估计量的方法：矩估计法和最大似然估计法。

（一）矩估计法

设 X 为连续型随机变量，其概率密度为 $f(x;\theta_1,\theta_2,\cdots,\theta_k)$，或 X 为离散型随机变量，其分布律为 $P\{X=x\}=p(x;\theta_1,\theta_2,\cdots,\theta_k)$，其中 $\theta_1,\theta_2,\cdots,\theta_k$ 为待估参数，$X_1,X_2,\cdots,X_n$ 是来自 X 的样本。假设总体 X 的前 k 阶矩

$$\mu_l=E(X^l)=\int_{-\infty}^{\infty}x^l f(x;\theta_1,\theta_2,\cdots,\theta_k)dx\ (X\text{ 连续型})\ l=1,2,\cdots,k$$

或

$$\mu_l=E(X^l)=\sum_{x\in R_X}x^l p(x;\theta_1,\theta_2,\cdots,\theta_k)\ (X\text{ 离散型})\ l=1,2,\cdots,k$$

存在。一般来说，它们是 $\theta_1,\theta_2,\cdots,\theta_k$ 的函数。基于样本矩

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

依概率收敛于相应的总体矩 $\mu_l(l=1,2,\cdots,k)$，样本矩的连续函数依概率收敛于相应的总体矩的连续函数，我们就用样本矩作为相应的总体矩的估计量，而以样本矩的连续函数作为相应的总体矩的连续函数的估计量。这种估计方法称为**矩估计法**。矩估计法的具体做法如下：设

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

这是一个包含 k 个未知参数 $\theta_1,\theta_2,\cdots,\theta_k$ 的联立方程组。一般来说，可以从中解出 $\theta_1,\theta_2,\cdots,\theta_k$，得到

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

以 A_i 分别代替上式中的 $\mu_i, i = 1, 2, \cdots, k$，就以

$$\widehat{\theta_i} = \theta_i(A_1, A_2, \cdots, A_k),\ i = 1, 2, \cdots, k$$

分别作为 $\theta_i, i = 1, 2, \cdots, k$ 的估计量，这种估计量称为**矩估计量**，矩估计量的观察值称为**矩估计值**。

（二）最大似然估计法

若总体 X 属离散型，其分布律 $P\{X = x\} = p(x; \theta), \theta \in \Theta$ 的形式为已知，θ 为待估参数，Θ 是 θ 可能取值的范围，设 $X_1, X_2, \cdots, X_n$ 是来自 X 的样本，则 $X_1, X_2, \cdots, X_n$ 的联合分布律为：

$$\prod_{i=1}^{n} p(x_i; \theta).$$

又设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的一个样本值，易知样本 $X_1, X_2, \cdots, X_n$ 取到观察值 $x_1, x_2, \cdots, x_n$ 的概率，亦即事件 $\{X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n\}$ 发生的概率为：

$$L(\theta) = L(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta), \theta \in \Theta$$

这一概率随 θ 的取值而变化，它是 θ 的函数，$L(\theta)$ 称为样本的**似然函数**。固定样本观察值 $x_1, x_2, \cdots, x_n$，在 θ 取值的可能范围 Θ 内挑选使似然函数达到最大的参数值 $\hat{\theta}$，作为参数 θ 的估计值，即取 $\hat{\theta}$ 使

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

这样得到的 $\hat{\theta}$ 与样本值 $x_1, x_2, \cdots, x_n$ 有关，常记为 $\hat{\theta}(x_1, x_2, \cdots, x_n)$，称为参数 θ 的**最大似然估计值**，而相应的统计量 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 称为参数 θ 的**最大似然估计量**。

若总体 X 属连续型，其概率密度 $f(x; \theta), \theta \in \Theta$ 的形式已知，θ 为待估参数，Θ 是 θ 可能取值的范围。设 $X_1, X_2, \cdots, X_n$ 是来自 X 的样本，则 $X_1, X_2, \cdots, X_n$ 的联合密度为：

$$\prod_{i=1}^{n} f(x_i; \theta).$$

设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的一个样本值，则随机点 $(X_1, X_2, \cdots, X_n)$ 落在点 $(x_1, x_2, \cdots, x_n)$ 的邻域（边长分别为 $dx_1, dx_2, \cdots, dx_n$ 的 n 维立方体）内的概率近似地为：

$$\prod_{i=1}^{n} f(x_i; \theta) dx_i$$

其值随 θ 的取值而变化。与离散型的情况一样，取 θ 的估计值 $\hat{\theta}$ 使上式概率取到最大值，但因子 $\prod_{i=1}^{n} dx_i$ 不随 θ 而变，故只需考虑函数

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

的最大值，这里 $L(\theta)$ 称为样本的**似然函数**，若

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

则称 $\hat{\theta}(x_1, x_2, \cdots, x_n)$ 为 θ 的**最大似然估计值**，称 $\hat{\theta}(X_1, X_2, \cdots, X_n)$ 为 θ 的最大似然估计量。

在很多情形下，$p(x; \theta)$ 和 $f(x; \theta)$ 关于 θ 可微，这时 $\hat{\theta}$ 可从方程

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

解的。又因 $L(\theta)$ 与 $lnL(\theta)$ 在同一处取到极值，因此，θ 的最大似然估计也可从方程

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

求得，后一方程求解比较方便，称为**对数似然方程**。