第 5 章

统计方法与分析

从统计方法角度可将统计分为描述统计和推断统计。前者主要研究如何取得、整理和表现数据资料，进而通过综合、概括与分析反映客观现象的数量特征，包括数据的收集与整理、数据的显示方法、数据分布特征的描述与分析方法等；后者研究如何根据样本数据推断总体数量特征的方法，包括参数估计、假设检验、方差分析及回归分析等。描述统计是统计研究工作的前提，推断统计则是统计研究工作的核心和关键。

5.1 常见随机变量的分布

常见随机变量的分布主要包括离散型随机变量分布与连续型随机变量分布。

5.1.1 离散型随机变量分布

一个随机变量的取值如果是有限个或可列无限个，则称这个随机变量为离散型随机变量。对一个离散型随机变量 X，其所有可能取值记为 $x_1, x_2, \cdots, x_n, \cdots$，并把每个可能取值的概率记为 $P\{X=x_i\}=p_i$，其中 $i=1,2,3,\cdots$，称 $P\{X=x_i\}=p_i$ 为随机变量 X 的概率分布或分布律。

通常用表格形式来表示离散型随机变量的概率分布，如表 5-1 所示。

表 5-1 离散型随机变量的概率分布

X	x_1	x_2	$\cdots$	x_n	$\cdots$
p_i	p_1	p_2	$\cdots$	p_n	$\cdots$

离散型随机变量由其所有可能取值及取这些值的概率唯一确定，且概率分布具有如下性质。

(1) $P\{X=x_k\}=p_k \geqslant 0$。

(2) $\sum\limits_{k=1}^{\infty} P\{X=x_i\}=\sum\limits_{k=1}^{\infty} p_k=1$。

下面介绍常用的离散型随机变量的分布。

1. 0-1 分布

如果随机变量 X 的分布律为 $P\{X=0\}=1-p$，$P\{X=1\}=p$，$0<p<1$，则称随机变量

X 服从参数为 p 的 0-1 分布，也称为伯努利分布。

0-1 分布也可记作 $P\{X=k\}=p^k(1-p)^{n-k}(k=0,1)$，如表 5-2 所示。

表 5-2　0-1 分布的概率分布

X	0	1
P	$1-p$	p

记为 $X\sim b(1,p)$，其中 $0\leqslant p\leqslant 1$。

2. 二项分布

如果随机变量 X 的分布律为 $P\{X=k\}=C_n^k p^k(1-p)^{n-k}$，$k=0,1,\cdots,n$，则称随机变量 X 服从参数为(n,p)的二项分布，记为 $X\sim b(n,p)$。其中参数 n 表示实验的总次数，p 表示每次实验事件发生的概率，且满足如下性质。

(1) $P\{X=k\}\geqslant 0$，$k=0,1,\cdots,n$。

(2) $\sum\limits_{k=0}^{n}P\{X=k\}=\sum\limits_{k=0}^{\infty}C_n^k p^k(1-p)^{n-k}=(p+1-p)^n=1$。

显然，当 $n=1$ 时 $X\sim b(1,p)$，说明伯努利分布是二项分布的一个特例。

3. 泊松分布

如果随机变量 X 的分布律为

$$P\{X=k\}=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\quad(k=0,1,2,\cdots)$$

其中，参数 $\lambda>0$，则称随机变量 X 服从参数为 λ 的泊松分布，记为 $X\sim\pi(\lambda)$，且满足如下性质。

(1) $P\{X=k\}\geqslant 0$，$k=0,1,2,\cdots$。

(2) $\sum\limits_{k=0}^{\infty}P\{X=k\}=\sum\limits_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}=\mathrm{e}^{-\lambda}\sum\limits_{k=0}^{\infty}\frac{\lambda^k}{k!}=\mathrm{e}^{-\lambda}\mathrm{e}^{\lambda}=1$。

泊松分布是概率论中最重要的分布之一，实际问题中许多随机现象都服从或近似服从泊松分布。泊松分布往往适用于在单位时间、空间、区域或连续时间间隔内事件发生次数的概率。例如，电话总机在某一时间间隔内收到的呼叫次数、放射物在某一时间间隔内发射的粒子数、容器在某一时间间隔内产生的细菌数、某一时间间隔内来到某服务台要求服务的人数等。

5.1.2　连续型随机变量分布

如果对随机变量 X 的分布函数 $F(x)$，存在非负函数 $f(x)$，使得对于任意实数 x 有 $F(x)=P\{X\leqslant x\}=\int_{-\infty}^{x}f(t)\mathrm{d}t$，则称 X 为连续型随机变量，称 $f(x)$ 为 X 的概率密度函数，简称为概率密度或密度函数。概率密度 $f(x)$ 具有如下性质。

(1) $f(x)\geqslant 0$。

(2) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$。

(3) $P\{x_1 < X \leqslant x_2\} = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\mathrm{d}x$，$x_1 \leqslant x_2$。

(4) 若 $f(x)$在点 x 处连续，则有 $F'(x) = f(x)$。

下面介绍常用的连续型随机变量分布。

1. 均匀分布

若连续型随机变量 X 的概率密度为

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

则称 X 在区间(a,b)上服从均匀分布，$X \sim U(a,b)$。其分布函数为

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

2. 指数分布

若随机变量 X 的概率密度为 $f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x > 0 \\ 0, & \text{其他} \end{cases}$，$\lambda > 0$，则称 X 服从参数为 λ 的指数分布，简记为 $X \sim \exp(\lambda)$。

指数分布的分布函数为

$$F(x) = \begin{cases} 0, & x \leqslant 0 \\ 1 - \mathrm{e}^{-\lambda x}, & x > 0 \end{cases}$$

3. 正态分布

若随机变量 X 的概率密度为 $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$，$-\infty < x < \infty$，其中 μ 和 $\sigma(\sigma > 0)$都是常数，则称服从参数为 μ 和 σ^2 的正态分布，记为 $X \sim N(\mu, \sigma^2)$。

正态分布是概率论中最重要的连续型分布，在 19 世纪前叶由高斯加以推广，故又常称为高斯分布。一般来说，一个随机变量如果受到许多随机因素的影响，而其中每个因素都不起主导作用(作用微小)，则它服从正态分布。例如，产品的质量指标，元件的尺寸，某地区成年男子的身高、体重，测量误差，射击目标的水平或垂直偏差，信号噪声，农作物的产量等，都服从或近似服从正态分布。

正态分布密度函数具有如下性质。

(1) 曲线关于直线 $x = \mu$ 对称，这表明对于任意的 $h > 0$，有

$$P\{\mu - h < X \leqslant \mu\} = P\{\mu < X \leqslant \mu + h\}$$

(2) 当 $x = \mu$ 时，$f(x)$取到最大值 $f(\mu) = 1/\sqrt{2\pi}\sigma$，离 μ 越远，$f(x)$的值就越小。这表

明，对于同样长度的区间，当区间离 μ 越远时，随机变量 X 落在该区间中的概率就越小。

(3) 曲线 $y=f(x)$ 在 $x=\mu\pm\sigma$ 处有拐点；曲线 $y=f(x)$ 以 Ox 轴为渐近线。

(4) 若 σ 固定，改变 μ 的值，则 $f(x)$ 的图形沿 x 轴平行移动，但不改变其形状，因此 $y=f(x)$ 图形的位置完全由参数 μ 确定。

(5) 若 μ 固定，改变 σ 的值，则 σ 越大，$f(x)$ 的图形越陡，随机变量 X 落在 μ 附近的概率就越大；反之，X 的取值越分散。

当 $\mu=0$，$\sigma=1$ 时正态分布称为标准正态分布，此时其密度函数和分布函数常用 $\varphi(x)$ 和 $\phi(x)$ 表示，即

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}},\quad \phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$$

标准正态分布的重要性在于，任何一个一般的正态分布都可以通过线性变换转换为标准正态分布。

例 5.1　若随机变量 $X\sim N(-3,0.9^2)$，试求该正态随机变量概率密度函数包含 95% 取值的区间。

解

```
#Python 代码
import numpy as np
from scipy import stats
mu = -3
sigma = 0.9
myDistribution = stats.norm(mu, sigma)
P = 0.95
significanceLevel = 1 - P
myDistribution.ppf([significanceLevel/2, 1 - significanceLevel/2])
Out[1]: array([ -4.76396759, -1.23603241])
```

对 $X\sim N(\mu,\sigma^2)$，可计算出

$$P\{\mu-k\sigma<X<\mu+k\sigma\}=P\left\{-k<\frac{X-\mu}{\sigma}<k\right\}=\phi(k)-\phi(-k)=2\phi(k)-1$$

5.1.3　其他连续型分布

1. 对数正态分布

设 X 是取值为正数的连续随机变量，若 $\ln X\sim N(\mu,\sigma^2)$，$X$ 的概率密度为

$$f(x,\mu,\sigma)=\begin{cases}\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{1}{2\sigma^2}(\ln x-\mu)^2\right], & x>0\\ 0, & x\leqslant 0\end{cases}$$

则称随机变量 X 服从对数正态分布，记为 $\ln X\sim N(\mu,\sigma^2)$。

2. 韦布尔分布

韦布尔分布又称韦氏分布或威布尔分布，是可靠性分析和寿命检验的理论基础。韦布

尔分布在可靠性工程中被广泛应用，尤其适用于机电类产品的磨损累计失效的分布形式。由于它可以利用概率值容易地推断分布参数，因此被广泛应用于各种寿命试验的数据处理。韦布尔分布的概率密度为

$$f(x;\lambda,k)=\begin{cases}\dfrac{k}{\lambda}\left(\dfrac{x}{\lambda}\right)^{k-1}\mathrm{e}^{-(x/\lambda)^k}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

其中，$\lambda>0$ 是比例参数，$k>0$ 是形状参数。显然，它的累积分布函数是扩展的指数分布函数，而且韦布尔分布与许多分布都有关系。例如，当 $k=1$ 时，它是指数分布；当 $k=2$ 时，它是瑞利分布。

3. 抽样分布

统计量的分布为抽样分布，常见的抽样分布有如下三种。

1) χ^2 分布

设 $X_1,X_2,\cdots,X_n$ 是取自总体 $N(0,1)$ 的样本，则称统计量 $\chi^2=X_1^2+X_2^2+\cdots+X_n^2$ 为服从自由度为 n 的 χ^2 分布，记为 $\chi^2\sim\chi^2(n)$。$\chi^2(n)$ 分布的概率密度为

$$f(x)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{1}{2}x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中，$\Gamma(\cdot)$ 为 Gamma 函数，$\Gamma(\alpha)=\int_0^{+\infty}x^{\alpha-1}\mathrm{e}^{-x}\,\mathrm{d}x$，$\alpha>0$。

2) t 分布

设 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立，则称 $T=\dfrac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布，记为 $t\sim t(n)$。$t(n)$ 分布的概率密度为

$$f(x)=\frac{\Gamma[(n+1)/2]}{\sqrt{\pi n}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\quad -\infty<t<+\infty$$

t 分布具有如下性质。

(1) $f(x)$ 的图形关于纵轴对称，且 $\lim\limits_{x\to\infty}f(x)=0$。

(2) 当 n 充分大时，t 分布近似于标准正态分布。

3) F 分布

设 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且 X 与 Y 相互独立，则称 $F=\dfrac{X/n_1}{Y/n_2}$ 服从自由度为 (n_1,n_2) 的 F 分布，记为 $F\sim F(n_1,n_2)$。$F(n_1,n_2)$ 分布的概率密度为

$$f(x)=\begin{cases}\dfrac{\Gamma[(n_1+n_2)/2](n_1/n_2)^{n_1/2}x^{(n_1/2)-1}}{\Gamma(n_1/2)\Gamma(n_2/2)[1+(n_1x/n_2)]^{(n_1+n_2)/2}}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

F 分布具有如下性质。

(1) 若 $X\sim t(n)$，则 $X^2\sim F(1,n)$。

(2) 若 $F\sim F(n_1,n_2)$，则 $\frac{1}{F}\sim F(n_2,n_1)$。

5.2　描述统计

描述统计是指运用制表、分类、图形及概括性数据来描述数据特征的各项活动，是社会科学研究中的常用方法。

5.2.1　统计量及数据分布特征

统计分析主要分为两类，一类是对收集的数据整理后进行描述性统计，另一类是根据所得的数据资料进行分析、研究，从而对所研究对象的性质、特点做出推断，即推断统计。我们在介绍描述统计前先介绍样本的相关概念。

统计中，人们都是通过从总体中抽取一部分个体，然后根据获得的数据来对总体分布得出推断的，被抽出的部分个体叫作总体的一个样本。为了使样本能很好地反映总体的情况，需满足两个条件：每个个体和总体具有相同分布，即代表性；每次抽样的结果既不影响其他各次抽样的结果，也不受其他各次抽样结果的影响，即独立性。这种随机的、独立的抽样方法称为简单随机抽样。由此得到的样本称为简单随机样本。

设总体 X 是具有某一分布函数的随机变量，如果随机变量 $X_1,X_2,\cdots,X_n$ 相互独立，且都与 X 具有相同的分布，则称 $X_1,X_2,\cdots,X_n$ 为来自总体 X 的简单随机样本，简称样本。n 称为样本容量。

样本是进行统计推断的依据。在应用时，往往不是直接使用样本本身，而是针对不同的问题构造样本的适当函数，即统计量。

设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的一个样本，$g(X_1,X_2,\cdots,X_n)$ 是 $X_1,X_2,\cdots,X_n$ 的函数，若 g 中不含未知参数，则称 $g(X_1,X_2,\cdots,X_n)$ 是一个统计量。

常见的统计量有以下几种。

(1) 样本均值：$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$。

(2) 样本方差：$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)$。

(3) 样本标准差：$s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$。

(4) 样本(k 阶)原点矩：$a_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,k=1,2,3,\cdots$。

(5) 样本(k 阶)中心矩：$b_k=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^k$，$k=1,2,3,\cdots$。

数据分析中样本构造的函数往往反映了数据分布的形状和特征，这可以从三方面进行测度和描述：一是分布的集中趋势，反映各数据向其中心值靠拢或聚集的程度；二是分布的离散程度，反映各数据远离其中心值的趋势；三是分布的形状，反映数据分布的偏态和峰态。

下面给出在统计分析中常用的一些反映数据分布特征的指标。

众数(mode)是一组数据中出现次数最多的变量值。众数也适用于作为顺序数据及数值型数据集中趋势的测度值。一般情况下，只有在数据量较大的情况下，众数才有意义。

样本中位数也是一个很常见的统计量，它是次序统计量的函数，通常定义为

$$m_{0.5}=\begin{cases}x_{\left(\frac{n+1}{2}\right)}, & n\text{ 为奇数}\\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right), & n\text{ 为偶数}\end{cases}$$

中位数将全部数据等分成两部分，每部分包含 50%的数据，一部分数据比中位数大，另一部分则比中位数小。中位数主要用于测度顺序数据的集中趋势，当然也适用于测度数值型数据的集中趋势，但不适用于分类数据。根据未分组数据计算中位数时，要先对数据进行排序，然后确定中位数的位置，最后确定中位数的具体数值。与中位数类似的还有四分位数等，一般地，p 分位数 m_p 定义为

$$m_p=\begin{cases}x_{(np+1)}, & np\text{ 不是整数}\\ \dfrac{1}{2}(x_{(np)}+x_{(np+1)}), & np\text{ 是整数}\end{cases}$$

样本极差定义为 $R_n=x_{(n)}-x_{(1)}$，样本中程定义为 $[x_{(n)}-x_{(1)}]/2$。样本极差是一个很常用的统计量，其分布只在很少场合可用初等函数表示。

方差和标准差是反映数据离散程度的绝对值，它们与原变量值的计量单位相同，采用不同计量单位计量的变量值，其离散程度的测度值也就不同。故对于计量单位不同的不同组别的变量值，不能用标准差直接比较其离散程度。为消除变量值水平高低和计量单位不同对离散程度的影响，需要计算离散系数。离散系数也称为变异系数(coefficient of variation)，它是一组数据的标准差与其相应的平均数之比，其计算公式为

$$v=s/\bar{x}$$

离散系数是测度数据离散程度的相对统计量，主要用于比较不同样本数据的离散程度。离散系数大，说明数据的离散程度也大；离散系数小，说明数据的离散程度也小。

要全面了解数据分布的特点，还需要知道数据分布的形状是否对称、偏斜的程度及分布的扁平程度等。偏态和峰态就是对分布形状的测度。

偏态是对数据分布对称性的测度。测度偏态的统计量是偏态系数，计算公式通常为

$$\gamma_1=\frac{b_3}{b_2^{3/2}}$$

如果一组数据的分布是对称的，则偏态系数等于0；如果偏态系数明显不等于0，则分布为非对称的。

峰态是对数据分布平峰或尖峰程度的测度。如果一组数据服从标准正态分布，则峰态系数的值等于0；若峰态系数的值明显不等于0，则表明分布比正态分布更平或更尖，通常称为平峰分布或尖峰分布。样本峰度反映了总体分布密度曲线在其峰值附近的陡峭程度。测度峰态的统计量是峰态系数，计算公式通常为

$$\gamma_2 = \frac{b_4}{b_2^2}$$

Python中的NumPy和SciPy库可以作为描述性统计的工具，例如

```
from numpy import array
from numpy.random import normal, randint
#使用List来创造一组数据
data = [1, 2, 3]
#使用ndarray来创造一组数据
data = array([1, 2, 3])
#创造一组服从正态分布的定量数据
data = normal(0, 10, size=10)
#创造一组服从均匀分布的定性数据
data = randint(0, 10, size=10)
```

描述性统计常用命令参见表5-3。

表5-3　描述性统计常用命令

库	方　法	功　能
numpy	array	创造一组数
numpy.random	normal	创造一组服从正态分布的定量数
numpy.random	randint	创造一组服从均匀分布的定性数
numpy	mean	计算均值
numpy	median	计算中位数
scipy.stats	mode	计算众数
numpy	ptp	计算极差
numpy	var	计算方差
numpy	std	计算标准差
numpy	cov	计算协方差
numpy	corrcoef	计算相关系数

除NumPy外，Python另一个强大的数据分析工具是Pandas，它基于NumPy且提供了大量高效处理数据的函数和方法。

例5.2　表5-4所示的数据为某高校42名学生身高，存放在文件“height.csv”中。根据这些数据来计算相关统计量。

表 5-4　42 名学生身高数据

序　号	身　高	序　号	身　高	序　号	身　高
1	189	15	183	29	182
2	170	16	193	30	188
3	189	17	178	31	175
4	163	18	173	32	179
5	183	19	174	33	183
6	171	20	183	34	193
7	185	21	183	35	182
8	168	22	168	36	183
9	173	23	170	37	177
10	183	24	178	38	185
11	173	25	182	39	188
12	173	26	180	40	188
13	175	27	183	41	182
14	178	28	178	42	185

解

```
#Python 代码
#spyder 脚本中运行
import numpy as np
import pandas as pd
data = pd.read_csv('D:\\ height.csv')
#读入文件后,在 IPython console 中运行
```

继续输入代码来提取身高所在的列数据。

```
heights = np.array(data['身高'])
print(heights)
```

结果如下。

```
[189  170  189  163  183  171  185  168  173  183  173  173  175  178  183
 193  178  173  174  183  183  168  170  178  182  180  183  178  182  188
 175  179  183  193  182  183  177  185  188  188  182  185]
```

对取到的数据进行描述性统计。

```
print("身高平均值: ", np.mean(heights))
print("身高方差:",np.var(heights))
print("身高标准差:",np.std(heights))
print("最小身高: ", np.min(heights))
print("最大身高: ", np.max(heights))
print("身高极差: ", np.ptp(heights))
print("四分之一分位数: ", np.percentile(heights, 25))
print("中位数: ", np.median(heights))
print("四分之三分位数: ", np.percentile(heights, 75))
print("众数: ", pd.Series(heights).mode())
```

```
print("偏度: ", pd.Series(heights).skew())
print("峰度: ", pd.Series(heights).kurt())
print("50 % 分位数(中位数): ", pd.Series(heights).quantile())
```

输出结果如下。

```
身高平均值: 179.73809523809524
身高方差: 48.05045351473922
身高标准差: 6.931843442745892
最小身高: 163
最大身高: 193
身高极差: 30
四分之一分位数: 174.25
中位数: 182.0
四分之三分位数: 183.0
众数: 183
偏度: -0.2520337007049649
峰度: -0.3823273337498865
50 % 分位数(中位数): 182.0
```

例 5.3　利用 pandas 对初始数据集[1,1,1,1,1,2,2,2,4]进行简单的统计分析。

解

```
#Python 代码
import pandas
data = [1, 1, 1, 1, 1, 2, 2, 2, 4]
df = pandas.DataFrame(data, columns = ['value'])
print(df.describe())
```

输出结果如下。

```
value
count    9.000000
mean     1.666667
std      1.000000
min      1.000000
25 %     1.000000
50 %     1.000000
75 %     2.000000
max      4.000000
```

5.2.2　Python 统计可视化

描述性统计中除用一些数据分布特征指标进行度量外,还可以用更加直观的统计图来描述。

直方图是频数分布的图形表示,它的横坐标表示所关心变量的取值区间,纵坐标有三种表示方法:频数、频率及最准确的频率/组距,它可使得诸长条矩形面积和为 1。这三种直方图的差别仅在于纵轴刻度的选择,直方图本身并无变化。

茎叶图于 20 世纪早期由英国统计学家 Arthur Bowley 设计,其思路是把每个数值分为两部分,前面部分(百位和十位)称为茎,后面部分(个位)称为叶,然后画一条竖线,在竖线的

左侧写上茎，右侧写上叶，就形成了茎叶图。例如，213 的茎叶图如表 5-5 所示。

表 5-5　213 的茎叶图

数　值	分　开	茎	叶
213	21|3	21	3

茎叶图表示数据的好处是从统计图上没有原始数据信息的损失，所有数据信息都可以从茎叶图中得到；此外，茎叶图中的数据可以随时记录，随时添加，方便记录与表示。但是茎叶图只便于表示个位之前相差不大的数据，当样本量较大、数据很分散、横跨二三个数量级时，茎叶图并不适用。

箱线图是一种用作显示一组数据分散情况的统计图，于 1977 年由美国著名统计学家 John Tukey 发明，因形状如箱子而得名。箱线图主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。箱线图的绘制方法是：先找出一组数据的上边缘、下边缘、中位数和两个四分位数；然后连接两个四分位数画出箱体；再将上边缘和下边缘与箱体相连接，中位数在箱体中间。即用五分位数概括：最小观测值 $x_{\min}=x_{(1)}$，最大观测值 $x_{\max}=x_{(n)}$，中位数 $m_{0.5}$，第一四分位数 $Q_1=m_{0.25}$，第三四分位数 $Q_3=m_{0.75}$。

直方图、茎叶图、箱线图是传统意义上常见的统计图，实际数据分析中根据不同需求可以有多种展示，即对数据进行可视化。制作可视化的工具很多，如 R、Python、MATLAB 等。这里结合 Python 中的库来介绍数据可视化的制作。

例 5.4　数据文件“exfilms.csv”存储了四部电影《阿飞正传》《建党伟业》《西虹市首富》《建军大业》上映后的共 85 条相关数据，如表 5-6 所示。

表 5-6　电影上映后的数据

序　号	电影名称	当日票房/百万元	排片场次	场均人次	上座率	时　间
1	阿飞正传	1.48	13	36	0.3	2018/6/24 0：00
2	阿飞正传	318.57	6570	15	0.153	2018/6/25 0：00
3	阿飞正传	256.28	6743	12	0.114	2018/6/26 0：00
4	阿飞正传	216.53	6867	10	0.096	2018/6/27 0：00
5	建党伟业	1.69	14	41	0.35	2018/6/27 0：00
6	阿飞正传	183.56	6813	9	0.086	2018/6/28 0：00
7	建党伟业	2.72	9	104	0.241	2018/6/28 0：00
8	阿飞正传	124.97	2841	13	0.128	2018/6/29 0：00
9	建党伟业	1.37	9	49	0.481	2018/6/29 0：00
10	阿飞正传	177.55	3119	17	0.161	2018/6/30 0：00
11	阿飞正传	139.76	3145	13	0.13	2018/7/1 0：00
…	…	…	…	…	…	…
83	西虹市首富	374.71	10151	11	0.132	2018/8/29 0：00
84	西虹市首富	377.87	10767	11	0.126	2018/8/30 0：00
85	西虹市首富	149.83	3477	14	0.174	2018/8/31 0：00

试对影片《建党伟业》每日场均人次绘制散点图。

解

```
#Python 代码
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
data = pd.read_csv("exfilms.csv",encoding = "gbk")
data["时间"] = pd.to_datetime(data["时间"])
df = data[(data["电影名称"] == "建党伟业")]
```

变量 df 将原始数据中的影片《建党伟业》单独筛选出来，其输出结果如下。

```
In [61]: df
Out[61]:
电影名称 当日票房 排片场次 场均人次 上座率 时间
4 建党伟业 1.69 14 41 0.35 2018-06-27
6 建党伟业 2.72 9 104 0.241 2018-06-28
8 建党伟业 1.37 9 49 0.481 2018-06-29
13 建党伟业 3.81 6 210 0.833 2018-07-03
17 建党伟业 1.38 10 38 0.467 2018-07-06
23 建党伟业 1.41 17 27 0.355 2018-07-11
34 建党伟业 1.23 4 122 0.804 2018-07-20
```

取得电影《建党伟业》的数据后，以放映日期为横轴，以场均人次为纵轴绘制散点图，代码如下。

```
fig,ax = plt.subplots(figsize = (10,5))
plt.scatter(df["时间"],df["场均人次"])
plt.xticks(rotation = 30)
plt.title("《建党伟业》上映后场均人次散点图", fontproperties = 'simhei', fontsize = 18)
plt.grid(False)
plt.ylabel('场均人次',fontproperties = 'simhei', fontsize = 16)
plt.xlabel('时间',fontproperties = 'simhei', fontsize = 16)
plt.show()
```

最终生成的散点图如图 5-1 所示。

注　散点图 scatter 函数的参数可以设定诸如点的颜色、大小、形状、透明度等，其完整格式为

```
matplotlib.pyplot.scatter(x, y, s = None, c = None, marker = None, cmap = None, norm = None, vmin = None, vmax = None, alpha = None, linewidths = None, verts = None, edgecolors = None, * , data = None, ** kwargs)
```

例如，参数 s 表示点的大小；c 表示点的颜色；marker 表示点的形状等。在本例中通过下面的代码可以生成不同的散点图。

```
plt.scatter(df["时间"],df["场均人次"],c = 'b',s = 30,marker = 'v')
plt.xticks(rotation = 30)
plt.title("《建党伟业》上映后场均人次散点图", fontproperties = 'simhei', fontsize = 18)
plt.grid()
plt.ylabel('场均人次',fontproperties = 'simhei', fontsize = 16)
```

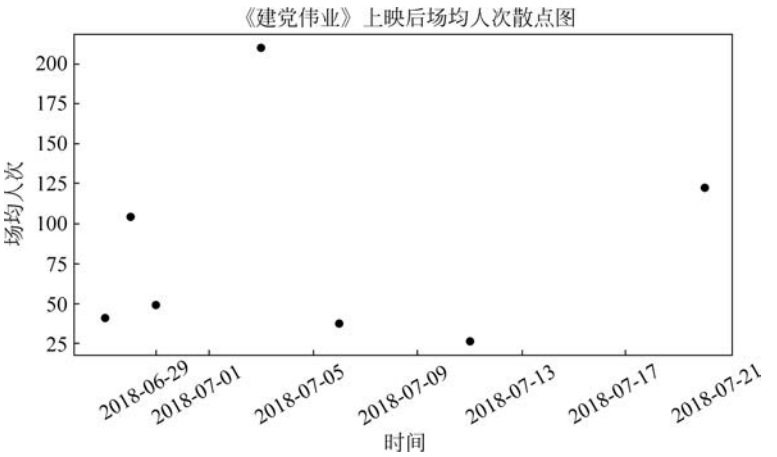

图 5-1　scatter 方法散点图

```
plt.xlabel('时间',fontproperties = 'simhei', fontsize = 16)
```

运行结果如图 5-2 所示。

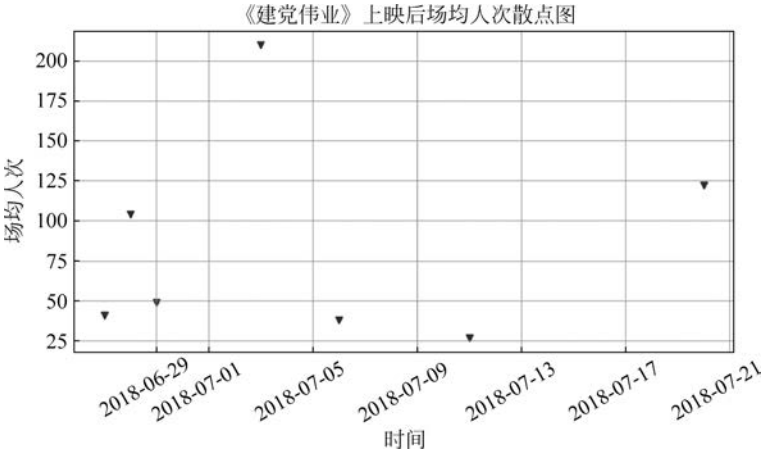

图 5-2　不同参数的 scatter 方法散点图

例 5.5　利用例 5.4 中数据文件“exfilms.csv”，绘制影片《阿飞正传》上映后的场均人次折线图。

解

```
#Python 代码
df = data[(data["电影名称"] == "阿飞正传")]#前述代码同例 5.4
fig,ax = plt.subplots(figsize = (10,5))
plt.plot(df["时间"],df["场均人次"])
plt.xticks(rotation = 30)
plt.title("《阿飞正传》场均人次折线图", fontproperties = 'simhei', fontsize = 18)
plt.show()
```

运行结果如图 5-3 所示。

图 5-3　折线图

注：plot 函数参数除表示坐标轴数据外，还可以设置曲线的颜色、线型和标记样式等。颜色常用的值有“r/g/b/c/m/y/k/w”，如 r 表示红色，b 表示蓝色，k 表示黑色等；线型常用的值有“－/－－/：/－.”，如－表示实线，：表示点线；标记样式常用的值有“./,/o/v/^/s/*/D…”，如 v 表示三角标识，s 表示方块等。

单独运用标记样式参数，效果与散点图类似。

```
#Python 代码
plt.plot(df["时间"],df["场均人次"],'*') #定义标记样式
plt.xticks(rotation=30)
plt.title("《阿飞正传》场均人次折线图", fontproperties = 'simhei', fontsize=18)
plt.ylabel('场均人次',fontproperties = 'simhei', fontsize = 16)
plt.xlabel('时间',fontproperties = 'simhei', fontsize = 16)
plt.show()
```

运行结果如图 5-4 所示。

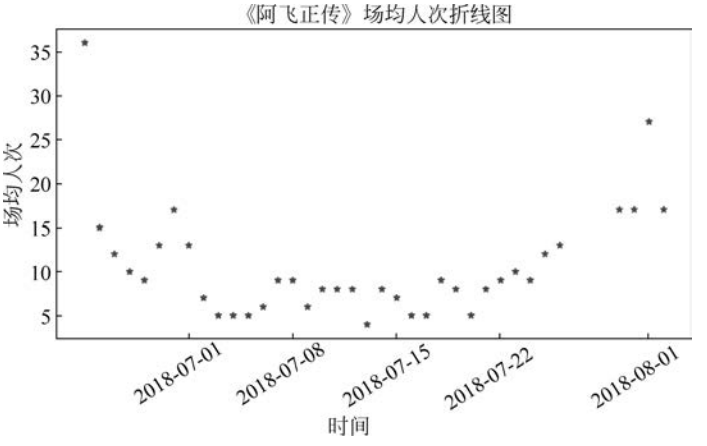

图 5-4　标记图

plot 函数参数中线型、颜色及标记可组合运用，例如

```
plt.plot(df["时间"],df["场均人次"],'-- *') #定义标记样式
```

后续代码同上，运行结果如图 5-5 所示。

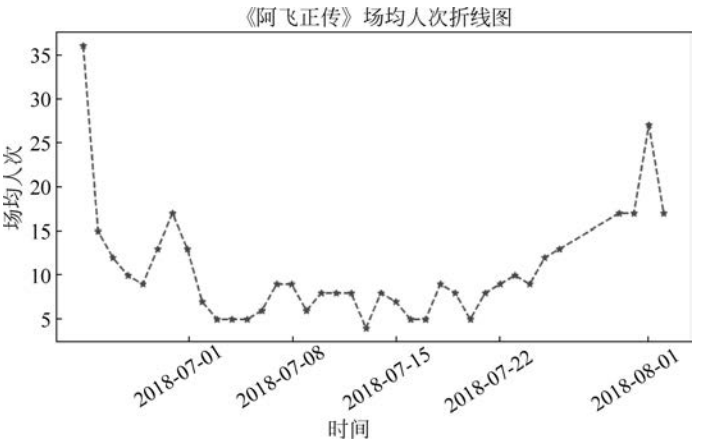

图 5-5　不同参数生成的折线图

例 5.6　利用例 5.4 中数据文件“exfilms.csv”，绘制文件中日期范围内四部影片平均上座率的柱状图。

解

```
#Python 代码
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
data = pd.read_csv("exfilms.csv",encoding = "gbk")
data["时间"] = pd.to_datetime(data["时间"])
data["上座率"] = pd.to_numeric(data["上座率"], errors = 'coerce').fillna(0)
# 文件中"上座率"为非数值类型,需转换;同时部分数据缺失,用 0 填充
df1 = data[data["电影名称"] == "阿飞正传"]['上座率'].mean()
df2 = data[(data["电影名称"] == "建党伟业")]['上座率'].mean()
df3 = data[(data["电影名称"] == "西虹市首富")]['上座率'].mean()
df4 = data[(data["电影名称"] == "建军大业")]['上座率'].mean()
fig,ax = plt.subplots(figsize = (10,6))
film_n = [1,2,3,4]
shangzuo = [df1,df2,df3,df4]
for x,y in zip(film_n,shangzuo):
    plt.bar(x,y,width = 0.5)
plt.ylabel('平均上座率', fontproperties = 'simhei',fontsize = 16)
plt.title('四部影片平均上座率', fontproperties = 'simhei', fontsize = 19)
film_name = ['阿飞正传','建党伟业','西虹市首富','建军大业']
plt.xticks(list(range(1,5)),
           film_name,
           #color = 'green',
           fontproperties = 'simhei',
```

```
            fontsize = 16,rotation = 0)
#plt.grid(False)
plt.show()plt.show()
```

运行结果如图 5-6 所示。

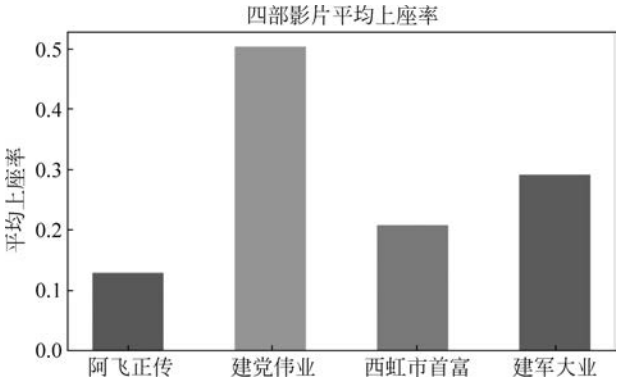

图 5-6　柱状图示例

例 5.7　目前国内汽车销售市场上 SUV 车型销售火爆，表 5-7 数据统计了 2020 年的某三个月中某一国产品牌和合资品牌各自的销售量，试绘制柱状图以直观比较。

表 5-7　2020 年国内 SUV 销售量

品　　牌	5 月	6 月	7 月
国产品牌 A	22691	23258	23723
合资品牌 B	18212	11217	14170
合资品牌 C	15910	15898	17967

解

```
#Python 代码
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
fig,ax = plt.subplots(figsize = (10,6))
df  =  pd.DataFrame({'国产品牌 A':(22691,23258,23723),
                     '合资品牌 B':(18212,11217,14170),
                     '合资品牌 C':(15910,15898,17967)})
index = np.arange(3)
plt.bar(index, df['国产品牌 A'],label = "国产品牌 A",width = 0.2)
plt.bar(index + 0.2, df['合资品牌 B'], label = "合资品牌 B", width  =  0.2)
plt.bar(index + 0.4, df['合资品牌 C'],label = "合资品牌 C", width  =  0.2)
plt.xticks([0.2,1.2,2.2],['5 月', '6 月', '7 月'],
           fontproperties = 'simhei', fontsize = 16,rotation = 20)
plt.ylabel('销量', fontproperties = 'simhei', fontsize = 16)
plt.title('SUV 销量', fontproperties = 'simhei', fontsize = 20)
plt.ylim(0,29000)
```

```
font = fm.FontProperties(fname = r'C:\Windows\Fonts\simkai.ttf')
plt.legend(prop = font)
plt.show()
```

运行结果如图 5-7 所示。

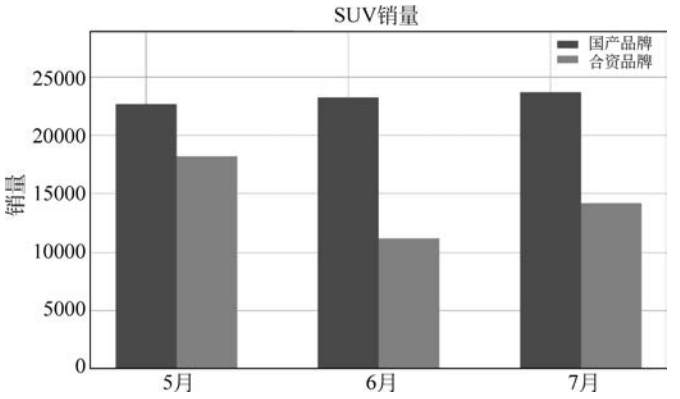

图 5-7　柱状图

例 5.8　根据例 5.7 中国内汽车销售数据，绘制不同品牌的汽车销售雷达图。

解

```
#Python 代码
import random
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.font_manager as fm
data = {'国产品牌 A': [22691,23258,23723],
'合资品牌 B': [18212,11217,14170],
'合资品牌 C': [15910,15898,17967]}
dataLength = len(data['国产品牌 A'])
angles = np.linspace(0, 2 * np.pi, dataLength, endpoint = False)
markers = '* v^so'
for col in data.keys():
color = '#' + ''.join(map('{0:02x}'.format,
np.random.randint(0,255,3)))
plt.polar(angles, data[col], color = color,
marker = random.choice(markers), label = col)
plt.thetagrids(angles[:12] * 180/np.pi,
list(map(lambda i:'%d 月' % i, range(5,9))),
fontproperties = 'simhei')
font = fm.FontProperties(fname = r'C:\Windows\Fonts\simkai.ttf')
plt.legend(prop = font)
plt.show()
```

运行结果如图 5-8 所示。

例 5.9　给出一组数据，绘制其直方图。数据量比较少时直方图效果不明显，为演示用，此处利用正态分布随机数进行模拟。

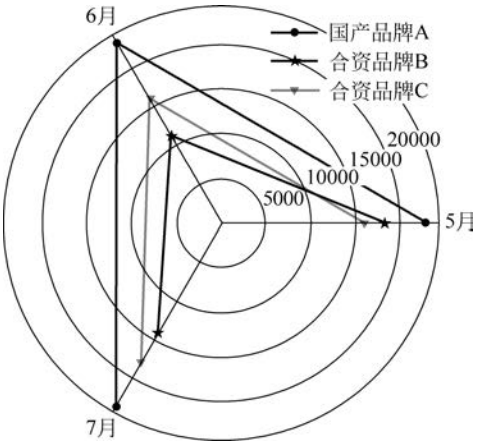

图 5-8　雷达图

解

```
#Python 代码
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style = "ticks")
plt.rcParams['axes.unicode_minus'] = False
fig,ax = plt.subplots(figsize = (10,7))
x = np.random.randn(1000)
plt.hist(x,bins = 15)
plt.title("正态分布随机数直方图",fontproperties = 'simhei', fontsize = 18)
plt.xlabel("随机变量取值",fontproperties = 'simhei', fontsize = 16)
plt.ylabel("频数",fontproperties = 'simhei', fontsize = 16)
plt.show()
```

运行结果如图 5-9 所示。

例 5.10　绘制数据[335,232,232,245,235,338,341,225,241,240,172,222,170,220,231,334]的茎叶图。

解

```
#Python 代码
from itertools import groupby
data = [335, 232, 232, 245, 235, 338, 341, 225, 241, 240, 172, 222, 170, 220, 231, 334]
for i, j in groupby(sorted(data), key = lambda x: int(x) // 10):
    lst = map(str, [int(y) % 10 for y in list(j)])
    print (i, '|', ' '.join(lst))
```

运行结果如图 5-10 所示。

例 5.11　根据例 5.4 中数据文件“exfilms.csv”,绘制四部电影《阿飞正传》《建党伟业》《建军大业》《西虹市首富》上座率的箱线图。

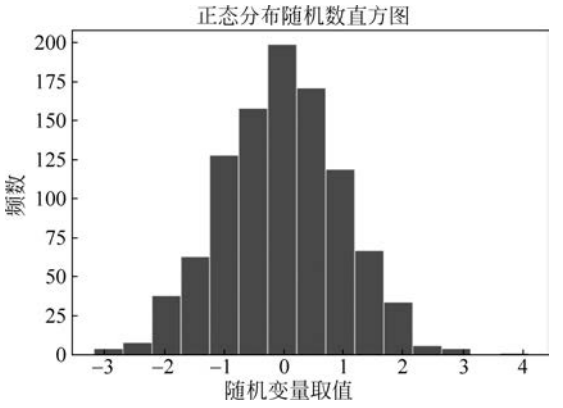

图 5-9　直方图

```
17 | 0 2
22 | 0 2 5
23 | 1 2 2 5
24 | 0 1 5
33 | 4 5 8
34 | 1
```

图 5-10　茎叶图

解

```
#Python 代码
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.font_manager import FontProperties
myfont = FontProperties(fname = r'C:\Windows\Fonts\simheii.ttf',size = 12)
sns.set(style = "ticks",font = myfont.get_name())
fig,ax = plt.subplots(figsize = (8,5))
data = pd.read_csv("exfilms.csv",encoding = "gbk")
data["时间"] = pd.to_datetime(data["时间"])
data["上座率"] = pd.to_numeric(data["上座率"], errors = 'coerce').fillna(0)
sns.boxplot(data["电影名称"],data["上座率"])
plt.show()
```

运行结果如图 5-11 所示。

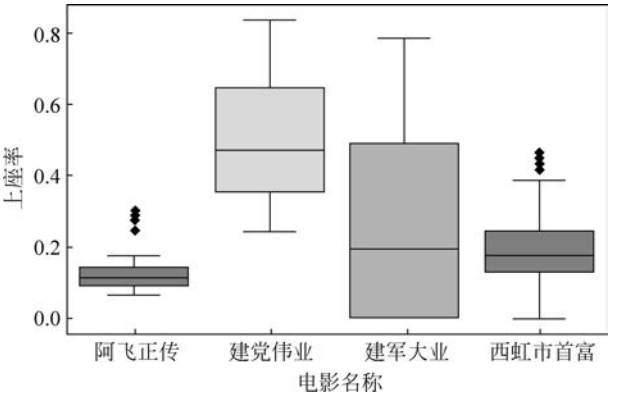

图 5-11　四部电影箱线图

5.3　推断统计

在研究对象的总体数量很大，甚至无限时，由于经济成本、时间及精力等各种原因只能通过部分个体，也即样本来对总体特征进行估计、检验及分析，这就是推断统计。

5.3.1　参数估计与假设检验

参数估计与假设检验都是基本的统计推断问题，首先介绍参数估计问题。

1. 参数估计

参数估计有两类，即点估计和与区间估计。点估计问题就是用某一个函数值作为总体未知参数的估计值；区间估计就是对于未知参数给出一个范围，并且在一定的可靠度下使这个范围包含未知参数。

点估计方法常用的有矩估计和极大似然估计。矩估计原理就是用样本 k 阶矩替换相应的总体矩，方法如下。

步骤 1：根据总体分布汇总未知参数的个数，写出总体矩方程，即 $\mu_k = E(X^k)$，$k = 1, 2, \cdots, l$。这里的 l 是参数个数。显然表达式的期望结果中含有未知参数，我们将其认为是参数的函数，写为

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

步骤 2：对上述方程组求出参数的表达式，即

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

步骤 3：用 $A_k = \dfrac{1}{n}\sum_{i=1}^{n} {X_i}^k$ 替换 μ_k，$k = 1, 2, \cdots, l$，即得估计值。

例 5.12　设总体 X 的均值 μ 及方差 σ^2 都存在，且有 $\sigma^2 > 0$，但 μ、σ^2 均为未知，X_1，X_2，$\cdots$，X_n 为样本，试求 μ、σ^2 的矩估计。

解　由矩估计法

$$\mu_1 = E(X) = \mu$$

$$\mu_2 = E(X^2) = D(X) + (E(X))^2 = \sigma^2 + \mu^2$$

即

$$\begin{cases}\mu=\mu_1\\ \sigma^2+\mu^2=\mu_2-\mu_1^2\end{cases}$$

用 $A_k=\frac{1}{n}\sum_{i=1}^{n}X_i{}^k$ 替换 μ_k，$k=1,2$，得

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

综上，不管总体 X 服从何种分布，总体期望和方差的矩估计量分别为样本均值、样本二阶中心矩。

例 5.13 某涂料的干燥时间服从正态分布 $X\sim N(\mu,\sigma^2)$，现对该涂料干燥时间进行 5 次测量，观察值(单位：秒)为：92,105,103,94,106，求未知参数 μ 与 σ^2 的矩估计。

解 将样本观察值代入例 5.12 中的矩估计量中得

$$\hat{\mu}=\frac{1}{5}(92+105+103+94+106)=100$$

$$\hat{\sigma}^2=\frac{1}{5}[(92-100)^2+(105-100)^2+(103-100)^2+(94-100)^2+(106-100)^2]=34$$

除矩估计外，点估计还有极大似然估计法。极大似然估计法是根据极大似然原理得出的估计方法，即在随机试验中，概率大的事件发生的可能性大于概率小的事件。如果在一次试验中，一个事件发生了，于是有理由认为此事件比其他事件发生的概率大。根据这一原理，对分布中未知参数 θ 选择一个合适的值，使取得的样本观测值出现的概率最大，用这个值作为未知参数的估计，此种方法称为极大似然估计法。下面分别就随机变量的类型分别给出估计方法。

(1) 总体为离散型分布。设总体 X 分布律为 $P\{X=x\}=p(x;\theta)$，其中 θ 为未知参数。$X_1,X_2,\cdots,X_n$ 是来自总体的样本，其观察值记为 $x_1,x_2,\cdots,x_n$，可知样本联合分布为 $P\{X_1=x_1,\cdots,X_n=x_n\}=\prod_{i=1}^{n}p\{X_i=x_i\}$，则记

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

上述函数称为参数的似然函数。若有 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 使 $L(\hat{\theta})=\max\limits_{\theta}L(\theta)$，则称 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 为 θ 的极大似然估计值。

(2) 总体为连续型分布。总体 X 概率密度为 $f(x,\theta)$，其中 θ 为未知参数，$X_1,X_2,\cdots,X_n$ 是取自总体 X 的样本，则样本的联合概率密度函数为

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

记 $L(\theta)=\prod_{i=1}^{n}f(x_i,\theta)$ 为似然函数。若有 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 使得 $L(\hat{\theta})=\max_{\theta}L(\theta)$，则称 $\hat{\theta}(x_1,x_2,\cdots,x_n)$ 为参数 θ 的极大似然估计值，称 $\hat{\theta}(X_1,X_2,\cdots,X_n)$ 为 θ 的极大似然估计量。

极大似然估计的一般步骤如下。

步骤 1：写出似然函数 $L(\theta)=L(x_1,x_2,\cdots,x_n,\theta)$。

步骤 2：令 $\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}=0$ 或 $\frac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=0$，求出驻点（未知参数多个时求对应偏导）。

步骤 3：判断并求出最大值点。

例 5.14　设总体 X 的概率函数为 $p(x;\theta)=\sqrt{\theta}x^{\sqrt{\theta}-1}$，$0<x<1$，$\theta>0$。$x_1,x_2,\cdots,x_n$ 是取自总体 X 的容量为 n 的样本，求未知参数 θ 的极大似然估计。

解　写出似然函数

$$L(\theta)=\sqrt{\theta}^{n}(x_1,x_2,\cdots,x_n)^{\sqrt{\theta}-1}$$

对数似然函数为

$$\ln L(\theta)=\frac{n}{2}\ln\theta+(\sqrt{\theta}-1)\sum_{i=1}^{n}\ln x_i$$

将 $\ln L(\theta)$ 关于 θ 求导并令其为 0，即得到似然方程

$$\frac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=\frac{n}{2\theta}+\frac{1}{2\sqrt{\theta}}\sum_{i=1}^{n}\ln x_i=0$$

解得 θ 的极大似然估计为

$$\hat{\theta}=\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_i\right)^{-2}$$

除了点估计，实际问题中还需要估计参数所在的区间，即区间估计。

首先介绍置信区间的概念。设总体 X 分布函数 $F(x;\theta)$ 中含有未知参数 θ，对给定的 $\alpha(0<\alpha<1)$，若由样本 $X_1,X_2,\cdots,X_n$ 确定的两个统计量

$$\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)\quad\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$$

使得 $P\{\hat{\theta}_1<\theta<\hat{\theta}_2\}=1-\alpha$，则称随机区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 为 θ 的 $1-\alpha$ 置信区间，称 $1-\alpha$ 为置信水平，称 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 分别为 θ 的置信下限与置信上限。

对于置信区间需要注意以下两点。

（1）若对样本 $X_1,X_2,\cdots,X_n$ 进行多次观察，对应每个样本值都可确定一置信区间，每个这样的区间要么包含了 θ 的真值，要么不包含 θ 的真值。由大数定律，当抽样次数充分大时，包含 θ 真值的区间大约有 $100(1-\alpha)\%$ 个。

（2）置信水平 $1-\alpha$ 越大，置信区间包含参数真值的概率就越大，但区间 $(\hat{\theta}_1,\hat{\theta}_2)$ 的长度变大意味着参数估计精度变差，因此原则是在设定置信水平前提件下尽可能使得参数位于较小的估计区间中。

求置信区间的一般步骤如下。

(1) 找到一个包含样本 $X_1,X_2,\cdots,X_n$ 和参数的函数 $Z(X_1,X_2,\cdots,X_n,\theta)$,此函数不包含其他未知参数,且 Z 的分布已知,并不依赖于任何未知参数。

(2) 对给出的置信水平 $1-\alpha(0<\alpha<1)$,利用 Z 的分布定出两个临界值 z_1、z_2 使

$$P(z_1 \leqslant Z(X_1,X_2,\cdots,X_n,\theta) \leqslant z_2)=1-\alpha$$

(3) 从不等式 $z_1 \leqslant Z(X_1,X_2,\cdots,X_n,\theta) \leqslant z_2$ 中解出等价的不等式 $\hat{\theta}_1 \leqslant \theta \leqslant \hat{\theta}_2$,其中 $\hat{\theta}_1=\hat{\theta}_1(X_1,X_2,\cdots,X_n)$,$\hat{\theta}_2=\hat{\theta}_2(X_1,X_2,\cdots,X_n)$,则得所求的置信区间为$(\hat{\theta}_1,\hat{\theta}_2)$。

单个正态总体 $X \sim N(\mu,\sigma^2)$的参数区间估计如下。

(1) σ^2已知时均值 μ 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\overline{X}-\frac{\sigma}{\sqrt{n}}u_{\alpha/2},\overline{X}+\frac{\sigma}{\sqrt{n}}u_{\alpha/2}\right]$$

(2) σ^2未知时均值 μ 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\overline{X}-t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\overline{X}+t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right]$$

(3) μ 已知时均值σ^2 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2}}(n)},\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2}}(n)}\right]$$

(4) μ 未知时均值σ^2 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

两个正态总体下的参数区间估计如下。

(1) σ_1^2、σ_2^2 已知时 $\mu_1-\mu_2$ 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\overline{X}-\overline{Y}-u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\overline{X}-\overline{Y}+u_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

(2) $\sigma_1^2=\sigma_2^2=\sigma^2$ 未知时 $\mu_1-\mu_2$ 的置信水平为 $1-\alpha$ 的置信区间

$$\left((\overline{X}-\overline{Y})-t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},(\overline{X}-\overline{Y})+t_{\alpha/2}(n_1+n_2-2)S_w\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

其中,$S_w=\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$。

(3) μ_1、μ_2 未知时 σ_1^2/σ_2^2 的置信水平为 $1-\alpha$ 的置信区间

$$\left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)},\quad\frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2}(n_1-1,n_2-1)}\right]$$

2. 假设检验

假设检验问题分为参数假设检验和非参数假设检验。一般情况下,假设检验主要指参

数假设检验问题。假设检验主要考虑的是对总体分布函数中未知参数提出的假设进行检验。

1）单个正态总体 $N(\mu,\sigma^2)$下参数的假设检验

（1）方差 σ^2 已知，μ 的检验——u 检验。

检验统计量 $U=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$，给定显著性水平 α 下：

对检验问题 $H_0:\mu=\mu_0$；$H_1:\mu\neq\mu_0$，拒绝域 $W=\{|U|>u_{\alpha/2}\}=\{U<-u_{\alpha/2}\}\cup\{U>u_{\alpha/2}\}$。

对检验问题 $H_0:\mu\leqslant\mu_0$；$H_1:\mu>\mu_0$，拒绝域 $W=\{U>u_\alpha\}$。

对检验问题 $H_0:\mu\geqslant\mu_0$；$H_1:\mu<\mu_0$，拒绝域 $W=\{U<u_{1-\alpha}\}$。

（2）方差 σ^2 未知，μ 的检验—— t 检验。

检验统计量 $T=\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$，给定显著性水平 α 下：

对检验问题 $H_0:\mu=\mu_0$；$H_1:\mu\neq\mu_0$，拒绝域 $W=\{|T|>t_{\alpha/2}(n-1)\}$。

对检验问题 $H_0:\mu\leqslant\mu_0$；$H_1:\mu>\mu_0$，拒绝域 $W=\{T>t_\alpha(n-1)\}$。

对检验问题 $H_0:\mu\geqslant\mu_0$；$H_1:\mu<\mu_0$，拒绝域 $W=\{T<-t_\alpha(n-1)\}$。

（3）μ 未知，σ^2 的检验——χ^2 检验。

检验统计量 $\chi^2=\dfrac{(n-1)S^2}{\sigma_0^2}$，给定显著性水平 α 下：

对检验问题 $H_0:\sigma^2=\sigma_0^2$；$H_1:\sigma^2\neq\sigma_0^2$，拒绝域 $W=\{\chi^2<\chi^2_{1-\alpha/2}(n-1)\}\cup\{\chi^2>\chi^2_{\alpha/2}(n-1)\}$。

对检验问题 $H_0:\sigma^2\leqslant\sigma_0^2$；$H_1:\sigma^2>\sigma_0^2$，拒绝域 $W=\{\chi^2>\chi^2_\alpha(n-1)\}$。

对检验问题 $H_0:\mu\geqslant\mu_0$；$H_1:\mu<\mu_0$，拒绝域 $W=\{\chi^2<\chi^2_{1-\alpha}(n-1)\}$。

2）两个正态总体下参数的区间估计

考虑如下检验假设

$$H_0:\sigma_1^2=\sigma_2^2;\ H_1:\sigma_1^2\neq\sigma_2^2$$

取检验统计量 $F=\dfrac{S_1^2}{S_2^2}\sim F(n_1-1,n_2-1)$。给定显著性水平 α，由 F 分布分位点的定义可知 $P\{[F<F_{1-\alpha/2}(n_1-1,n_2-1)]\cup[F>F_{\alpha/2}(n_1-1,n_2-1)]\}=\alpha$，于是推得拒绝域 $W=\{F<F_{1-\alpha/2}(n_1-1,n_2-1)\}\cup\{F>F_{\alpha/2}(n_1-1,n_2-1)\}$。

上面的检验均为正态分布下的参数检验。实际中，还需要对获取到的数据进行分布类型验证，一般情况下，往往进行正态性检验。正态性检验的方法比较多，常见的主要有以下几种。

（1）Q-Q 图。

Q-Q 图（quantile-quantile plot）用来判断数据的正态分布性，横坐标表示实际数据的分

位数，纵坐标表示正态分布的对应分位数。在样本服从正态分布的条件下，图中的散点应该是呈现出45°的直线。除了进行正态检验，Q-Q图还能进行如Beta分布、卡方分布、指数分布等的检验，其步骤如下。

步骤1：将数据样本从小到大排序，记为 $x_1, x_2, \cdots, x_n$。

步骤2：分位数 $Q_i = \dfrac{x_i - \bar{x}}{\sigma}$。

步骤3：对百分比 $P_i = \dfrac{i - 0.5}{n}$，找出对应标准正态分布的分位数 Q_i'。

步骤4：以 Q_i 为横轴，以 Q_i' 为纵轴画图检验。

(2) 皮尔逊 χ^2 检验。

皮尔逊 χ^2 检验适用于样本比较大的场合，如 $n \geqslant 50$。设 $F(x)$ 为总体 X 的分布函数，$F_0(x)$ 为一个已知的分布函数，$X_1, X_2, \cdots, X_n$ 为总体 X 的一个样本，检验的原假设是 $H_0: F(x) = F_0(x)$，取检验统计量

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

式中，$\hat{p}_i$ 是由 $\hat{F}_0(x)$ 计算出来的理论频率，$\hat{F}_0(x)$ 是由 $F_0(x)$ 中未知参数估计出的分布函数，f_i/n 为实际频率，且有如下定理。

定理 5-1 若 n 足够大，当 H_0 成立时，统计量 χ^2 总是近似地服从自由度为 k-r-1 的 χ^2 分布，其中 r 是已知的分布函数 $F_0(x)$ 中未知参数的个数。

皮尔逊 χ^2 检验的步骤如下。

步骤1：把总体 X 划分成 k 个不交的区间 $A_i (i = 1, 2, \cdots, k)$，使得每个区间包含的理论频数满足 $np_i \geqslant 5$，否则将区间适当调整。

步骤2：H_0 成立时，计算各理论频率，即概率 p_i 的值

$$p_i = P(A_i) = F_0(y_i) - F_0(y_{i-1}), \quad i = 1, 2, \cdots, k$$

式中，y_{i-1} 与 y_i 为区间 A_i 的端点，即 $A_i = (y_{i-1}, y_i]$。

步骤3：找出 A_i 的频数 f_i，并计算

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

步骤4：对于给定的显著性水平 α，找出临界值 $\chi_\alpha^2(k-1)$。

步骤5：若 $\chi^2 > \chi_\alpha^2(k-1)$，则拒绝 H_0，否则可接受 H_0。

由于皮尔逊 χ^2 检验对数据进行分组，会丢失部分信息，且除了正态分布还可以检验其他的分布，对正态分布没有特效性，在效果上并不是很理想。

(3) Shapiro-Wilk W 检验。

检验步骤如下。

步骤1：数据样本从小到大排序，记为 $x_1, x_2, \cdots, x_n$。

步骤 2：对分布为正态的原假设问题，选择统计量

$$W=\frac{\left(\sum_{i=1}^{\left[\frac{n}{2}\right]}a_i(x_{n+1-i}-x_i)\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

式中，系数 a_i 可查 Shapiro-wilk W 检验的系数表。

步骤 3：对样本容量 n 及给定的检验水平 α，查找 Shapiro-wilk W 检验的临界值 W_α。

步骤 4：根据样本值计算出 W 值，若 $W<W_\alpha$，则拒绝原假设。

Shapiro-wilk W 检验易受异常值的影响，且一般适用样本容量为 3～50 的场合。

除了上述几种检验方法，用于正态检验的还有 Kolmogorov-Smirnov 检验、Lilliefor 检验、AD 检验、D 检验、偏度检验和峰度检验等，具体步骤不再叙述。

5.3.2　方差分析

方差分析是通过试验的结果数据来推断各个有关因素对试验结果是否产生显著影响的一种统计方法。方差分析中涉及几个概念，即因素或因子，指影响试验指标的条件，也就是研究变量，常表示为因素 A、因素 B 等；水平指因素所处的状态或内容，如因素 A 的水平记为 A_1、A_2 等；单因素方差分析指实验中涉及一个因素；多因素方差分析指试验中的因素多于一个。

1. 单因素方差分析

单因素方差分析可用的数据如表 5-8 所示。

表 5-8　单因素方差分析可用的数据

试验次数	A_1	…	A_i	…	A_r
1	x_{11}	…	x_{i1}	…	x_{r1}
2	x_{12}	…	x_{i2}	…	x_{r2}
…	…	…	…	…	…
j	x_{1j}	…	x_{ij}	…	x_{rj}
…	…	…	…	…	…
n_i	x_{1n1}	…	x_{ini}	…	x_{rnr}

例如，为了考察不同种类的肥料对某地区玉米亩产量是否有影响，做了如下实验，数据如表 5-9 所示。

表 5-9　玉米亩产量数据

实验次数	A_1	A_2	A_3
1	251	268	257
2	263	261	265
3	259	272	255
4	248	269	273

其中，因素即为肥料，A_1、A_2、A_3 表示三种肥料，即三个水平，表 5-9 中数据如 263 可记为 x_{12}。

方差分析中要求：各水平下的总体都服从正态分布；各水平下的总体方差相等，即方差齐性；数据的取得是相互独立的。于是

$$X_{ij} \sim N(\mu_j, \sigma^2)$$

记 $\varepsilon_{ij} = X_{ij} - \mu_j$，则有 $X_{ij} = \mu_j + \varepsilon_{ij}$，$\varepsilon_{ij} \sim N(0, \sigma^2)$，且各 ε_{ij} 相互独立，$j = 1, 2, \cdots, n_i$，$i = 1, 2, \cdots, r$。其中，$\mu_1, \mu_2, \cdots, \mu_r$ 及 σ^2 为未知参数。

方差分析的任务是检验 r 个总体的均值是否相等，也即检验假设 H_0：$\mu_1 = \mu_2 = \cdots = \mu_s$；$H_1$：$\mu_1, \mu_2, \cdots, \mu_s$ 不全相等。

下面叙述简要步骤。

步骤 1：计算平均值。组内平均值 $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$，总平均 $\bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} x_{ij}$。

步骤 2：计算离差平方和，即

总离差平方和 SS_T(sum of squares for total)

$$SS_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

组间离差平方和 SS_A(sum of square for factor A)

$$SS_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i(\bar{x}_i - \bar{x})^2$$

组内离差平方和 SS_e(sum of square for error)

$$SS_e = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

总离差平方和 SS_T 表示了各试验值与总平均值的偏差的平方和，反映了试验结果之间存在的总差异；组间离差平方和 SS_A 反映了各组内平均值之间的差异程度，这是由因素 A 不同水平的不同作用造成的。组内离差平方和 SS_e 反映了在各水平内试验值之间的差异程度，由随机误差的作用产生。

三种离差平方和之间关系为

$$SS_T = SS_A + SS_e$$

步骤 3：计算如下自由度(degree of freedom)。

总自由度：$df_T = n - 1$。

组间自由度：$df_A = r - 1$。

组内自由度：$df_e = n - r$。

三者关系：$df_T = df_A + df_e$。

步骤 4：计算平均平方，均方等于离差平方和除以对应的自由度，即

$$MS_A = SS_A/df_A, \quad MS_e = SS_e/df_e$$

其中，MS_A 表示组间均方，MS_e 表示组内均方/误差的均方。

步骤 5：F 检验

$$F_A = \frac{\text{组间均方}}{\text{组内均方}} = \frac{MS_A}{MS_e}$$

上述统计量服从自由度为(df_A，df_e)的 F 分布，对于给定的显著性水平 α，从 F 分布表查得临界值 $F_\alpha(df_A, df_e)$，如果 $F_A > F_\alpha(df_A, df_e)$，则认为因素 A 对实验结果有显著影响，否则认为因素 A 对试验结果没有显著影响，如表 5-10 所示。

表 5-10　单因素方差分析表

差异源	SS	df	MS	F
组间(因素 A)	SS_A	$r-1$	$MS_A = SS_A/(r-1)$	MS_A/MS_e
组内(误差)	SS_e	$n-r$	$MS_e = SS_e/(n-r)$	—
总和	SS_T	$n-1$	—	—

按照上述步骤，对表 5-9 中的玉米亩产量数据进行单因素方差分析，计算得其描述数据与方差分析表，如表 5-11 和表 5-12 所示。

表 5-11　玉米亩产量的描述数据

组	观　测　数	求　　和	平　　均	方　　差
A_1	4	1021	255.25	48.25
A_2	4	1070	267.5	21.66667
A_3	4	1050	262.5	67.66667

表 5-12　玉米亩产量的方差分析(显著水平 0.05)

差异源	SS	df	MS	F	P 值	F 临界值
组间	303.5	2	151.75	3.308904	0.083715	4.256495
组内	412.75	9	45.86111	—	—	—
总计	716.25	11	—	—	—	—

F 值 3.308904 小于临界值 4.256495，即认为三种肥料在 0.05 的水平下对亩产量没有显著影响。通过表 5-12 中的 P 值也可以得到同样结论。

2. 双因素方差分析

双因素无重复方差分析如表 5-13 所示。

表 5-13　双因素无重复实验数据

	B_1	B_2	…	B_s
A_1	x_{11}	x_{12}	…	x_{1s}
A_2	x_{21}	x_{22}	…	x_{2s}
…	…	…	…	…
A_r	x_{r1}	x_{r2}	…	x_{rs}

不同于单因素方差分析，这里考虑两个因素对实验结果的影响。例如，在某化工产品的得率实验中，对添加剂浓度取四种不同水平值，温度取三种不同水平值做如下实验以考察浓度和温度这两个因素对得率是否有显著影响，表 5-14 给出了不同温度和浓度下的产品率。

表 5-14 不同温度和浓度下的产品率

浓度	温度		
	B_1	B_2	B_3
A_1	2.7	2.1	2
A_2	1.9	1.9	2.1
A_3	0.6	1.2	1.6
A_4	0.4	0.2	0.5

双因素无重复情况下的检验步骤如下。

步骤 1：计算各个平均值，即

水平 A_i 平均值 $\bar{x}_{i\cdot}=\frac{1}{s}\sum_{j=1}^{s}x_{ij}$，水平 B_j 平均值 $\bar{x}_{\cdot j}=\frac{1}{r}\sum_{i=1}^{r}x_{ij}$，总平均 $\bar{x}=\frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}x_{ij}$。

步骤 2：计算离差平方和，即

总离差平方和

$$SS_T=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x})^2$$

因素 A 离差平方和

$$SS_A=\sum_{j=1}^{s}\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2=s\sum_{i=1}^{r}(\bar{x}_{i\cdot}-\bar{x})^2$$

因素 B 离差平方和

$$SS_B=\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x}_{\cdot j}-\bar{x})^2=r\sum_{j=1}^{s}(\bar{x}_{\cdot j}-\bar{x})^2$$

误差平方和

$$SS_e=\sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2$$

离差平方和之间关系

$$SS_T=SS_A+SS_B+SS_e$$

步骤 3：计算如下自由度。

总自由度：$df_T=n-1=rs-1$。

SS_A 自由度：$df_A=r-1$。

SS_B 自由度：$df_B=s-1$。

SS_e 自由度：$df_e=(r-1)(s-1)$。

自由度之间关系：$df_T=df_A+df_B+df_e$。

步骤 4：计算平均平方，均方等于离差平方和除以对应的自由度，即

$$MS_A=SS_A/df_A,\quad MS_B=SS_B/df_B,\quad MS_e=SS_e/df_e$$

其中，MS_A 表示组间均方，MS_e 表示组内均方/误差的均方。

步骤 5：F 检验

$$F_A=\frac{MS_A}{MS_e},\quad F_B=\frac{MS_B}{MS_e}$$

上述统计量 F_A 服从自由度为(df_A,df_e)的 F 分布；F_B 服从自由度为(df_B,df_e)的 F 分布。对于给定的显著性水平 α，从 F 分布表查得临界值$F_A(df_A,df_e)$及 $F_B(df_B,df_e)$。则若 $F_A>F_A(df_A,df_e)$，认为因素 A 使实验结果产生了显著影响，同样，若 $F_B>F_B(df_B,df_e)$，则认为因素 B 对实验有显著影响，表 5-15 给出了双因素方差分析结果。

表 5-15　双因素方差分析结果

差异源	SS	df	MS	F
因素 A	SS_A	$r-1$	$MS_A=SS_A/(r-1)$	$F_A=MS_A/MS_e$
因素 B	SS_B	$s-1$	$MS_B=SS_B/(s-1)$	$F_B=MS_B/MS_e$
误差	SS_e	$(r-1)(s-1)$	$MS_e=SS_e/(r-1)(s-1)$	—
总和	SS_T	$rs-1$	—	—

对表 5-14 中不同温度和浓度下的产品率的实验数据，通过上述步骤计算得到的数据统计如表 5-16 所示。

表 5-16　两个因素实验数据的描述

	观　测　数	求　　和	平　　均	方　　差
浓度 1	3	6.8	2.266667	0.143333
浓度 2	3	5.9	1.966667	0.013333
浓度 3	3	3.4	1.133333	0.253333
浓度 4	3	1.1	0.366667	0.023333
温度 1	4	5.6	1.4	1.193333
温度 2	4	5.4	1.35	0.736667
温度 3	4	6.2	1.55	0.536667

方差分析如表 5-17 所示。

表 5-17　方差分析(显著水平为 0.05)

差异源	SS	df	MS	F	P 值	F 临界值
浓度	6.62	3	2.206667	16.97436	0.002458	4.757063
温度	0.086667	2	0.043333	0.333333	0.729	5.143253
误差	0.78	6	0.13	—	—	—
总计	7.486667	11	—	—	—	—

如表 5-17 所示，比较 F 值和临界值可以看出，在 0.05 的显著水平下，浓度对产品得率有显著性影响，而温度对实验结果没有显著影响。

5.3.3　回归分析

回归分析是研究变量之间相关关系的一种统计方法。相关关系也称为不确定性关系，

变量之间的相关关系通过大量实验和观测可表现出某种规律性，回归分析可得到变量之间的关系方程。

1. 一元线性回归

回归方程中的变量 y 作为因变量，x 为自变量，它们之间的关系为

$$y=a+bx+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2)$$

其中，参数 a、b、σ^2 为不依赖于自变量 x 的待估参数。这就是一元线性回归模型。由上式可知，$y\sim N(a+bx,\sigma^2)$，它依赖于 x 的值。若通过实验数据估计出了参数 a、b 的值 $\hat{a}$、$\hat{b}$，则称关系式 $\hat{y}=\hat{a}+\hat{b}x$ 为一元线性回归方程或回归方程，它的图形称为回归直线，一般把 b 称为回归系数。

通过实验或观察，得到变量 x 和变量 y 的一组数据，即 $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，于是回归模型可以改写为如下形式

$$y_i=a+bx_i+\varepsilon_i$$

其中，$\varepsilon_i\sim N(0,\sigma^2)$，各 ε_i 相互独立，$i=1,2,\cdots,n$。

回归分析主要解决的问题：对 a、b 和 σ 作点估计，对回归方程作假设检验，在点 x_0 处作出 y 的预测。

由 $\varepsilon_i=y_i-a-bx_i\sim N(0,\sigma^2)$ 知，误差期望为零。若记

$$Q=\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

称 Q 为误差平方和，它反映了 y 与 $a+bx$ 之间在 n 次观察中总的误差程度。又由 $Y_i\sim N(a+bx_i,\sigma^2)$，$i=1,2,\cdots,n$，且 $Y_1,Y_2,\cdots,Y_n$ 相互独立，故可写出 $Y_1,Y_2,\cdots,Y_n$ 的联合概率密度或似然函数。通过极大似然估计的方法即可求出参数，这与最小二乘原理得到的结果一致。下面利用最小二乘法来推导，也就是找到使得 Q 达到最小值的 a 和 b，即 a 和 b 满足

$$\min_{a,b}Q=\min_{a,b}\sum_{i=1}^{n}(y_i-a-bx_i)^2$$

对二元函数 Q 求驻点有

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)=0\\[2ex]\dfrac{\partial Q}{\partial b}=-2\displaystyle\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{cases}$$

可得

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i{}^2-n\bar{x}^2$$

$$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}$$

显然有

$$\begin{cases}na+n\bar{x}b=n\bar{y}\\ n\bar{x}a+b\sum_{i=1}^{n}x_i{}^2=\sum_{i=1}^{n}x_iy_i\end{cases}$$

称上面的表达式为正规方程组，解之得 a 和 b 的最小二乘估计为

$$\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i{}^2-n\bar{x}^2}=\frac{L_{xy}}{L_{xx}}$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}$$

利用样本的观察数据和上面估计出的参数 $\hat{a}$ 和 $\hat{b}$，即可写出变量 x 和变量 y 之间的回归方程为 $\hat{y}_i=\hat{a}+\hat{b}x_i(i=1,2,\cdots,n)$。

定义 $y_i-\hat{y}_i$ 为 x_i 处的残差，其平方和

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$$

称为残差平方和。由于

$$\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)$$

可知 $E\left(\frac{Q_e}{\sigma^2}\right)=n-2$，于是得 σ^2 的无偏估计 $\hat{\sigma}^2=\frac{Q_e}{n-2}$。由

$$\begin{aligned}Q_e&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\\&=\sum_{i=1}^{n}(y_i-\bar{y}-\hat{b}(x_i-\bar{x}))^2\\&=L_{yy}-2\hat{b}L_{xy}+\hat{b}^2L_{xx}\\&=L_{yy}-\hat{b}L_{xy}\end{aligned}$$

得 σ^2 的无偏估计 $\hat{\sigma}^2$ 为

$$\hat{\sigma}^2=\frac{L_{yy}-\hat{b}L_{xy}}{n-2}$$

目前已经对回归方程的系数给出了估计。但是建立的方程是否有效或显著，还需要进

行假设检验才能确定，即需要对回归系数 b 是否为零进行检验。若 $b=0$，则变量 y 与变量 x 没有关系；若 $b\neq 0$，则说明 y 与 x 之间存在线性相关关系，于是假设检验问题是

$$H_0: b=0; \quad H_1: b\neq 0$$

对上述问题进行假设检验常用的方法有 r 检验法和 t 检验法等。

(1) r 检验法。

计算检验统计量

$$r=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

给定显著性水平 α，当 $|r|>r_{1-\alpha}$ 时拒绝原假设，认为建立的回归方程显著，这里临界值 $r_{1-\alpha}=\sqrt{\frac{1}{1+(n-2)/F_{1-\alpha}(1,n-2)}}$。

(2) t 检验法。

检验统计量 $t=\frac{\hat{b}}{\hat{\sigma}}\sqrt{L_{xx}}\sim t(n-2)$，给定显著性水平 α，当 $|t|\geqslant t_{\alpha/2}(n-2)$ 时拒绝原假设，认为建立的回归方程显著。

建立回归方程后，一个重要的作用就是对给定的自变量 x 取值来预测 y 的值，同时给出置信区间。假定给出的值为 x_0，由于

$$E(\hat{y}_0)=E(\hat{a}+\hat{b}x_0)=a+bx_0=E(y_0)$$

故预测值 $\hat{y}_0$ 是 $E(y_0)$ 的无偏估计。由

$$\hat{y}_0=\hat{a}+\hat{b}x_0\sim N\left(a+bx_0,\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

知

$$y_0-\hat{y}_0\sim N\left(0,\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}\right)\sigma^2\right)$$

又因为 $\frac{Q_e}{\sigma^2}\sim\chi^2(n-2)$，且 Q_e 和 $\hat{y}_0$ 相互独立，也即 Q_e 和 $y_0-\hat{y}_0$ 相互独立，故有

$$t=\frac{y_0-\hat{y}_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}}}\Bigg/\sqrt{\frac{Q_e}{\sigma^2(n-2)}}=\frac{y_0-\hat{y}_0}{\sqrt{\frac{Q_e}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}}}\sim t(n-2)$$

若置信水平为 $1-\alpha$，则求得 y_0 的置信水平为 $1-\alpha$ 的置信区间为

$$[\hat{y}_0-\delta(x_0),\hat{y}_0+\delta(x_0)]$$

其中，$\delta(x_0)=t_{\alpha/2}(n-2)S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}}$，$S=\sqrt{\frac{Q}{n-2}}$ 称为剩余标准差。

在取得观察数据后，曲线 $y_1(x)=\hat{y}-\delta(x)$ 和 $y_2(x)=\hat{y}+\delta(x)$ 之间的区域即为 $y=a+bx+\varepsilon$ 的置信水平是 $1-\alpha$ 的预测带域。

2. 多元线性回归

前面介绍的是一元线性回归方程，实际问题中随机变量 Y 往往与多个自变量相关，如果这种关系是线性的，即

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k$$

称为多元线性回归，建立的方程称为回归平面方程，一般模型为

$$\begin{cases} Y=X\beta+\varepsilon \\ E(\varepsilon)=0,\operatorname{cov}(\varepsilon,\varepsilon)=\sigma^2 I_n \end{cases}$$

其中

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\quad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},\quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},\quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

对未知参数 β 和 σ^2 作点估计，作离差平方和

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_k x_{ik})^2$$

选择 $\beta_0,\beta_1,\cdots,\beta_k$ 使 Q 达到最小。采用最小二乘法进行求解，得到估计值

$$\hat{\beta}=(X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}Y)$$

将估计值 $\hat{\beta}_i$ 代入回归平面方程得经验回归平面方程为

$$y=\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k$$

$\hat{\beta}_i$ 称为经验回归系数。

多元回归方程的显著性检验中原假设为

$$H_0:\beta_0=\beta_1=\cdots=\beta_k=0$$

采用的假设检验为 F 检验法。统计量为

$$F=\frac{U/k}{Q_{\mathrm{e}}/(n-k-1)}\sim F(k,n-k-1)$$

式中，$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，$Q_{\mathrm{e}}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$。如果 $F>F_{1-\alpha}(k,n-k-1)$，则拒绝 H_0，认为在显著水平 α 下建立的多元线性回归方程是显著的。

建立回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k$ 后，对于自变量的取值可以给出点的预测，如对于 $x_1^*,x_2^*,\cdots,x_k^*$，可得 $y^*=\beta_0+\beta_1 x_1^*+\cdots+\beta_k x_k{}^*+\varepsilon$。若置信水平取 $1-\alpha$，则因变量 y 的置信水平为 $1-\alpha$ 的置信区间为 $(\hat{y}\pm\delta)$，其中

$$\delta=\hat{\sigma}_{\mathrm{e}}\sqrt{1+\sum_{i=0}^{k}\sum_{j=0}^{k}c_{ij}x_ix_j}\,t_{1-\alpha/2}(n-k-1)$$

$$\hat{\sigma}_{\mathrm{e}}=\sqrt{\frac{Q_{\mathrm{e}}}{n-k-1}},\quad C=L^{-1}=(c_{ij}),\quad L=X'X$$

5.3.4 聚类分析

聚类分析又称群分析、点群分析，是研究分类问题即相似元素集合的一种多元统计方法。实际应用中往往把聚类分析与判别分析、主成分分析等多元分析方法结合起来，以取得较好的效果。

为将研究对象分类，需要度量样品之间的关系，比较主流的有两种衡量方法，一种方法是求相似系数；另一种方法是将一个样品看作 P 维空间的一个点，并在空间定义距离，距离相近的点归为一类。聚类分析的一般思想是在样品之间分类采用距离，在指标之间分类采用相似系数。样品距离表明样品之间的相似度，指标之间的相似系数刻画指标之间的相似度。对研究对象按相似度进行归类，关系密切的聚集到较小的一类，关系疏远的聚集到较大的一类，直到所有的样品或指标聚集完毕。

聚类问题的一般提法是：设由 n 个样品的 p 元观测数据组成一个数据矩阵，$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$，$i=1,2,\cdots,n$。这里每个样品可看成 p 元空间的一个点，即一个 p 维向量。矩阵形式为

$$\boldsymbol{HX}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\ddots&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}$$

式中，$x_{ij}(i=1,2,\cdots,n;\ j=1,2,\cdots,p)$为第 i 个样品的第 j 个指标的观测数据。第 i 个样品 X_i 为矩阵 $\boldsymbol{HX}$ 的第 i 行所描述，所以任何两个样品之间的相似性，可以通过矩阵 $\boldsymbol{HX}$ 中对应的行之间相似程度来刻画；任何两个指标之间的相似性，可以通过对应的列之间相似程度来刻画。

样品分类和指标分类采用的方法类似，下面以样品分类为例进行介绍。

常见的距离有以下几种。

(1) 明氏(Minkowski)距离

$$d_{ij}(q)=\left(\sum_{a=1}^{p}|\,x_{ia}-x_{ja}\,|^{q}\right)^{1/q}$$

当 $q=1$ 时，即为绝对距离，当 $q=2$ 时即为欧氏距离

(2) 切比雪夫距离

$$d_{ij}(\infty)=\max_{1\leqslant a\leqslant p}|\,x_{ia}-x_{ja}\,|$$

(3) 马氏(Mahalanobis)距离

$$d(x_i,x_j)=(x_i-x_j)^{\mathrm{T}}\Delta^{-1}(x_i-x_j)$$

这里 Δ 为样品的协方差矩阵。

(4) 方差加权距离

$$d(x_i,x_j)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2/s_k^2\right]^{1/2}$$

其中，$s_k^2=\frac{1}{n-1}\sum_{j=1}^{n}(x_{jk}-\bar{X}_k)^2$，$\bar{X}_k=\frac{1}{n}\sum_{j=1}^{n}x_{jk}$。

除距离外，描述相似程度的量还有相似系数，一般有如下两种。

(1) 夹角余弦。

把研究对象 X_i 与 X_j 看作 p 维空间的两个向量，这两个向量的夹角余弦表示为 $\cos\theta_{ij}$，于是

$$\cos\theta_{ij}=\frac{\sum_{a=1}^{p}x_{ia}x_{ja}}{\sqrt{\sum_{a=1}^{p}x_{ia}^2\cdot\sum_{a=1}^{p}x_{ja}^2}}$$

若 $\cos\theta_{ij}=1$，则说明两个样品 X_i 与 X_j 完全相似；$\cos\theta_{ij}$ 接近 1，说明 X_i 与 X_j 相似密切；若 $\cos\theta_{ij}$ 接近于 0，说明 X_i 与 X_j 差别大，如果夹角余弦等于零，则说明 X_i 与 X_j 完全不一样。

(2) 相关系数。

第 i 个样品与第 j 个样品之间的相关系数定义为

$$r_{ij}=\frac{\sum_{a=1}^{p}(x_{ia}-\bar{x}_i)(x_{ja}-\bar{x}_j)}{\sqrt{\sum_{a=1}^{p}(x_{ia}-\bar{x}_i)^2\cdot\sum_{a=1}^{p}(x_{ja}-\bar{x}_j)^2}}$$

式中，$\bar{x}_i=\frac{1}{p}\sum_{a=1}^{p}x_{ia}$，$\bar{x}_j=\frac{1}{p}\sum_{a=1}^{p}x_{ja}$，$-1\leqslant r_{ij}\leqslant 1$。事实上，相关系数也是另一种形式的夹角余弦。

上面给出了样品之间的距离定义。在类与类之间也有各种定义的距离。常用的系统聚类方法中有最短距离法、最长距离法、类平均法、重心法、可变法、离差平方和法等。这些不同的距离在归类时有类似的步骤，只不过是采用的类之间的距离不同。下面介绍常见的几个类之间的距离。

(1) 最短距离

$$D_{pq}=\min_{i\in G_p,j\in G_q}d_{ij}$$

即用两类中样品之间的距离最短者作为两类间距离。

(2) 最长距离

$$D_{pq}=\max_{i\in G_p,j\in G_q}d_{ij}$$

即用两类中样品之间的距离最长者作为两类间距离。

（3）类平均距离

$$D_{pq}=\frac{1}{n_p n_q}\sum_{i\in G_p}\sum_{j\in G_q}d_{ij}$$

即用两类中所有两两样品之间距离的平均作为两类间距离。

（4）重心距离

$$D_{pq}=d(\bar{x}_p,\bar{x}_q)=\sqrt{(\bar{x}_p-\bar{x}_q)^{\mathrm{T}}(\bar{x}_p-\bar{x}_q)}$$

式中，$\bar{x}_p$、$\bar{x}_q$ 分别是 G_p、G_q 的重心，这是用两类重心之间的欧氏距离作为类间距离。

例 5.15 对 5 个样品测得某指标值为 1.5，2.5，4，7.5，9.5，用最短距离方法对其进行聚类分析。

解 首先各个指标值自成一类，记为 G_i，$i=1,2,3,4,5$。计算方便起见，取样品之间距离为绝对距离，距离矩阵如表 5-18 所示。

表 5-18 距离矩阵

	$G_1=\{1.5\}$	$G_2=\{2.5\}$	$G_3=\{4\}$	$G_4=\{7.5\}$	$G_5=\{9.5\}$
$G_1=\{1.5\}$	0				
$G_2=\{2.5\}$	1	0			
$G_3=\{4\}$	2.5	1.5	0		
$G_4=\{7.5\}$	6	5	3.5	0	
$G_5=\{9.5\}$	8	7	5.5	2	0

表 5-18 中找到对角线之外的最小元素为 1，于是把 G_1、G_2 归为一类，记为 $G_6=\{1.5,2.5\}$。新的距离矩阵如表 5-19 所示。

表 5-19 新的距离矩阵（一）

	$G_6=\{1.5,2.5\}$	$G_3=\{4\}$	$G_4=\{7.5\}$	$G_5=\{9.5\}$
$G_6=\{1.5,2.5\}$	0			
$G_3=\{4\}$	1.5	0		
$G_4=\{7.5\}$	5	3.5	0	
$G_5=\{9.5\}$	7	5.5	2	0

表 5-19 中找到对角线之外的最小元素为 1.5，于是把 G_3、G_6 归为一类，记为 $G_7=\{1.5,2.5,4\}$。

新的距离矩阵如表 5-20 所示。

表 5-20 新的距离矩阵（二）

	$G_7=\{1.5,2.5,4\}$	$G_4=\{7.5\}$	$G_5=\{9.5\}$
$G_7=\{1.5,2.5,4\}$	0		
$G_4=\{7.5\}$	3.5	0	
$G_5=\{9.5\}$	5.5	2	0

表 5-20 中找到对角线之外的最小元素为 2，于是把 G_4、G_5 归为一类，记为 $G_8=\{7.5,$

9.5}。新的距离矩阵如表 5-21 所示。

表 5-21　新的距离矩阵(三)

	$G_7=\{1.5,2.5,4\}$	$G_8=\{7.5,9.5\}$
$G_7=\{1.5,2.5,4\}$	0	
$G_8=\{7.5,9.5\}$	3.5	0

最终 G_7、G_8 归为一类。上面的聚类过程表示为图 5-12。

聚类方法通过比较距离的大小逐步归类,最终合并为一个类,分类的结果也是唯一的,同时在样品数即样本容量比较大时计算量也越来越大,为了提升效率,克服计算速度变慢的问题,产生了动态聚类方法,即 K 均值聚类法。在采取 K 均值聚类法之前,需要根据实际问题给出分类的个数,在每类中选择有代表性的样品,这样的样品称为聚点。

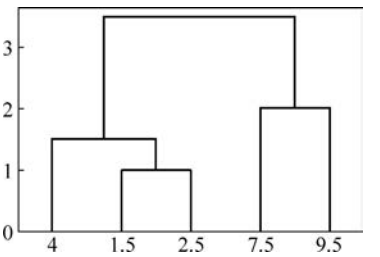

图 5-12　最短距离聚类示例

K 均值聚类的步骤如下。

步骤 1:所有样品初始化为 m 个类,初始聚点的集合记为

$$W^{(0)}=\{x_1^{(0)},x_2^{(0)},\cdots,x_m^{(0)}\}$$

并记 $G_i^{(0)}=\{x:d(x,x_i^{(0)})\leqslant d(x,x_j^{(0)}),j=1,2,\cdots,m,j\neq i\},i=1,2,\cdots,m$。则初始分类表示为 $G^{(0)}=\{G_1^{(0)},G_2^{(0)},\cdots,G_m^{(0)}\}$。

步骤 2:由 $G^{(0)}$ 开始,计算 $x_i^{(1)}=\dfrac{1}{n_i}\sum\limits_{x_l\in G_i^{(0)}}x_l,i=1,2,\cdots,m$,这里 n_i 表示第 i 个类中 $G_i^{(0)}$ 的样品数。聚点的新集合为 $W^{(1)}=\{x_1^{(1)},x_2^{(1)},\cdots,x_m^{(1)}\}$,从 $W^{(1)}$ 开始再进行分类,将样品作新的分类,记

$$G_i^{(1)}=\{x:d(x,x_i^{(1)})\leqslant d(x,x_j^{(1)}),\quad j=1,2,\cdots,m,j\neq i\},i=1,2,\cdots,m$$

于是,类中的样品进行了更新 $G^{(1)}=\{G_1^{(1)},G_2^{(1)},\cdots,G_m^{(1)}\}$。

重复上述两个步骤,直到分类稳定为止。

5.3.5　Python 示例

例 5.16　对某地区 100 名青少年的身高进行了统计,数据如表 5-22 所示。

表 5-22　青少年身高数据

人　数	身高/cm	人　数	身高/cm
1	153	6	161
3	156	7	162
2	157	6	163
1	159	10	164
4	160	8	165

续表

人　数	身高/cm	人　数	身高/cm
7	166	7	173
5	167	3	174
7	168	2	176
5	169	1	178
6	170	1	180
3	171	1	181
4	172		

在显著性水平 $\alpha=0.05$ 下是否可以认为该地区青少年身高服从正态分布？

解　由数据可知本题是对收集的数据进行正态性检验，为了输入数据方便，将表 5-22 中数据去掉中文标签，以列的形式存于文件 normtest.xlsx 中。程序如下。

```
#Python 代码
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
df = pd.read_excel("normtest.xlsx ")
data = np.empty([0,0])
for i in range(len(df['f'])):
data = np.append(data,np.tile(df['value'][i],df['f'][i]))
fig, ax = plt.subplots(figsize = (10, 8))
sns.distplot(data,ax = ax)
ax.set_xlabel('身高', fontsize = 16)
ax.set_ylabel('频率', fontsize = 16)
ax.set_title("直方图", fontsize = 18)
```

程序运行结果如图 5-13 所示。

可以看出，上述直方图与正态分布曲线比较近似。为了进一步直观判断，画出 Q-Q 图，代码如下。

```
#Python 代码(续)
import statsmodels.api as sm
fig, axs = plt.subplots(figsize = (10, 8));sm.qqplot(data, line = 's',ax = axs)
axs.set_xlabel('理论分位数');axs.set_ylabel('样本分位数')
axs.set_title('Q-Q 图')
```

运行结果如图 5-14 所示。

由图 5-14 可看出，基本上生成了一条直线，初步判断数据服从正态分布，为了给出更加可信的结果，进行 W 检验，程序如下。

```
#Python 代码(续)
from scipy.stats import shapiro
shapiro(data)
```

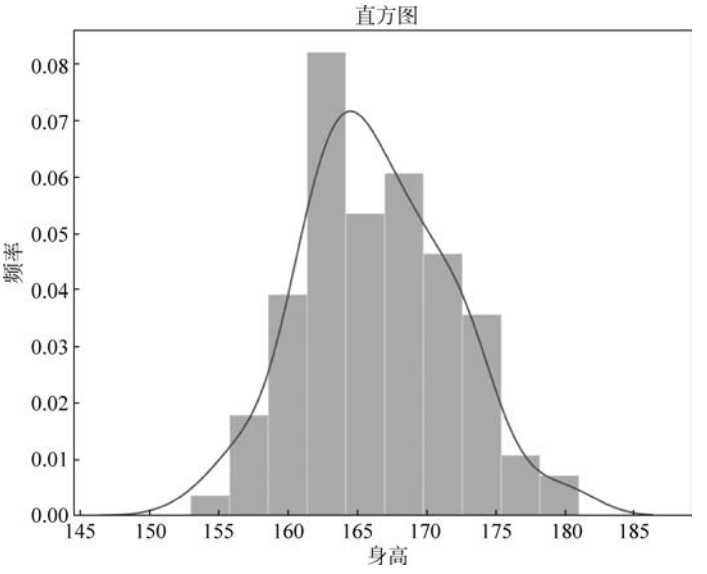

图 5-13　身高数据生成的直方图

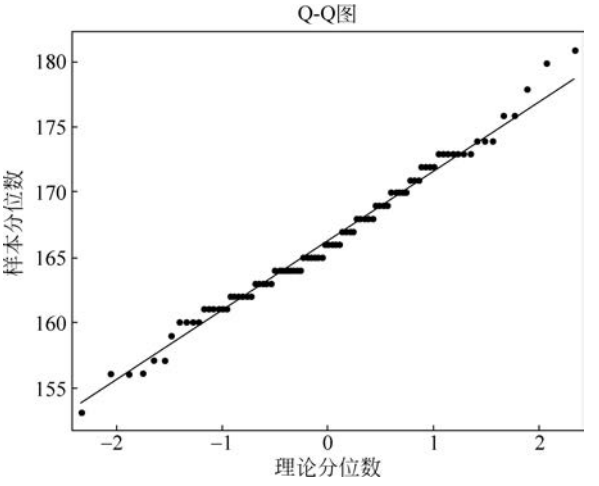

图 5-14　身高数据的 Q-Q 图

运行结果为

```
In [289]: from scipy.stats import shapiro
    ...: shapiro(data)
Out[289]: (0.9882553815841675, 0.5272228717803955)
```

上述结果表明，做出接受原假设(样本数据的分布与正态分布无显著差异)的结论。

例 5.17　为考察某化工产品的得率是否受到酸碱值和添加剂浓度的影响，在四个酸碱值水平和三个添加剂浓度水平下进行实验，数据如表 5-23 所示。

表 5-23 两个因素下的实验数据

酸碱值	添加剂浓度		
	B1	B2	B3
A1	3.5	2.3	2
A2	2.6	2	1.9
A3	2	1.5	1.2
A4	1.4	0.8	0.3

解

```
#Python 代码
import pandas as pd
import statsmodels.api as sm
df = pd.DataFrame([[1,1,3.5],
                   [1,2,2.3],
                   [1,3,2],
                   [2,1,2.6],
                   [2,2,2],
                   [2,3,1.9],
                   [3,1,2],
                   [3,2,1.5],
                   [3,3,1.2],
                   [4,1,1.4],
                   [4,2,0.8],
                   [4,3,0.3]],columns=['A','B','data'])
model = sm.formula.ols("data~C(A)+C(B)", df[['A','B','data']]).fit()
anovat = sm.stats.anova_lm(model)
print(anovat)
```

运行结果如下。

```
             df      sum_sq     mean_sq          F        PR(>F)
C(A)        3.0    5.289167    1.763056  40.948387      0.000217
C(B)        2.0    2.221667    1.110833  25.800000      0.001130
Residual    6.0    0.258333    0.043056        NaN           NaN
```

由 PR(>F)的值可知,在显著水平 0.05 下,因素 A 和因素 B 均对实验结果产生了显著影响。

例 5.18 (续例 5.15)对 5 个样品测得某指标值为 1.5,2.5,4,7.5,9.5,下面给出用最短距离方法对其进行聚类分析的 Python 代码。

解

```
#Python 代码
from sklearn import preprocessing as pp
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt
data = np.array([[1.5],[2.5],[4],[7.5],[9.5]])
data1 = [1.5,2.5,4,7.5,9.5]
```

```
dist = sch.distance.pdist(data,'cityblock')          #样品之间的绝对距离
#dist_matrix = sch.distance.squareform(data)         #距离的矩阵形式
spec = sch.linkage(dist)
#print(spec)
sch.dendrogram(spec,labels = data1)
plt.show()
```

第 5 章习题

1. 自行在网络中搜索某一汽车品牌 2020 年的月销售数量和 2021 年第一季度的销售数量，要求解答以下问题。

（1）对 2020 年的 12 个月销售数量进行简单统计描述分析，给出各项数据指标和特征。

（2）画出 2020 年全年销售数据的直方图。

（3）对 2021 年和 2020 年的第一季度的销售数据进行对比分析，给出统计结论。

2. 对某厂家生产的零部件抽取两组，取得如下测定值。

A：8.6，10.0，9.9，8.8，9.1，9.1

B：8.7，8.4，9.2，8.9，7.4，8.0，7.3，8.1，6.8

已知 A 组测定值无系统误差，试检验 B 组测定值是否有系统误差。

3. 在某放射性粒子实验过程中，通过捕捉粒子进行计数来检验装置是否正常工作。具体过程是粒子从信号发送端通过信道传送到接收端，接收端放置粒子分离器，并将分离后的粒子穿过金箔后，射到荧光屏上产生一个个的闪光点，这些闪光点可用显微镜来观察。粒子计数如表 5-24 所示。由于物理环境中的噪声及其他不可控因素，粒子发生器可能受到这些因素影响而产生异常，对收集的数据利用统计方法进行异常值检测是其中一种有效方法。

（1）对粒子数量进行异常值检测，并给出检测依据。

（2）如果存在异常值，请选择合适的方法进行替换，并说明原因。

（3）若不存在异常值或经过相应的替换后，检验粒子数是否服从正态分布。

表 5-24　放射粒子数据

序号	数量	序号	数量	序号	数量	序号	数量	序号	数量
1	5	10	8	19	60	28	122	37	70
2	11	11	3	20	39	29	103	38	81
3	16	12	0	21	28	30	73	39	111
4	23	13	0	22	26	31	47	40	101
5	36	14	2	23	22	32	35	41	73
6	58	15	11	24	11	33	11	42	40
7	29	16	27	25	21	34	5	43	20
8	20	17	47	26	40	35	16	44	16
9	10	18	63	27	78	36	34	45	5

续表

序号	数量	序号	数量	序号	数量	序号	数量	序号	数量
46	11	85	10.2	124	1.8	163	59.1	202	2.7
47	22	86	24.1	125	8.5	164	44	203	5
48	40	87	82.9	126	16.6	165	47	204	24.4
49	60	88	132	127	36.3	166	30.5	205	42
50	80.9	89	130.9	128	49.6	167	16.3	206	63.5
51	83.4	90	118.1	129	64.2	168	7.3	207	53.8
52	47.7	91	89.9	130	67	169	37.6	208	62
53	47.8	92	66.6	131	70.9	170	74	209	48.5
54	30.7	93	60	132	47.8	171	139	210	43.9
55	12.2	94	46.9	133	27.5	172	111.2	211	18.6
56	9.6	95	41	134	8.5	173	101.6	212	5.7
57	10.2	96	21.3	135	13.2	174	66.2	213	3.6
58	32.4	97	16	136	56.9	175	44.7	214	1.4
59	47.6	98	6.4	137	121.5	176	17	215	9.6
60	54	99	4.1	138	138.3	177	11.3	216	47.4
61	62.9	100	6.8	139	103.2	178	12.4	217	57.1
62	85.9	101	14.5	140	85.7	179	3.4	218	103.9
63	61.2	102	34	141	64.6	180	6	219	80.6
64	45.1	103	45	142	36.7	181	32.3	220	63.6
65	36.4	104	43.1	143	24.2	182	54.3	221	37.6
66	20.9	105	47.5	144	10.7	183	59.7	222	26.1
67	11.4	106	42.2	145	15	184	63.7	223	14.2
68	37.8	107	28.1	146	40.1	185	63.5	224	5.8
69	69.8	108	10.1	147	61.5	186	52.2	225	16.7
70	106.1	109	8.1	148	98.5	187	25.4	226	44.3
71	100.8	110	2.5	149	124.7	188	13.1	227	63.9
72	81.6	111	0	150	96.3	189	6.8	228	69
73	66.5	112	1.4	151	66.6	190	6.3	229	77.8
74	34.8	113	5	152	64.5	191	7.1	230	64.9
75	30.6	114	12.2	153	54.1	192	35.6	231	35.7
76	7	115	13.9	154	39	193	73	232	21.2
77	19.8	116	35.4	155	20.6	194	85.1	233	11.1
78	92.5	117	45.8	156	6.7	195	78	234	5.7
79	154.4	118	41.1	157	4.3	196	64	235	8.7
80	125.9	119	30.1	158	22.7	197	41.8	236	36.1
81	84.8	120	23.9	159	54.8	198	26.2	237	79.7
82	68.1	121	15.6	160	93.8	199	26.7	238	114.4
83	38.5	122	6.6	161	95.8	200	12.1	239	109.6
84	22.8	123	4	162	77.2	201	9.5	240	88.8

续表

序号	数量	序号	数量	序号	数量	序号	数量	序号	数量
241	67.8	251	83.9	261	112.3	271	104.5	281	154.7
242	47.5	252	69.4	262	53.9	272	66.6	282	140.5
243	30.6	253	31.5	263	37.5	273	68.9	283	115.9
244	16.3	254	13.9	264	27.9	274	38	284	66.6
245	9.6	255	4.4	265	10.2	275	34.5	285	45.9
246	33.2	256	38	266	15.1	276	15.5	286	17.9
247	92.6	257	141.7	267	47	277	12.6	287	13.4
248	151.6	258	190.2	268	93.8	278	27.5	288	29.2
249	136.3	259	184.8	269	105.9	279	92.5	289	100.2
250	134.7	260	159	270	105.5	280	155.4		

4. 猫粮生产商为丰富产品线，对生产的猫粮在满足各种营养成分的前提下，调整牛肉和鸡肉的比例，共推出三种类型的猫粮，即 A、B、C。表 5-25 所示是猫在三种类型的猫粮喂养下的体重数据，试根据这些数据判断不同类型的猫粮是否对体重有显著性影响。

表 5-25　不同类型的猫粮喂养数据

类型	A	B	C	B	B	A	C	A	C	C	A
体重	45.5	64.3	61.2	57	64.8	68.1	61.6	54.4	66.3	75.8	55.6
类型	C	A	B	C	C	A	B	B	B	B	
体重	60.8	47.8	63	64.8	60.6	52.3	48.9	49.8	60	61	

5. (续例 5.15)对 5 个样品测得某指标值为 1.5，2.5，4，7.5，9.5，分别用最长距离和类平均距离方法对其进行聚类分析。