第 3 章

概率统计基础

概率统计知识在人工智能领域发挥着非常重要的作用，如深度学习理论、概率图模型等都依赖于概率分布作为框架的基本建模语言，与此同时，复杂的深度学习等算法又超越了传统统计分析方法，在人工智能与大数据领域发挥了不可替代的作用。本章将从概率论与数理统计的基本概念和原理出发，重点介绍与人工智能有关的数理统计方法，并以 Python 语言为例设计对应的实验。

3.1 概率论

3.1.1 概率与条件概率

在我们的日常生活和工作中，经常会遇到许多包含不确定结果的现象，如股票价格是涨还是跌呢？已经布满乌云的天空是否会下雨呢？某项工程是否能够按期完成呢？投资项目盈利可能性有多大呢？上述现象都有两个特点：①事先不能确定哪一个结果会出现；②各种结果在多次重复过程中可能会体现某种规律。

人们把这类现象称为随机现象或者不确定现象，同时使用一个数值来度量随机现象中某一结果出现可能性的大小，这个数值就被称为概率(Probability)。概率的取值在 0 到 1 之间，概率值等于 0 表明该现象不可能发生，概率值等于 1 表明该现象必然发生，介于 0 和 1 之间的概率则说明该现象出现可能性的不同程度。

随机试验是对随机现象的观察、记录、实验的统称，是在相同条件下对某随机现象进行的大量重复观测，具有以下特点：①在试验前不能断定其将发生什么结果，但可明确指出或说明试验的全部可能结果是什么；②在相同的条件下试验可大量地重复；③重复试验的结果是以随机方式或偶然方式出现。

人们把随机试验 E 的所有可能结果组成的集合称为 E 的样本空间，记为 S。样本空间的元素即 E 的每个结果称为样本点。

假设在相同条件下，进行了 n 次试验，在这 n 次试验中，人们把事件 A 发生的次数称为事件 A 发生的频数，A 发生的次数与试验次数的比值称为事件 A 发生的频率。

当试验次数增加时，随机事件 A 发生的频率趋于一个稳定值，记为 p，p 就称为该事件发生的概率，记为 $P(A)=p$。

条件概率(Conditional Probability)是一种带有附加条件的概率，例如，如果事件 A 与事件 B 是相依事件，即事件 A 的概率随事件 B 是否发生而变化，记为 $P(A|B)$，表示在事件

B发生的条件下，事件A发生的概率，相当于A在B中所占的比例。

【例3-1】 一个家庭中有两个小孩，已知至少一个是女孩，问两个都是女孩的概率是多少（假定生男生女的可能性是相等的）？

解：由题意，样本空间为

$$S=\{(兄,弟),(兄,妹),(姐,弟),(姐,妹)\}$$
$$B=\{(兄,妹),(姐,弟),(姐,妹)\}$$
$$A=\{(姐,妹)\}$$

由于事件B已经发生，所以这时试验的所有可能只有3种，而事件A包含的基本事件只占其中的一种，所以有 $P(A|B)=1/3$，即在已知至少一个是女孩的情况下，两个都是女孩的概率为1/3。

在这个例子中，如果不知道事件B发生，则事件A发生的概率为 $P(A)=1/4$。这里 $P(A)\neq P(A|B)$，其原因在于事件B的发生改变了样本空间，使它由原来的S缩减为新的样本空间 $S_B=B$。

3.1.2 随机变量

随机变量（Random Variable）表示随机试验各种结果的实值单值函数，即能用数学分析方法来研究随机现象。例如某一时间内公共汽车站等车的乘客人数、淘宝在一定时间内的交易次数、百度在一段时间内某个关键词搜索次数等，都是随机变量的实例。这些例子中所提到的量，尽管它们的具体内容各式各样，但从数学观点来看，它们表现了同一种情况，就是每个变量都可以随机地取得不同的数值，而在进行试验或测量之前，我们要预言这个变量将取得某个确定的数值是不可能的。

按照随机变量可能取得的值，可以把它们分为两种基本类型：离散型和连续型。

离散型随机变量即在一定区间内变量取值为有限个或可数个。例如某地区某年人口的出生数、死亡数，某药治疗某病病人的有效数、无效数等。离散型随机变量根据不同的概率分布有伯努利分布、二项分布、几何分布、泊松分布、超几何分布等。

连续型随机变量即在一定区间内变量取值有无限个，或数值无法一一列举出来。例如某地区男性健康成人的身高值、体重值，一批传染性肝炎患者的血清转氨酶测定值等。连续型随机变量根据不同的概率分布有均匀分布、指数分布、正态分布、伽马分布等。

随机变量的性质主要有两类：一类是大而全的性质，这类性质可以详细描述所有可能取值的概率，例如描述连续型随机变量的累积分布函数（Cumulative Distribution Function，CDF）、概率密度函数（Probability Density Function，PDF），描述离散型随机变量的概率质量分布函数（Probability Mass Function，PMF）等；另一类是找到该随机变量的一些特征或代表值，例如随机变量的方差（Variance）、期望（Expectation）、置信区间等数字特征。

SciPy、NumPy、Matplotlib是Python中使用最为广泛的科学计算工具包，基本上可以处理大部分的统计与可视化作图等任务。SciPy有个stats模块，这个模块中包含了概率论及统计相关的函数。下面我们用Python来模拟实现各种随机变量的分布及其图形化显示。

3.1.3 离散随机变量分布 Python 实验

1. 伯努利分布

伯努利分布(Bernoulli Distribution)又称两点分布或 0-1 分布,其样本空间中只有两个点,一般取为{0,1},不同的伯努利分布只是取到这两个值的概率不同。伯努利分布只有一个参数 p,记作 $X\sim$Bernoulli(p),或 $X\sim$B(1,p),读作 X 服从参数为 p 的伯努利分布。

下面以抛硬币为例模拟伯努利分布。如果将抛一次硬币看作一次伯努利实验,且将正面朝上记为 1,反面朝上记为 0。那么伯努利分布中的参数 p 就表示硬币正面朝上的概率。

伯努利分布的概率分布用 Python 代码绘制如下:

```
#第 3 章/bernoulli_pmf.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def bernoulli_pmf(p = 0.0):
    ber_dist = stats.bernoulli(p)
    x = [0, 1]
    x_name = ['0', '1']
    pmf = [ber_dist.pmf(x[0]), ber_dist.pmf(x[1])]
    plt.bar(x, pmf, width = 0.15)
    plt.xticks(x, x_name)
    plt.ylabel('Probability')
    plt.title('PMF of bernoulli distribution')
    plt.show()

bernoulli_pmf(p = 0.3)
```

图 3-1 是该程序的运行结果图。Probability 为概率,横坐标表示随机变量 X,1 表示正面朝上。

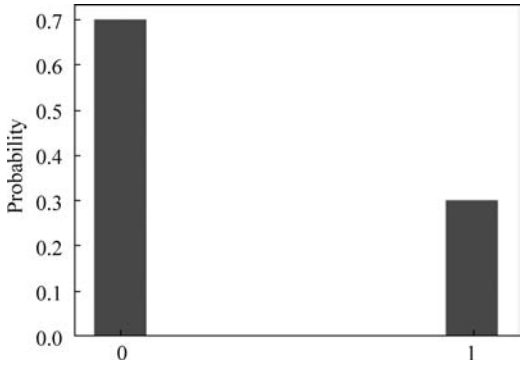

图 3-1 伯努利分布 B(1,0.3)的 PMF 柱状图

通常为了得到比较准确的某个服从伯努利分布的随机变量的期望,需要大量重复伯努利试验,例如重复 n 次,然后利用“正面朝上的次数/n”来估计 p。

2．二项分布

如果把一个伯努利分布独立地重复 n 次，就得到了一个二项分布。二项分布是最重要的离散型概率分布之一。随机变量 X 要满足这个分布有两个重要条件：①各次试验的条件是稳定的；②各次试验之间是相互独立的。

还是以利用抛硬币的例子来比较伯努利分布和二项分布的区别。如果将抛一次硬币看作一次伯努利实验，且将正面朝上记为1，反面朝上记为0。那么抛 n 次硬币，记录正面朝上的次数 Y，Y 就服从二项分布。假如硬币是均匀的，Y 的取值应该大部分集中在 $n/2$ 附近，而非常大或非常小的值都很少。由此可见，二项分布关注的是计数，而伯努利分布关注的是比值(正面朝上的计数/n)。一个随机变量 X 服从参数为 n 和 p 的二项分布，记作 $X\sim$ Binomial(n，p)或 $X\sim$B(n，p)。

下面利用 Python 代码模拟抛一枚不均匀的硬币 20 次，设正面朝上的概率为 0.6。

```
#第 3 章/binom_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def binom_dis(n = 1, p = 0.1):
    binom_dis = stats.binom(n, p)
    x = np.arange(binom_dis.ppf(0.0001), binom_dis.ppf(0.9999))
    print(x)#[ 0.  1.  2.  3.  4.]
    fig, ax = plt.subplots(1, 1)
    ax.plot(x, binom_dis.pmf(x), 'bo', label = 'binom pmf')
    ax.vlines(x, 0, binom_dis.pmf(x), colors = 'b', lw = 5, alpha = 0.5)
    ax.legend(loc = 'best', frameon = False)
    plt.ylabel('Probability')
    plt.title('PMF of binomial distribution(n = {}, p = {})'.format(n, p))
    plt.show()

binom_dis(n = 20, p = 0.6)
```

以上定义了一个 $n=20$，$p=0.6$ 的二项分布，表示每次试验抛硬币，该硬币正面朝上的概率大于背面朝上的概率，共抛 20 次并记录正面朝上的次数。该分布的概率质量分布函数如图 3-2 所示，Probability 为概率，X 表示随机变量：正面朝上的次数。

从图 3-2 中可以看出，该分布的概率质量分布函数 PMF 图明显向右边偏移，在 $x=12$ 处取到最大概率，这是因为这个硬币正面朝上的概率大于反面朝上的概率。

为了比较准确地得到某个服从二项分布的随机变量的期望，需要大量重复二项分布试验，例如有 m 个人进行试验，每人抛 n 次，然后利用“所有人得到的正面次数之和/m”来估计 n 和 p，总共相当于做了 $n\times m$ 次伯努利实验。

3．泊松分布

日常生活中很多事件的发生是有固定频率的，例如某医院平均每小时出生的婴儿数，某网站平均每分钟的访问数、某一服务设施在一定时间内收到的服务请求的次数、汽车站台的候客人数等。它们的特点是可以预估这些事件在某个时间段内发生的总次数，但是没法知

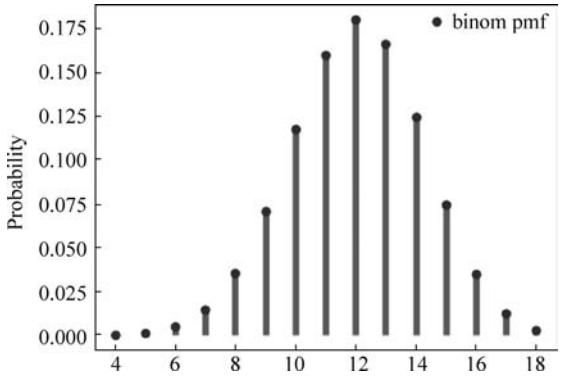

图 3-2　二项分布 B(20,0.6)的 PMF 图

道具体的发生时间。如果某事件以固定强度 λ 随机且独立地出现,该事件在单位时间内出现的次数(个数)可以看成是服从泊松分布。我们把一个随机变量 X 服从参数为 λ 的泊松分布,记作 $X\sim$Poisson(λ),或 $X\sim$P(λ)。泊松分布适合于描述单位时间内随机事件发生次数的概率分布。

下面是参数 $\mu=8$ 时的泊松分布 Python 实现,在 SciPy 中将泊松分布的参数表示为 μ。

```
#第 3 章/poisson_pmf.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def poisson_pmf(mu = 3):
    poisson_dis = stats.poisson(mu)
    x = np.arange(poisson_dis.ppf(0.001), poisson_dis.ppf(0.999))
    print(x)
    fig, ax = plt.subplots(1, 1)
    ax.plot(x, poisson_dis.pmf(x), 'bo', ms = 8, label = 'poisson pmf')
    ax.vlines(x, 0, poisson_dis.pmf(x), colors = 'b', lw = 5, alpha = 0.5)
    ax.legend(loc = 'best', frameon = False)
    plt.ylabel('Probability')
    plt.title('PMF of poisson distribution(mu = {})'.format(mu))
    plt.show()

poisson_pmf(mu = 8)
```

代码运行后输出的概率质量分布 PMF 如图 3-3 所示。

如果仅仅看二项分布与泊松分布的概率质量分布图,也可以发现它们的相似度非常高。泊松分布可以作为二项分布的极限。一般来说,若 $X\sim$B(n,p),其中 n 很大,p 很小,而 $np=\lambda$ 不太大时,则 X 的分布接近于泊松分布 P(λ)。

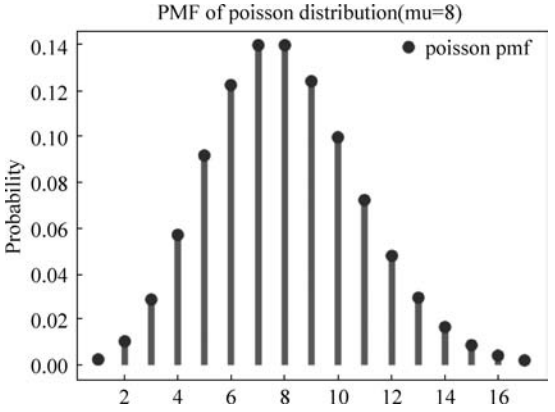

图 3-3　二项分布 B(20,0.6)的 PMF 图

3.1.4　连续随机变量分布 Python 实验

1. 均匀分布

均匀分布(Uniform Distribution)是最简单的连续型概率分布,因为其概率密度是一个常数,不随随机变量 X 取值的变化而变化。如果连续型随机变量 X 具有如下的概率密度函数,则称 X 服从 $[a,b]$ 上的均匀分布,记作 $X\sim \mathrm{U}(a,b)$或 $X\sim \mathrm{Unif}(a,b)$。

$$f_X(x)=\begin{cases}\dfrac{1}{b-a} & a<x<b\\ 0 & x<a \text{ 或 } x>b\end{cases} \tag{3-1}$$

从式(3-1)可以看出,定义一个均匀分布需要两个参数,定义域区间的起点 a 和终点 b,在 Python 中用 location 和 scale 分别表示起点和区间长度,代码如下:

```
#第 3 章/uniform_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def uniform_distribution(loc = 0, scale = 1):
    uniform_dis = stats.uniform(loc = loc, scale = scale)
    x = np.linspace(uniform_dis.ppf(0.01),
                    uniform_dis.ppf(0.99), 100)
    fig, ax = plt.subplots(1, 1)

    #直接传入参数
    ax.plot(x, stats.uniform.pdf(x, loc = 2, scale = 4), 'r-',
            lw = 5, alpha = 0.6, label = 'uniform pdf')

    #从冻结的均匀分布取值
    ax.plot(x, uniform_dis.pdf(x), 'k-',
            lw = 2, label = 'frozen pdf')
```

```
    #计算 ppf 分别等于 0.001、0.5 和 0.999 时的 x 值
    vals = uniform_dis.ppf([0.001, 0.5, 0.999])
    print(vals)#[ 2.004  4.     5.996]

    #检测 cdf 和 ppf 的精确度
    print(np.allclose([0.001, 0.5, 0.999], uniform_dis.cdf(vals)))     #结果为 True

    r = uniform_dis.rvs(size = 10000)
    ax.hist(r, normed = True, histtype = 'stepfilled', alpha = 0.2)
    plt.ylabel('Probability')
    plt.title(r'PDF of Unif({}, {})'.format(loc, loc + scale))
    ax.legend(loc = 'best', frameon = False)
    plt.show()

uniform_distribution(loc = 2, scale = 4)
```

上面的代码采用两种方式：直接传入参数和先冻结一个分布，画出来均匀分布的概率分布函数，此外还从该分布中取了 10000 个值作直方图，运行结果如图 3-4 所示。

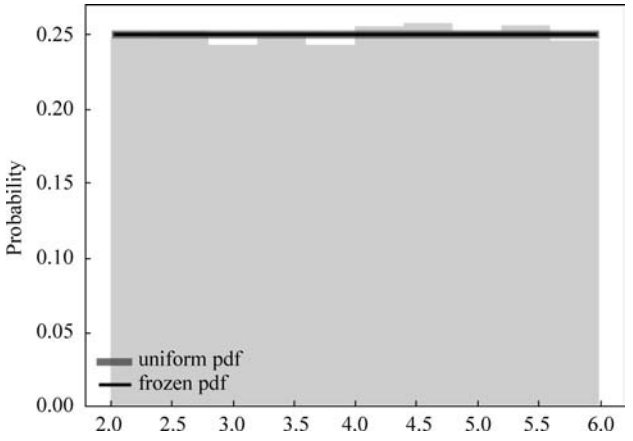

图 3-4　均匀分布 U(2,6)的 PDF 图

2. 指数分布

指数分布和离散型的泊松分布之间有很大的关系。泊松分布表示单位时间(或单位面积)内随机事件的平均发生次数，指数分布则可以用来表示独立随机事件发生的时间间隔。由于发生次数只能是自然数，所以泊松分布自然就是离散型的随机变量，而时间间隔则可以是任意的实数，因此其定义域是(0，+∞)。

如果一个随机变量 X 的概率密度函数满足以下形式，就称 X 为服从参数 λ 的指数分布(Exponential Distribution)，记作 $X\sim E(\lambda)$或 $X\sim Exp(\lambda)$。指数分布只有一个参数 λ，且 $\lambda>0$。

$$f_X(x)=\begin{cases}\lambda e^{-\lambda e} & x>0\\ 0 & \text{其他}\end{cases} \tag{3-2}$$

指数分布的一个显著的特点是其具有无记忆性。例如，如果排队的顾客接受服务的时间长短服从指数分布，那么无论你已经排了多长时间的队，在排 t 分钟的概率始终是相同

的。用公式表示为 $P(X \geqslant s+t \mid X \geqslant s)=P(X \geqslant t)$，代码如下：

```
#第 3 章/exponential_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def exponential_dis(loc = 0, scale = 1.0):
    """
    指数分布,exponential continuous random variable
    按照定义,指数分布只有一个参数 lambda,这里的 scale = 1/lambda
    :param loc:定义域的左端点,相当于将整体分布沿 x 轴平移 loc
    :param scale: lambda 的倒数,loc + scale 表示该分布的均值,scale^2 表示该分布的方差
    :return:
    """
    exp_dis = stats.expon(loc = loc, scale = scale)
    x = np.linspace(exp_dis.ppf(0.000001),
                    exp_dis.ppf(0.999999), 100)
    fig, ax = plt.subplots(1, 1)

    #直接传入参数
    ax.plot(x, stats.expon.pdf(x, loc = loc, scale = scale), 'r-',
            lw = 5, alpha = 0.6, label = 'uniform pdf')

    #从冻结的均匀分布取值
    ax.plot(x, exp_dis.pdf(x), 'k-',
            lw = 2, label = 'frozen pdf')

    #计算 ppf 分别等于 0.001、0.5 和 0.999 时的x 值
    vals = exp_dis.ppf([0.001, 0.5, 0.999])
    print(vals)#[ 2.004  4.     5.996]

    #检测 cdf 和 ppf 的精确度
    print(np.allclose([0.001, 0.5, 0.999], exp_dis.cdf(vals)))

    r = exp_dis.rvs(size = 10000)
    ax.hist(r, normed = True, histtype = 'stepfilled', alpha = 0.2)
    plt.ylabel('Probability')
    plt.title(r'PDF of Exp(0.5)')
    ax.legend(loc = 'best', frameon = False)
    plt.show()

exponential_dis(loc = 0, scale = 2)
```

Exp(0.5)的 PDF 如图 3-5 所示。

3. 正态分布

正态分布(Normal Distribution)，也称常态分布，又名高斯分布(Gaussian Distribution)，最早由 A.棣莫弗在求二项分布的渐近公式中得到。C.F.高斯在研究测量误差时从另一个角度导出了它。P.S.拉普拉斯和高斯研究了它的性质。正态分布是一个在数学、物理及工程等领

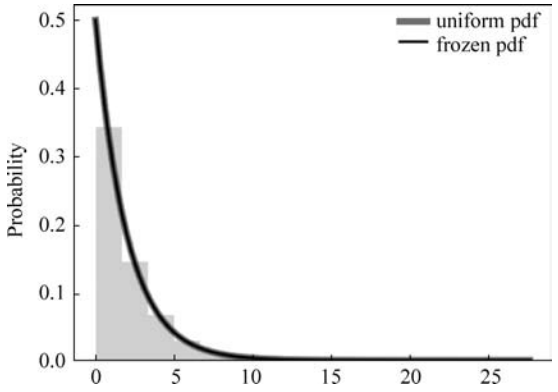

图 3-5　二项分布 B(20,0.6)的 PDF 图

域都非常重要的概率分布,在统计学的许多方面有着重大的影响力。正态曲线呈钟形,两头低,中间高,左右对称因其曲线呈钟形,因此人们又经常称其为钟形曲线。

若随机变量 X 服从一个数学期望为 μ、方差为 σ^2 的正态分布,记为 $N(\mu,\sigma^2)$。其概率密度函数为正态分布的期望值 μ 决定了其位置,其标准差 σ 决定了分布的幅度。当 $\mu=0$,$\sigma=1$ 时的正态分布是标准正态分布。其概率密度函数为

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3-3}$$

当 $\mu=0$,$\sigma=1$ 时,该正态分布称为标准正态分布。由于标准正态分布在统计学中的重要地位,它的累积分布函数 CDF 有一个专门的表示符号:Φ。正态分布,Python 代码如下:

```
#第 3 章/normal_dis.py
#绘制正态分布概率密度函数
import numpy as np
import matplotlib.pyplot as plt
import math

u = 0                    #均值 μ
u01 = -2
sig = math.sqrt(0.2)  #标准差 δ

x = np.linspace(u - 3 * sig, u + 3 * sig, 50)
y_sig = np.exp(-(x - u) ** 2 / (2 * sig ** 2)) / (math.sqrt(2 * math.pi) * sig)
print(x)
print("=" * 20)
print(y_sig)
plt.plot(x, y_sig, "r-", linewidth=2)
plt.grid(True)
plt.show()
```

该程序运行结果如图 3-6 所示。

正态分布中的两个参数含义如下:

当固定 σ,改变 μ 的大小时,$f(x)$图形的形状不变,只是沿着 x 轴作平移变换,因此 μ

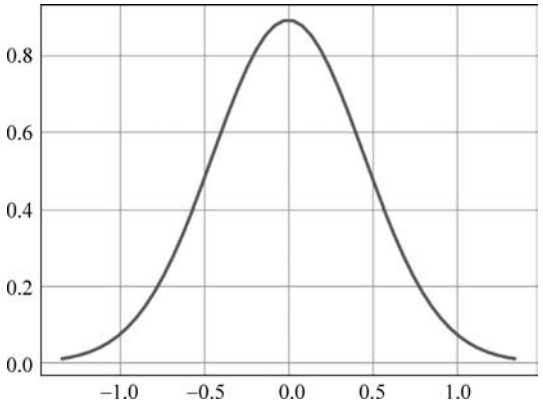

图 3-6　标准正态分布图

被称为位置参数(决定对称轴的位置)；

当固定 μ，改变 σ 的大小时，$f(x)$图形的对称轴不变，形状改变，σ 越小，图形尖峰越陡峭。σ 越大，图形越平坦，因此 σ 被称为尺度参数，决定曲线的分散程度。

3.2　数理统计基础

数理统计(Mathematics Statistics)是一门应用性很强的数学分支，它以概率论为基础，根据试验或者观察得到的数据来研究随机现象，对随机现象的概率 p 做出一些合理的估计和判断。数理统计包含许多内容，但其核心部分是统计推断，它包括两个基本问题，即参数估计和假设检验。

3.2.1　总体和样本

统计学中的概念很多，其中有几个概念是经常用到的，包括总体、样本、参数、统计量、变量等。在数理统计中，称研究对象的全体为总体，组成总体的每个基本单元叫个体。从总体 X 中随机抽取一部分个体 $X_1, X_2, \cdots, X_n$，称 $X_1, X_2, \cdots, X_n$ 为取自 X 的容量为 n 的样本。

例如，为了研究某厂生产的一批元件质量的好坏，规定使用寿命低于 1000 小时的为次品，则该批元件的全体就为总体，每个元件就是个体。实际上，数理统计学中的总体是指与总体相联系的某个(或某几个)数量指标 X 取值的全体。例如，该批元件的使用寿命 X 的取值全体就是研究对象的总体。显然 X 是随机变量，这时就称 X 为总体。

为了判断该批元件的次品率，最精确的办法是取出全部元件，对元件的寿命做试验。然而，寿命试验具有破坏性，即使某些试验是非破坏性的，因此只能从总体中抽取一部分，对 n 个个体进行试验。试验结果可得到一组数值集合$\{x_1, x_2, \cdots, x_n\}$，其中每个 x_i 是第 i 次抽样观察的结果。由于要根据这些观察结果来对总体进行推断，所以对每次抽样需要有一定的要求，要求每次抽取样本必须是随机的、独立的，这样才能较好地反映总体情况。所谓随机是指每个个体被抽到的机会是均等的，这样抽到的个体才具有代表性。若 $x_1, x_2, \cdots, x_n$ 相互独立，且每个 x_i 与 X 同分布，则称 $x_1, x_2, \cdots, x_n$ 为简单随机样本，简称样本。通常把

n 称为样本容量。

值得注意的是，样本具有两重性，即当在一次具体的抽样后它是一组确定的数值。但在一般叙述中样本也是一组随机变量，因为抽样是随机的。一般地，用 $X_1, X_2, \cdots, X_n$ 表示随机样本，它们取到的值记为 $x_1, x_2, \cdots, x_n$ 称为样本观测值。

样本作为随机变量有一定的概率分布，这个概率分布称为样本分布。显然，样本分布取决于总体的性质和样本的性质。

3.2.2 统计量与抽样分布

1. 统计量

数理统计的任务是采集和处理带有随机影响的数据，或者说收集样本并对之进行加工，以此对所研究的问题做出一定的结论，这一过程称为统计推断。在统计推断中，对样本进行加工整理，实际上就是根据样本计算出一些量，使得这些量能够将所研究问题的信息集中起来。这种根据样本计算出的量就是下面将要定义的统计量，因此，统计量是样本的某种函数。

设 $X_1, X_2, \cdots, X_n$ 是总体 X 的一个简单随机样本，$T(X_1, X_2, \cdots, X_n)$ 为一个 n 元连续函数，且 T 中不包含任何关于总体的未知参数，则称 $T(X_1, X_2, \cdots, X_n)$ 是一个统计量，称统计量的分布为抽样分布。下面列出几个常用的统计量。

设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观测值。统计量定义如下。

样本均值

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3-4}$$

样本方差

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3-5}$$

样本标准差

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{3-6}$$

样本 k 阶(原点)矩

$$A_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k \quad k = 1, 2, \cdots \tag{3-7}$$

样本 k 阶中心矩

$$B_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k \quad k = 2, 3, \cdots \tag{3-8}$$

2. 三大抽样分布

统计量是样本的函数，它是一个随机变量。统计量的分布称为抽样分布，在使用统计量进行统计推断时常需知道它的分布。当总体的分布函数已知时，抽样分布是确定的，然而要求出统计量的精确分布比较困难。下面介绍几个常用的统计量分布。

- 卡方(χ^2)分布

设 $X_1, X_2, \cdots, X_n$ 是来自总体 $N(0,1)$ 的样本，则统计量 $\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2$ 所服从的分布称为自由度为 n 的 χ^2 分布(χ^2-distribution)，记为 $\chi^2 \sim \chi^2(n)$。$\chi^2(n)$ 分布的概率密度函数为

$$f(y) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)} y^{\frac{n}{2}-1} \mathrm{e}^{\frac{-y}{2}} & y > 0 \\ 0 & \text{其他} \end{cases} \tag{3-9}$$

不同参数的卡方分布，Python 实现代码如下：

```
#第 3 章/binom_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def diff_chi2_dis():
    """
    不同参数下的卡方分布
    :return:
    """
    #chi2_dis_0_5 = stats.chi2(df = 0.5)
    chi2_dis_1 = stats.chi2(df = 1)
    chi2_dis_4 = stats.chi2(df = 4)
    chi2_dis_10 = stats.chi2(df = 10)
    chi2_dis_20 = stats.chi2(df = 20)

    #x1 = np.linspace(chi2_dis_0_5.ppf(0.01), chi2_dis_0_5.ppf(0.99), 100)
    x2 = np.linspace(chi2_dis_1.ppf(0.65), chi2_dis_1.ppf(0.9999999), 100)
    x3 = np.linspace(chi2_dis_4.ppf(0.000001), chi2_dis_4.ppf(0.999999), 100)
    x4 = np.linspace(chi2_dis_10.ppf(0.000001), chi2_dis_10.ppf(0.99999), 100)
    x5 = np.linspace(chi2_dis_20.ppf(0.00000001), chi2_dis_20.ppf(0.9999), 100)
    fig, ax = plt.subplots(1, 1)
    #ax.plot(x1, chi2_dis_0_5.pdf(x1), 'b-', lw = 2, label = r'df = 0.5')
    ax.plot(x2, chi2_dis_1.pdf(x2), 'g-', lw = 2, label = 'df = 1')
    ax.plot(x3, chi2_dis_4.pdf(x3), 'r-', lw = 2, label = 'df = 4')
    ax.plot(x4, chi2_dis_10.pdf(x4), 'b-', lw = 2, label = 'df = 10')
    ax.plot(x5, chi2_dis_20.pdf(x5), 'y-', lw = 2, label = 'df = 20')
    plt.ylabel('Probability')
    plt.title(r'PDF of $\chi^2$ Distribution')
    ax.legend(loc = 'best', frameon = False)
    plt.show()

diff_chi2_dis()
```

当自由度 df 等于 1 或 2 时，函数图像都呈单调递减的趋势；当 df 大于或等于 3 时，呈先增后减的趋势。从定义上来看，df 的值只能取正整数，如图 3-7 所示。

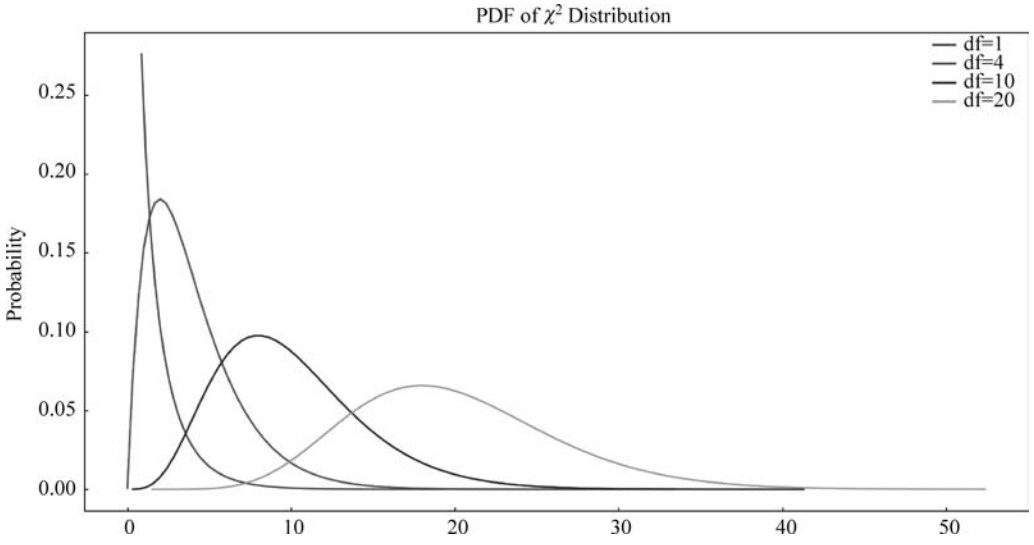

图 3-7　不同参数下卡方分布的 PDF 图

3. t 分布

设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，并且 X 和 Y 独立，则称随机变量 $t=\dfrac{X}{\sqrt{\dfrac{Y}{n}}}$ 服从自由度为 n 的 t 分布(t-distribution)，记为 $t\sim t(n)$。

$t(n)$分布的概率密度函数为

$$h(t)=\frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\,\Gamma(n/2)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \qquad -\infty<t<\infty \tag{3-10}$$

不同参数下的 t 分布，Python 实现代码如下：

```
#第 3 章/binom_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

def diff_t_dis():
    """
    不同参数下的 t 分布
    :return:
    """
    norm_dis = stats.norm()
    t_dis_1 = stats.t(df = 1)
    t_dis_4 = stats.t(df = 4)
    t_dis_10 = stats.t(df = 10)
    t_dis_20 = stats.t(df = 20)

    x1 = np.linspace(norm_dis.ppf(0.000001), norm_dis.ppf(0.999999), 1000)
    x2 = np.linspace(t_dis_1.ppf(0.04), t_dis_1.ppf(0.96), 1000)
```

```
    x3 = np.linspace(t_dis_4.ppf(0.001), t_dis_4.ppf(0.999), 1000)
    x4 = np.linspace(t_dis_10.ppf(0.001), t_dis_10.ppf(0.999), 1000)
    x5 = np.linspace(t_dis_20.ppf(0.001), t_dis_20.ppf(0.999), 1000)
    fig, ax = plt.subplots(1, 1)
    ax.plot(x1, norm_dis.pdf(x1), 'r-', lw=2, label=r'N(0, 1)')
    ax.plot(x2, t_dis_1.pdf(x2), 'b-', lw=2, label='t(1)')
    ax.plot(x3, t_dis_4.pdf(x3), 'g-', lw=2, label='t(4)')
    ax.plot(x4, t_dis_10.pdf(x4), 'm-', lw=2, label='t(10)')
    ax.plot(x5, t_dis_20.pdf(x5), 'y-', lw=2, label='t(20)')
    plt.ylabel('Probability')
    plt.title(r'PDF of t Distribution')
    ax.legend(loc='best', frameon=False)
    plt.show()

diff_t_dis()
```

从图 3-8 中可以看到，$t(1)$与标准正态分布之间的差别还是比较大的，但是当自由度 n 趋近于无穷大时，t 分布与标准正态分布没有差别(公式上的形式将变得完全相同，这里没有列出概率密度函数的公式)。较大的区别在于，当自由度 n 较小时，t 分布比标准正态分布的尾部(Fatter Tails)更宽，因此也比正态分布更慢地趋近于 0。关于这两类分布的异同将会在后面的假设检验部分详细阐述。

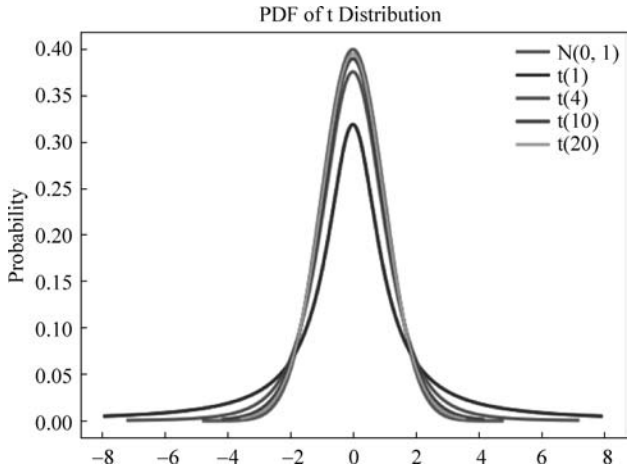

图 3-8　不同参数下的 t 分布 PDF 图

4. F 分布

设 $U\sim\chi^2(n_1)$，$V\sim\chi^2(n_2)$，且 U 和 V 独立，则称随机变量 $F=\dfrac{U/n_1}{V/n_2}$ 服从自由度为 (n_1,n_2)的 F 分布(F-distribution)，记：$F\sim F(n_1,n_2)$。

$F(n_1,n_2)$分布的概率密度为

$$\psi(y)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2]\,(n_1/n_2)^{n_1/2}\,y^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)\,[1+(n_1y/n_2)]^{(n_1+n_2)/2}} & y>0\\ 0 & \text{其他}\end{cases}\tag{3-11}$$

F 分布经常被用来对两个样本方差进行比较。它是方差分析的一个基本分布，也被用于回归分析中的显著性检验。

F 分布有两个参数：dfn 和 dfd，分别代表分子上的第一自由度和分母上的第二自由度。下面是不同参数下 F 分布的 Python 实现，代码如下：

```
#第 3 章/binom_dis.py
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt


def diff_f_dis():
    """
    不同参数下的 F 分布
    :return:
    """
#f_dis_0_5 = stats.f(dfn = 10, dfd = 1)
    f_dis_1_30 = stats.f(dfn = 1, dfd = 30)
    f_dis_30_5 = stats.f(dfn = 30, dfd = 5)
    f_dis_30_30 = stats.f(dfn = 30, dfd = 30)
    f_dis_30_100 = stats.f(dfn = 30, dfd = 100)
    f_dis_100_100 = stats.f(dfn = 100, dfd = 100)

#x1 = np.linspace(f_dis_0_5.ppf(0.01), f_dis_0_5.ppf(0.99), 100)
    x2 = np.linspace(f_dis_1_30.ppf(0.2), f_dis_1_30.ppf(0.99), 100)
    x3 = np.linspace(f_dis_30_5.ppf(0.00001), f_dis_30_5.ppf(0.99), 100)
    x4 = np.linspace(f_dis_30_30.ppf(0.00001), f_dis_30_30.ppf(0.999), 100)
    x6 = np.linspace(f_dis_30_100.ppf(0.0001), f_dis_30_100.ppf(0.999), 100)
    x5 = np.linspace(f_dis_100_100.ppf(0.0001), f_dis_100_100.ppf(0.9999), 100)
    fig, ax = plt.subplots(1, 1, figsize = (20, 10))
#ax.plot(x1, f_dis_0_5.pdf(x1), 'b - ', lw = 2, label = r'F(0.5, 0.5)')
    ax.plot(x2, f_dis_1_30.pdf(x2), 'g - ', lw = 2, label = 'F(1, 30)')
    ax.plot(x3, f_dis_30_5.pdf(x3), 'r - ', lw = 2, label = 'F(30, 5)')
    ax.plot(x4, f_dis_30_30.pdf(x4), 'm - ', lw = 2, label = 'F(30, 30)')
    ax.plot(x6, f_dis_30_100.pdf(x6), 'c - ', lw = 2, label = 'F(30, 100)')
    ax.plot(x5, f_dis_100_100.pdf(x5), 'y - ', lw = 2, label = 'F(100, 100)')

    plt.ylabel('Probability')
    plt.title(r'PDF of f Distribution')
    ax.legend(loc = 'best', frameon = False)
    plt.savefig('f_diff_pdf.png', dip = 500)
    plt.show()

diff_f_dis()
```

不同参数下 F 分布的 PDF 图如图 3-9 所示。

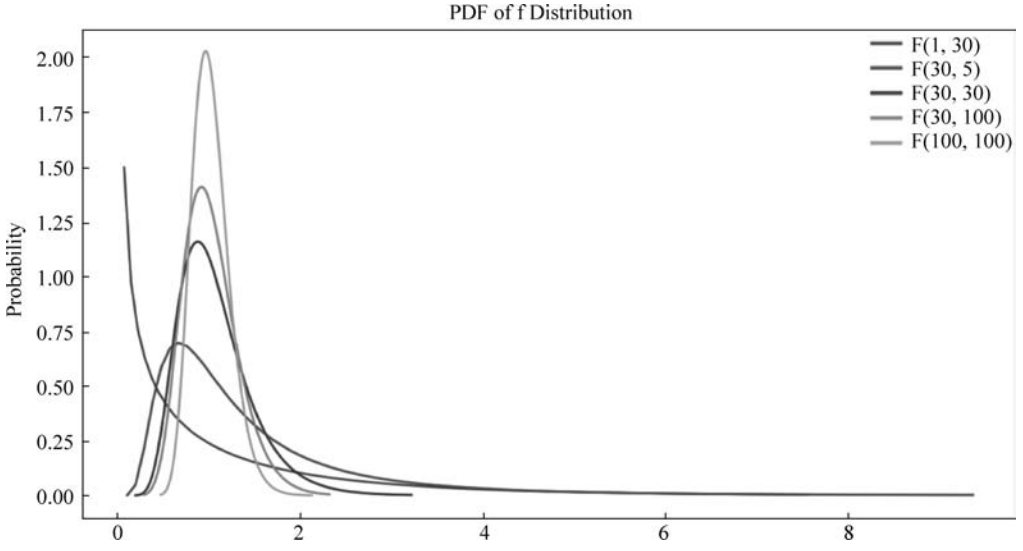

图 3-9　不同参数下 F 分布的 PDF 图

3.2.3　大数定律与中心极限定理

1. 大数定律

大数定律是一种描述当试验次数很大时所呈现的概率性质的定律，它由概率统计定义“频率收敛于概率”引申而来。简而言之，就是 n 个独立分布的随机变量其观察值的均值 $\overline{X}$ 依概率收敛于这些随机变量所属分布的理论均值，也就是总体均值。大数定律为这种后验认识世界的方式提供了坚实的理论基础。统计学中常采用大数定律用样本均值来估计总体的期望。例如，假设每次从 1、2、3 当中随机选取一个数字，随着抽样次数的增加，样本均值越来越趋近于总体期望((1＋2＋3)/3＝2)。

下面用程序模拟抛硬币的过程来辅助说明大数定律：用 random 模块生成区间[0,1)之间的随机数，如果生成的数小于 0.5，就记为硬币正面朝上，否则记为硬币反面朝上。由于 random.random()生成的数可以看作服从区间[0,1)上的均匀分布，所以以 0.5 为界限，随机生成的数中大于 0.5 或小于 0.5 的概率应该是相同的，相当于硬币出现正反面的概率是均匀的。这样就用随机数模拟出了实际的抛硬币试验。理论上试验次数越多，即抛硬币的次数越多，正反面出现的次数之比越接近于 1，也就是说正反面各占一半。

```
#第 3 章/binom_dis.py
import random
import matplotlib.pyplot as plt

def flip_plot(minExp, maxExp):
    """
    Assumes minExp and maxExp positive integers; minExp < maxExp
    Plots results of 2 ** minExp to 2 ** maxExp coin flips
    """
```

```
    #两个参数的含义,抛硬币的次数为 2 的 minExp 次方到 2 的 maxExp 次方,也就是一共做了
#(2 ** maxExp - 2 ** minExp)批次实验,每批次重复抛硬币 2 ** n 次

    ratios = []
    xAxis = []
    for exp in range(minExp, maxExp + 1):
        xAxis.append(2 ** exp)
    for numFlips in xAxis:
        numHeads = 0                        #初始化,硬币正面朝上的计数为 0
        for n in range(numFlips):
            if random.random() < 0.5:       #random.random()从[0, 1)随机地取出一个数
                numHeads += 1               #当随机取出的数小于 0.5 时,正面朝上的计数加 1
        numTails = numFlips - numHeads      #得到本次试验中反面朝上的次数
        ratios.append(numHeads/float(numTails))   #正反面计数的比值
    plt.title('Heads/Tails Ratios')
    plt.xlabel('Number of Flips')
    plt.ylabel('Heads/Tails')
    plt.plot(xAxis, ratios)
    plt.hlines(1, 0, xAxis[-1], linestyles = 'dashed', colors = 'r')
    plt.show()

flip_plot(4, 16)
```

该程序顺利执行后将会出现如图 3-10 所示界面。

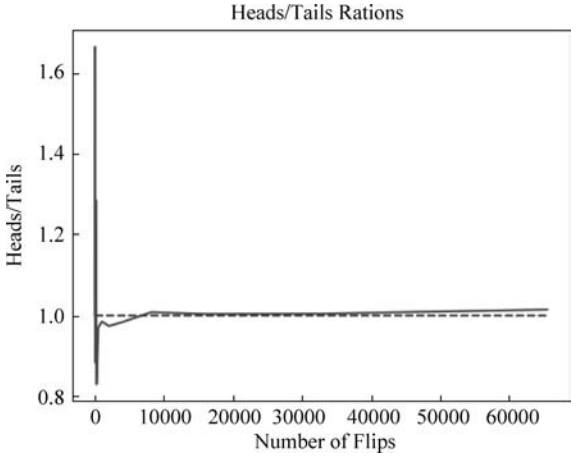

图 3-10　大数定律模拟输出

从图 3-10 可以看出,随着实验次数的增加,硬币正反面出现次数之比越来越接近于 1。

2. 中心极限定理

大数定律揭示了大量随机变量的平均结果,但没有涉及随机变量的分布问题。而中心极限定理说明的是在一定条件下,大量独立随机变量的平均数是以正态分布为极限的。中心极限定理指出大量的独立随机变量之和具有近似于正态的分布。因此,它不仅提供了计算独立随机变量之和的近似概率的简单方法,而且有助于解释为什么有很多自然群体的经验频率呈现出钟形(即正态)曲线这一事实,因此中心极限定理这个结论使正态分布在数理

统计中具有很重要的地位，也使正态分布有了广泛的应用。中心极限定理有辛钦中心极限定理、德莫佛-拉普拉斯中心极限定理、李亚普洛夫中心极限定理、林德贝尔格定理等表现形式。

下面使用 Python 来模拟并验证中心极限定理。假设我们现在观测一个人掷骰子。这个骰子是公平的，也就是说掷出 1～6 的概率都是相同的：1/6。他掷了一万次，用 Python 来模拟投掷过程，代码如下：

```
#第 3 章/central_limit.py
import numpy as np
random_data = np.random.randint(1, 7, 10000)
print random_data.mean()     #打印平均值
print random_data.std()      #打印标准差
```

生成的平均值：3.4927（注意：每次执行输出结果会略有不同），生成的标准差：1.7079。

平均值接近 3.5 很好理解，因为每次掷出来的结果是 1、2、3、4、5、6。每个结果的概率是 1/6，所以加权平均值就是 3.5。接下来随机抽取一组抽样，例如从生成的数据中随机抽取 10 个数字：

```
sample1 = []
for i in range(0, 10):
    sample1.append(random_data[int(np.random.random() * len(random_data))])

print sample1     #打印出来
```

这 10 个数字的结果是：[3,4,3,6,1,6,6,3,4,4]。

平均值：4.0。

标准差：1.54。

可以看到，只抽 10 个样本的时候，样本的平均值（4.0）距离总体的平均值（3.5）有所偏差。

下面抽取 1000 组，每组 50 个样本，并且把每组的平均值都算出来。

```
samples = []
samples_mean = []
samples_std = []

for i in range(0, 1000):
    sample = []
    for j in range(0, 50):
        sample.append(random_data[int(np.random.random() * len(random_data))])
    sample_np = np.array(sample)
    samples_mean.append(sample_np.mean())
    samples_std.append(sample_np.std())
    samples.append(sample_np)
```

```
samples_mean_np = np.array(samples_mean)
samples_std_np = np.array(samples_std)

print samples_mean_np
```

共1000组平均值大概是这样的：[3.44,3.42,3.22,3.2,2.94 … 4.08,3.74]，结果输出如下：

平均值：3.48494。

标准差：0.23506。

然后把这1000组数字用直方图画出来，如图3-11所示。

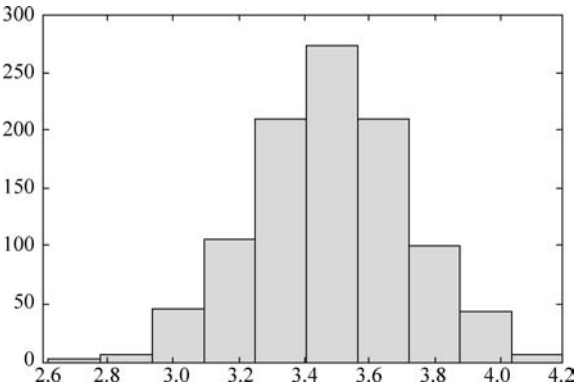

图3-11　大数定律下中心极限的正态分布图

从图3-11中可以看出，其结果已经非常接近正态分布了。

3.3　参数估计

参数估计是数理统计研究的主要问题之一。假设总体 $X\sim N(\mu,\sigma^2)$，μ 和 σ^2 是未知参数，$X_1,X_2,\cdots,X_n$ 是来自 X 的样本，样本值是 $x_1,x_2,\cdots,x_n$，要由样本值来确定 μ 和 σ^2 的估计值，这就是参数估计问题，参数估计分为点估计(Point Estimation)和区间估计(Interval Estimation)。

3.3.1　点估计

所谓点估计是指把总体的未知参数估计为某个确定的值或在某个确定的点上，所以点估计又称为定值估计。

设总体 X 的分布函数为 $F(x,\theta)$，θ 是未知参数，$X_1,X_2,\cdots,X_n$ 是 X 的一样本，样本值为 $x_1,x_2,\cdots,x_n$，构造一个统计量$(X_1,X_2,\cdots,X_n)$，用它的观察值$(x_1,x_2,\cdots,x_n)$作为 θ 的估计值，这种问题称为点估计问题。习惯上称随机变量$(X_1,X_2,\cdots,X_n)$为 θ 的估计量，称$(x_1,x_2,\cdots,x_n)$为估计值。

构造估计量$(X_1,X_2,\cdots,X_n)$的方法很多，下面介绍常用的矩估计法和极大似然估计法。

1. 矩估计法

矩估计法是一种古老的估计方法。矩是描述随机变量的最简单的数字特征，样本来自于总体，样本矩在一定程度上也反映了总体矩的特征，且在样本容量 n 增大的条件下，样本的 k 阶原点矩 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$ 依概率收敛到总体 X 的 k 阶原点矩 $\mu_k=E(X^k)$，即 $A_k\xrightarrow{p}\mu_k(n\to\infty)$，$k=1,2,\cdots$，因而自然想到用样本矩作为相应总体矩的估计量，而以样本矩的连续函数作为相应总体矩的连续函数的估计量，这种估计方法就称为矩估计法。

矩估计法的一般做法：设总体 $X\sim F(X;\theta_1,\theta_2,\cdots,\theta_l)$ 其中 $\theta_1,\theta_2,\cdots,\theta_l$ 均未知。

(1) 如果总体 X 的 k 阶矩 $\mu_k=E(X^k)(1\leqslant k\leqslant l)$ 均存在，则

$$\mu_k=\mu_k(\theta_1,\theta_2,\cdots,\theta_l),\quad(1\leqslant k\leqslant l)$$

(2) 令
$$\begin{cases}\mu_1(\theta_1,\theta_2,\cdots,\theta_l)A_1\\ \mu_2(\theta_1,\theta_2,\cdots,\theta_l)A_2\\ \vdots\\ \mu_i(\theta_1,\theta_2,\cdots,\theta_l)A_l\end{cases}$$

其中，$A_k(1\leqslant k\leqslant l)$ 为样本 k 阶矩。

求出方程组的解 $\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_l$，称 $\hat{\theta}_k=\hat{\theta}_k(X_1,X_2,\cdots,X_n)$ 为参数 $\theta_k(1\leqslant k\leqslant l)$ 的矩估计量，$\hat{\theta}_k=\hat{\theta}_k(x_1,x_2,\cdots,x_n)$ 为参数 θ_k 的矩估计值。

【例 3-2】 设总体 X 的密度函数为

$$f(x;\theta)=\begin{cases}e^{-(x-\theta)} & x>\theta\\ 0 & x<\theta\end{cases}$$

其中 θ 未知，样本为 $(X_1,X_2,\cdots,X_n)$，求参数 θ 的矩估计值。

解：$A_1=\overline{X}$。由 $\mu_1=A_1$ 及

$$\mu_1=E(X)=\int_{-\infty}^{+\infty}xf(x;\theta)\mathrm{d}x=\int_{0}^{\infty}xe^{-(x-\theta)}\mathrm{d}x=\theta+1$$

有 $\theta=\mu_1-1$，得 $A_1=\overline{X}$ 代替上式中的 μ_1，得到参数 θ 的矩估计量为

$$\hat{\theta}=\overline{X}-1$$

其中 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$。

2. 极大似然估计法

极大似然估计法(Maximum Likelihood Estimation)只能在已知总体分布的前提下进行，例如，假定一个盒子里装有许多大小相同的黑球和白球，并且假定它们的数目之比为 3∶1，但不知是白球多还是黑球多，现在有放回地从盒中抽了 3 个球，试根据所抽 3 个球中黑球的数目确定是白球多还是黑球多.

解：设所抽 3 个球中黑球数为 X，摸到黑球的概率为 p，则 X 服从二项分布

$$P\{X=k\}=C_3^k p^k(1-p)^{3-k}\quad k=0,1,2,3。$$

问题是 $p=1/4$ 还是 $p=3/4$ 呢？现根据样本中黑球数，对未知参数 p 进行估计。抽样后，共有 4 种可能结果，其概率如表 3-1 所示。

表 3-1 概论分布律

X	0	1	2	3
$p=1/4$ 时，$P\{X=k\}$	27/64	27/64	9/64	1/64
$p=3/4$ 时，$P\{X=k\}$	1/64	9/64	27/64	27/64

假如某次抽样中，只出现一个黑球，即 $X=1$，$p=1/4$ 时，$P\{X=1\}=27/64$；$p=3/4$ 时，$P\{X=1\}=9/64$，这时我们就会选择 $p=1/4$，即黑球数比白球数为 1∶3。因为在一次试验中，事件“1 个黑球”发生了。我们认为它应有较大的概率 27/64(27/64>9/64)，而 27/64 对应着参数 $p=1/4$，同样可以考虑 $X=0,2,3$ 的情形，最后可得

$$p=\begin{cases}\dfrac{1}{4} & \text{当 } x=0,1 \text{ 时}\\[2ex] \dfrac{3}{4} & \text{当 } x=2,3 \text{ 时}\end{cases}$$

1) 似然函数

在极大似然估计法中，最关键的问题是如何求得似然函数，有了似然函数，问题就简单了，下面分两种情形来介绍似然函数。

(1) 离散型总体

设总体 X 为离散型，$P\{X=x\}=p(x,\theta)$，其中 θ 为待估计的未知参数，假定 x_1，$x_2,\cdots,x_n$ 为样本 $X_1,X_2,\cdots,X_n$ 的一组观测值。

$$\begin{aligned}P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}&=P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\}\\&=p(x_1,\theta)p(x_2,\theta)\cdots p(x_n,\theta)\\&=\prod_{i=1}^{n}p(x_i,\theta)\end{aligned}$$

将 $\prod_{i=1}^{n}p(x_i,\theta)$ 看作参数 θ 的函数，记为 $L(\theta)$，即

$$L(\theta)=\prod_{i=1}^{n}p(x_i,\theta) \tag{3-12}$$

(2) 连续型总体

设总体 X 为连续型，已知其分布密度函数为 $f(x,\theta)$，θ 为待估计的未知参数，则样本 $(X_1,X_2,\cdots,X_n)$ 的联合密度为

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

将它也看作关于参数 θ 的函数，记为 $L(\theta)$，即

$$L(\theta)=\prod_{i=1}^{n}p(x_i,\theta) \tag{3-13}$$

由此可见，不管是离散型总体，还是连续型总体，只要知道它的概率分布或密度函数，我们总可以得到一个关于参数 θ 的函数 $L(\theta)$，称 $L(\theta)$ 为似然函数。

2) 极大似然估计

极大似然估计法的主要思想是：如果随机抽样得到的样本观测值为 $x_1,x_2,\cdots,x_n$，则我们应当这样来选取未知参数 θ 的值，使得出现该样本值的可能性最大，即使似然函数

$L(\theta)$取最大值，从而求参数θ的极大似然估计的问题就转化为求似然函数$L(\theta)$的极值点的问题，一般来说，这个问题可以通过求解下面的方程来解决：

$$\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}=0 \tag{3-14}$$

然而，$L(\theta)$是n个函数的连乘积，求导数比较复杂，由于$\ln L(\theta)$是$L(\theta)$的单调增函数，所以$L(\theta)$与$\ln L(\theta)$在θ的同一点处取得极大值。于是求解式(3-3)可转化为求解：

$$\frac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=0 \tag{3-15}$$

称$\ln L(\theta)$为对数似然函数，方程(3-15)为对数似然方程，求解此方程就可得到参数θ的估计值。

如果总体X的分布中含有k个未知参数：$\theta_1,\theta_2,\cdots,\theta_k$，则极大似然估计法也适用。此时，所得的似然函数是关于$\theta_1,\theta_2,\cdots,\theta_k$的多元函数$L(\theta_1,\theta_2,\cdots,\theta_k)$，解下列方程组，就可得到$\theta_1,\theta_2,\cdots,\theta_k$的估计值

$$\begin{cases}\dfrac{\partial \ln L(\theta_1,\theta_2,\cdots,\theta_k)}{\partial\theta_1}=0\\[2ex]\dfrac{\partial \ln L(\theta_1,\theta_2,\cdots,\theta_k)}{\partial\theta_2}=0\\[2ex]\vdots\\[2ex]\dfrac{\partial \ln L(\theta_1,\theta_2,\cdots,\theta_k)}{\partial\theta_k}=0\end{cases} \tag{3-16}$$

3.3.2　评价估计量的标准

设总体X服从$[0,\theta]$上的均匀分布，$\hat{\theta}_{矩}=2\overline{X}$，$\hat{\theta}_L \max\limits_{1\leqslant i\leqslant n}\{X_i\}$都是$\theta$的估计，这两个估计哪一个好呢？下面我们首先讨论衡量估计量好坏的标准问题。

1．无偏性

若估计量$(X_1,X_2,\cdots,X_n)$的数学期望等于未知参数θ，即

$$E(\hat{\theta})=\theta \tag{3-17}$$

则称$\hat{\theta}$为θ的无偏估计量(Non-deviation Estimator)。

估计量$\hat{\theta}$的值不一定就是θ的真值，因为它是一个随机变量，若$\hat{\theta}$是θ的无偏估计，则尽管$\hat{\theta}$的值随样本值的不同而变化，但平均来说它会等于θ的真值。

2．有效性

对于未知参数θ，如果有两个无偏估计量$\hat{\theta}_1$与$\hat{\theta}_2$，即$E(\hat{\theta}_1)=E(\hat{\theta}_2)=\theta$，那么在$\hat{\theta}_1$和$\hat{\theta}_2$中谁更好呢？此时我们自然希望$\theta$的平均偏差$E(\hat{\theta}-\theta)^2$越小越好，即一个好的估计量应该有尽可能小的方差，这就是有效性。

设$\hat{\theta}_1$和$\hat{\theta}_2$都是未知参数θ的无偏估计，若对任意的参数θ，有

$$D(\hat{\theta}_1)\leqslant D(\hat{\theta}_2) \tag{3-18}$$

则称$\hat{\theta}_1$比$\hat{\theta}_2$有效。

如果 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效，则虽然 $\hat{\theta}_1$ 还不是 θ 的真值，但 $\hat{\theta}_1$ 在 θ 附近取值的密集程度较 $\hat{\theta}_2$ 高，即用 $\hat{\theta}_1$ 估计 θ 精度要高些。

例如，对正态总体 $N(\mu,\sigma^2)$，$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$，$X_i$ 和 $\overline{X}$ 都是 $E(X)=\mu$ 的无偏估计量，但 $D(X)=\frac{\sigma^2}{n}\leqslant D(X_i)=\sigma^2$，故 $\overline{X}$ 较个别观测值 X_i 有效。实际中也是如此，例如要估计某个班学生的平均成绩，可使用两种方法，一种方法是在该班任意抽取一个同学，以该同学的成绩作为全班的平均成绩；另一种方法是在该班抽取 n 位同学，以这 n 个同学的平均成绩作为全班的平均成绩，显然第二种方法比第一种方法好。

3. 一致性

无偏性、有效性都是在样本容量 n 一定的条件下进行讨论的，然而 $(X_1,X_2,\cdots,X_n)$ 不仅与样本值有关，而且与样本容量 n 有关，记为 n，很自然，我们希望 n 越大时，n 对 θ 的估计应该越精确。如果 n 依概率收敛于 θ，即 $\forall\varepsilon>0$，有

$$\lim_{n\to\infty}P\{|\hat{\theta}_n-\theta|<\varepsilon\}=1 \tag{3-19}$$

则称 $\hat{\theta}_n$ 是 θ 的一致估计量(Uniform Estimator)。

3.3.3 区间估计

1. 区间估计的概念

3.3.1 节介绍了参数的点估计，假设总体 $X\sim N(\mu,\sigma^2)$，对于样本 $(X_1,X_2,\cdots,X_n)$，$\hat{\mu}=\overline{X}$ 是参数 μ 的矩法估计和极大似然估计，并且满足无偏性和一致性。但实际上 $\overline{X}=\mu$ 的可能性有多大呢？由于 $\overline{X}$ 是一连续型随机变量，$P\{X=\mu\}=0$，即 $\hat{\mu}=\mu$ 的可能性为 0，因此，我们希望给出 μ 的一个大致范围，使得 μ 有较高的概率在这个范围内，这就是区间估计问题。

设 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 及 $\hat{\theta}_2(X_1,X_2,\cdots,X_n)$ 是两个统计量，如果对于给定的概率 $1-\alpha(0<\alpha<1)$，有

$$P\{\hat{\theta}_1<\theta<\hat{\theta}_2\}=1-\alpha \tag{3-20}$$

则称随机区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 为参数 θ 的置信区间(Confidence Interval)，$\hat{\theta}_1$ 称为置信下限，$\hat{\theta}_2$ 称为置信上限，$1-\alpha$ 称为置信概率或置信度(Confidence Level)。

定义中的随机区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 的大小依赖于随机抽取的样本观测值，它可能包含 θ，也可能不包含 θ，式(3-9)的意义是指 $(\hat{\theta}_1,\hat{\theta}_2)$ 以 $1-\alpha$ 的概率包含 θ。例如，若取 $\alpha=0.05$，那么置信概率为 $1-\alpha=0.95$，这时，置信区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 的意义是指：在 100 次重复抽样所得到的 100 个置信区间中，大约有 95 个区间包含参数真值 θ，有 5 个区间不包含真值 θ，亦即随机区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 包含参数 θ 真值的频率近似为 0.95。

2. 正态总体参数的区间估计

由于在大多数情况下，所遇到的总体是服从正态分布的(有的是近似正态分布)，下面重点讨论正态总体参数的区间估计问题。在下面的讨论中，总假定 $X\sim N(\mu,\sigma^2)$，X_1，

$X_2,\cdots,X_n$ 为其样本。

分两种情况进行讨论。如果 σ^2 已知，则 μ 的置信区间为 $\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$，置信概率为 $1-\alpha$。

如果 σ^2 未知，不能使用式(3-7)作为置信区间，因为式(3-7)中区间的端点与 σ 有关，考虑到 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$ 是 σ^2 的无偏估计，将 $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 中的 σ 换成 S 得

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

对于给定的 α，由 t 分布分位数表可得上分位点 $t_{\sigma/2}(n-1)$，使得

$$P\left\{\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right|<t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

即

$$P\left\{\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)<\mu<\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right\}=1-\alpha$$

所以 μ 的置信概率为 $1-\alpha$ 的置信区间为

$$\left(\overline{X}-\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1),\overline{X}+\frac{S}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\right)\tag{3-21}$$

由于 $\frac{S}{\sqrt{n}}=\frac{S_0}{\sqrt{n-1}}$，$S_0=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}$，所以 μ 的置信区间也可写成

$$\left(\overline{X}-\frac{S_0}{\sqrt{n-1}}t_{\frac{\alpha}{2}}(n-1),\overline{X}+\frac{S_0}{\sqrt{n-1}}t_{\frac{\alpha}{2}}(n-1)\right)\tag{3-22}$$

以上仅介绍了正态总体的均值和方差两个参数的区间估计方法。在有些问题中并不知道总体 X 服从什么分布，要对 $E(X)=\mu$ 作区间估计，在这种情况下只要 X 的方差 σ^2 已知，并且样本容量 n 很大，由中心极限定理可知，$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ 近似地服从标准正态分布 $N(0,1)$，因而 μ 的置信概率为 $1-\alpha$ 的近似置信区间为 $\left(\overline{X}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$。

本章小结

本章从概率论基础知识开始，介绍了随机变量及随机变量的类型、性质等，并且详细介绍了离散型、连续型随机变量概率的 Python 实现。接着详细讲解了数理统计的基础知识，包括统计学的总体与样本、统计量的各种概念定义及三大抽样分布及其 Python 实现。同时对统计学两大基石——大数定律和中心极限定理进行了详细介绍，并给出对应的 Python 实现。最后对数理统计中参数的点估计与区间估计方法从理论层面进行详细讲解。

课后思考题

1. 什么是概率与条件概率？
2. 简述大数定律与中心极限定理。
3. 统计学中三大抽样分布与正态分布之间的关系是什么？
4. 简述评价估计量好坏的标准。
5. 简述样本量与置信水平、总体方差、估计误差之间的关系。
6. 什么是假设检验中的显著性水平？统计显著是什么意思？