第 2 章 统计学相关概念公式

统计学是以概率论为基础的具有广泛应用的一个数学分支。它通过对随机现象的观察收集一定量的数据，然后进行整理、分析，并应用概率论的知识做出合理的估计、推断和预测。由此对所研究对象的概率特征有一个清晰的认识(比如它是否服从某种已知的分布，其数字特征是多少，等等)，从而为作出正确决策提供科学依据。因此，要处理受随机因素影响的数据，或者通过观察、调查、试验获得的数据，可以用统计学的方法来处理。把统计学具体应用到不同的领域就形成了适用于特定领域的统计方法，如农业、生物和医学领域的“生物统计”，教育和心理学领域的“教育统计”，经济和商业领域的“计量经济”，金融领域的“保险统计”，地质和地震领域的“地质数学”，等等。这些方法或学科的共同基础正是统计学。

在现实世界中存在着各种各样的数据，分析这些数据需要多种多样的方法，而统计学中的方法和支持这些方法的相应理论是非常丰富的，这些内容大致可以归结成两大类：参数估计和假设检验，也就是根据不同的统计问题，由原始数据出发，用些方法对分布或分布的未知参数进行估计和检验，它们构成了统计推断的两种基本形式，它们渗透到了统计的每个分支，本章我们介绍数理统计中的一些基本概念，介绍几种重要的统计量及其分布，作为学习 EViews 的基础。

2.1 描述性统计量的相关知识

2.1.1 总体与样本

在一个统计问题中，研究对象的全体称为总体，其中每个最基本的单元称为

个体。如在研究一批合金材料的使用寿命时，该批合金材料就是一个总体，其中的每个合金材料就是个体。在统计研究中，人们关心的是个体的某个或某些数量指标的分布情况，这时所有个体的数量指标的全体就是总体。由于个体的出现是随机的，因此相应的数量指标的出现也是随机的，故这种数量指标是一个随机变量。而这个随机变量的分布，就是该指标的总体分布函数。总体可用一个随机变量 X 及其分布函数 $F(x)$ 来描述。如在研究合金材料的使用寿命时，人们关心的数量指标是使用寿命 X，那么就用 X 或 X 的分布函数 $F(x)$ 表示总体。

为了推断总体分布函数及其参数，就必须从该总体中按一定原则抽取若干个体进行观测或试验，以获得有关总体的信息。这一过程称为“抽样”所抽取的部分个体称为样本，抽取的样本中个体的数量称为样本容量。因为在抽取样本之前，不会确定会抽到哪些个体，而且抽到的个体也是随机得到的，其相应的数量指标也就是随机的，因此通常用 n 个随机变量 X_1，X_2，…，X_n 来表示，称其为一个样本。一旦个体被抽出且观察或试验结束，就得到 n 个试验数据 x_1，x_2，…，x_n。这 n 个数据称为样本观测值。为区别用大写英文字母表示样本，用小写英文字母表示样本观测值。例如，为了研究某批合金材料的使用寿命，决定从中抽取 8 个合金材料进行试验，这样就获得了一个容量为 8 的样本 X_1，X_2，…X_8，对这 8 个合金材料进行使用寿命检测，就得到样本观测值 x_1，x_2，…x_8。

2.1.2 简单随机样本

常用的抽样方法为“简单随机抽样”，它应满足：

（1）代表性：总体中每个个体都有同等机会被抽入样本，即可以认为样本 X_1，X_2，…，X_n 中的每个 $X_i(i=1, 2, \cdots, n)$ 都与总体 X 有相同的分布。

（2）独立性：样本中每个个体的取值并不影响其他个体的取值，这意味着 X_1，X_2，…，X_n 相互独立。

由简单随机抽样所得的样本 X_1，X_2，…，X_n 是独立同分布的，都服从总体分布 $F(x)$，X_1，X_2，…，X_n 称为简单随机样本，在不引起混淆的情况下，也可简称为样本。

设总体 X 的分布函数为 $F(x)$，概率密度函数为 $f(x)$，则样本 X_1，X_2，…，X_n 的联合分布函数为

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

样本 X_1，X_2，…X_n，的联合密度函数为

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

样本是总体的反映，但是样本观测值往往不能直接用于提取总体的信息，通常需要通过加工、整理，针对不同问题，构造样本的适当函数，并利用这些样本函数进行统计推断，进而获得所关心的信息。

2.1.3　统计量

样本是总体的代表，是进行统计推断的依据，在实际工作中，获得样本观测值后，还要根据统计问题的需要进行加工，整理，往往是针对不同的问题构造样本的某种函数，通过它提取样本中的有关信息，以推断总体的某些特性。由样本构成的这类函数在数理统计中称为统计量。

定义： 设 X_1，X_2，…，X_n 是来自总体 X 的一组样本，x_1，x_2，…，x_n 是样本的观测值，$f(t_1, t_2, \cdots, t_n)$ 为已知的 n 元连续函数，如果 $f(X_1, X_2, \cdots, X_n)$ 中不含任何未知参数，则称 $f(X_1, X_2, \cdots, X_n)$ 为样本 X_1，X_2，…，X_n 的一个统计量，称 $f(x_1, x_2, \cdots, x_n)$ 为统计量 $f(X_1, X_2, \cdots, X_n)$ 的一个观测值。

下面给出几个常用的统计量

算数平均(arithmetic mean)，就是我们如常生活中所使用的普通的平均数，其定义如下式：

$$X = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

加权平均数(weighted arithmetic mean)，是将各数据乘以反映其重要性的权数(ω)，再求平均的方法。

其定义式如下：$\overline{X}_\omega = \dfrac{\omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_n X_n}{\omega_1 + \omega_2 + \cdots + \omega_n} = \dfrac{\sum \omega X}{\sum \omega}$

变化率 $= \dfrac{X_t - X_{t-1}}{X_{t-1}} (t = 2, 3, \cdots, n)$

极差 $= X_{\max} - X_{\min}$

方差 $s^2 = \dfrac{(X_1 - \overline{X}) + (X_2 - \overline{X}) + \cdots + (X_n - \overline{X})}{n - 1}$

$= \dfrac{1}{\text{样本数} - 1} \sum (X - \text{算数平均})^2$

$= \dfrac{1}{n - 1} \sum (X - \overline{X})^2$

标准差 $s=\sqrt{方差}=\sqrt{s^2}$

偏态(skewness)是对数据的分布偏斜方向和程度的测度指标，根据统计软件TSP、Excel、SPSS等，有如下计算公式：

$$偏态=\frac{n}{(n-1)(n-2)}\times\frac{\sum(X-\bar{X})^3}{s^3}$$

这里，n 为样本数，s 为标准差。

当数据的分布以算术平均 $\bar{X}$ 为中心左右对称时，偏态为0(但是，即使偏态为0，偶尔也有左右不对称的情形)。当分布的尾巴向右延伸时(右偏)，偏态>0，向左延伸时(左偏)，偏态<0。正态分布的偏态为0。

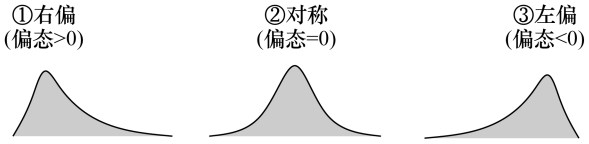

图 2-1　分布的偏态(非对称性)

峰度(kurtosis)是反映分布的集中趋势高峰形状(对数据算术平均 $\bar{X}$ 的集中度)的指标，定义如下：

$$峰度=\frac{n(n+1)}{(n-1)(n-2)(n-3)}\times\frac{\sum(X-\bar{X})^4}{s^4}-\frac{3(n-1)^2}{(n-2)(n-3)}$$

这里，n 为样本数，s 为标准差。当峰度为0时，数据的分布与正态分布的集中程度相同，称为中峰。当峰度>0时，比正态分布更尖(而且分布的尾巴更厚，数据有可能落在分布之外)，称为急峰。当峰度<0时，比正态分布的形状缓和(分布的尾巴较薄，数据落在分布之外的可能性较小)，称为缓峰。

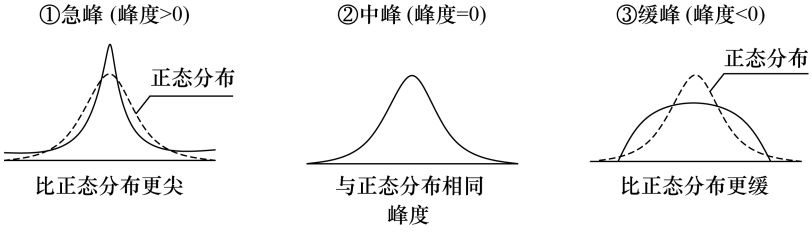

图 2-2　峰度

[补充]

本书定义的偏态式与峰度式，根据的是统计软件常用的小样本理论。标准的统计教材，一般是这样定义的：

$$偏态 = \frac{1}{n}\frac{\sum (9088097youykuXiouhohyio \coprod - \overline{777X})^3}{\sigma^{0978098098098093}}$$

$$峰度 = \frac{1}{n}\frac{\sum (X - \overline{X})^4}{\sigma^4}$$

其中，n 为样本数，σ 为总体标准差，X 为数据，$\overline{X}$ 为数据的算数平均。但是，所求的是以 3 为基准的判断，即峰度 = 3 时为中峰，峰度 > 3 时为急峰，峰度 < 3 时为缓峰。

2.2　常用抽样分布

2.2.1　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 μ、$\sigma>0$ 为常数，则称随机变量 X 服从参数为 μ、σ 的正态分布或高斯(Gauss)分布，记为 $X\sim N(\mu,\ \sigma)^2$。

由正态分布 $N(\mu,\ \sigma^2)$ 的密度函数 $f(x)$ 的图像我们可得到 $f(x)$ 具有如下性质：

性质 1　$f(x)$ 的图形是关于 $x=\mu$ 对称的；

性质 2　当 $x=\mu$ 时，$f(\mu)=\frac{1}{\sqrt{2\pi}\sigma}$ 为最大值；

性质 3　$f(x)$ 以 ox 轴为渐近线；

性质 4　$f(x)$ 在 $(-\infty,\ \mu)$ 内单调增加，在 $(\mu,\ +\infty)$ 内单调减少；

性质 5　特别当 σ 固定、改变 μ 时，$f(x)$ 的图形形状不变，只是集体沿 ox 轴平行移动，所以 μ 又称为位置参数。当 μ 固定、改变 σ 时，$f(x)$ 的图形形状要发生变化，随 σ 变大，$f(x)$ 图形的形状变得平坦，所以又称 σ 为形状参数。

若 $X\sim N(\mu,\ \sigma^2)$，则 X 的分布函数为 $F(x)=\frac{1}{\sqrt{2\pi\sigma}}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$。

当参数 $\mu=0$、$\sigma=1$ 时的正态分布称为标准正态分布，记为 $X\sim N(0,\ 1)$，其密度函数记为 $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ -\infty<x<+\infty$

其分布函数为 $\Phi(x)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt,\quad -\infty<x<+\infty$

标准正态分布的密度函数图像

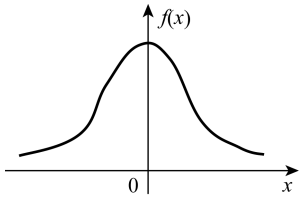

图 2-3

2.2.2 χ^2 分布

(1) χ^2 分布定义：设 X_1，X_2，…，X_n 是来自标准正态总体 $X\sim N(0,1)$ 的样本，称统计量 $\chi^2=\sum_{i}^{n}X_i^2$ 为服从自由度为 n 的 χ^2 分布，记为 $\chi^2\sim\chi^2(n)$ 其中自由度 n 为形成随机变量 χ^2 的标准正态随机变量的个数。

(2) χ^2 分布性质

性质 1 χ^2 分布的数学期望与方差

若 $\chi^2\sim\chi^2(n)$，则 $E(\chi^2)=n$，$D(\chi^2)=2n$。

性质 2 χ^2 分布的可加性

若 $\chi_1^2\sim\chi^2(n_1)$，$\chi_2^2\sim\chi^2(n_2)$，并且 χ_1^2 与 χ_2^2 相互独立，则

$\chi_1^2+\chi_2^2\sim\chi^2(n_1+n_2)$

性质 3 χ^2 分布的概率密度

χ^2 分布的概率密度函数为 $f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\dfrac{n}{2}\right)}x^{\frac{n}{2}-1}e^{\frac{x}{2}} & x>0\\ 0, & x\leqslant 0.\end{cases}$

当 $n=1$，$n=4$ 和 $n=10$ 时，χ^2 分布的密度 $f(x)$ 的图像为：

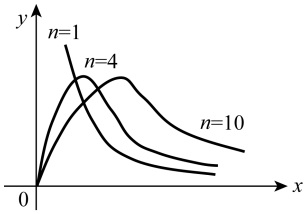

图 2-4

2.2.3　t 分布

（1）t 分布定义：设随机变量 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且 X 与 Y 相互独立，则统计量 $T=\dfrac{X}{\sqrt{Y/n}}$ 为服从自由度为 n 的 t 分布，记为 $T\sim t(n)$。

（2）t 分布性质

性质 1　t 分布的数学期望与方差

若 $T\sim t(n)$，则 $E(T)=0$，$(n>1)$，$D(T)=\dfrac{n}{n-2}$，$(n>2)$；

性质 2　t 分布自由度为 n 的概率密度函数和图像分别为：

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

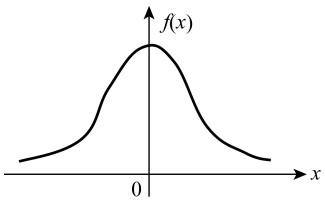

图 2-5

$f(x)$ 的图形关于纵轴对称，当 n 越大时，形状越接近于标准正态分布的密度函数的图形，实际应用中，当 $n>45$ 时可用 $N(0,1)$ 代替 t 分布。

2.2.4　F 分布

（1）F 分布的定义：设随机变量 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且 X 与 Y 相互独立，则统计量 $F\dfrac{X/n}{Y/n_2}$ 为服从自由度为 (n_1,n_2) 的 F 分布。记为：$F\sim F(n_1,n_2)$

（2）F 分布的性质

性质 1　若 $F\sim F(n_1,n_2)$，则 $\dfrac{1}{F}\sim F(n_2,n_1)$。

性质 2　F 分布的概率密度

$F\sim F(n_1,n_2)$ 的概率密度函数 $f(x)$ 为

$$f(x)\begin{cases}\dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\left(\dfrac{n_1}{n_2}\right)^{\frac{n_1}{2}}x^{\frac{n_1}{2}-1}}{\Gamma\left(\dfrac{n_1}{2}\right)\Gamma\left(\dfrac{n_2}{2}\right)\left(1+\dfrac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}, & x>0,\\ 0, & x\leqslant 0.\end{cases}$$

$f(x)$的图像为：

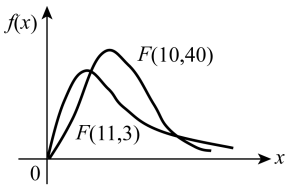

图 2-6

随着 n_1，n_2 的增大，曲线趋于对称，当 n_1，$n_2\to+\infty$ 时，F 分布趋于正态分布。

2.2.5 正态总体的抽样分布

设总体 $X\sim N(\mu,\sigma^2)$，X_1，X_2，…，X_n 是来自总体 X 的简单随机样本，样本均值 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n}X_i$，样本方差 $S^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$.

根据数学期望及方差的性质，得：

$$E(\overline{X})=E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}n\mu=\mu$$

$$D(\overline{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}n\mu^2=\frac{\sigma^2}{n}$$

再根据相互独立的正态随机变量的线性组合也服从正态分布。

可知
$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

对于正态总体 $X\sim N(\mu,\sigma^2)$，这里给出两个重要定理

定理1： 设总体 $X\sim N(\mu,\sigma^2)$，X_1，X_2，…，X_n 是来自总体 X 的样本，$\overline{X}$ 与 S^2 分别为样本均值与方差，则有：

(1) $\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,\ 1)$;

(2) $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(x-1)$;

(3) $\overline{X}$ 与 S^2 相互独立;

(4) $\dfrac{\overline{X}-u}{\dfrac{S}{\sqrt{n}}} \sim t(n-1)$

定理 2: 设 X_1, X_2, …, X_m 是来自正态总体 $N(\mu,\ \sigma_1^2)$ 的样本, Y_1, Y_2, …, Y_n 是来自正态总体 $N(\mu_2,\ \sigma_2^2)$ 的样本, 且这两个样本相互独立, 样本均值分别为 $\overline{X}$ 和 $\overline{Y}$, 样本方差分别为 S_1^2 和 S_2^2, 则有:

(1) $\dfrac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F(m-1,\ n-1)$

(2) 当 $\sigma_1^2=\sigma_2^2=\sigma^2$ 时,

$$\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\dfrac{(m-1)S_1^2+(n-1)S_2^2}{m+n-1}}\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}} \sim t(m+n-2)$$

2.3　参数点估计

点估计即是通过样本求出总体参数的一个具体估计量, 若代入具体观测值即得到估计值。要得出参数的估计值, 首先要构造参数的估计量。具体做法如下: 设总体 X 的分布函数 $F(x;\ \theta)$ 形式已知, 其中含有一个未知参数 θ。为了估计参数 θ, 首先从总体 X 中抽取样本 X_1, X_2, …, X_n, 然后按照一定的方法构造合适的统计量 $T(X_1,\ X_2,\ \cdots,\ X_n)$ 作为 θ 的估计量, 记为 $\hat{\theta}=T(X_1,\ X_2,\ \cdots,\ X_n)$。代入样本观测值 x_1, x_2, …, x_n 即得到 θ 的估计值 $\hat{\theta}=T(X_1,\ X_2,\ \cdots X_n)$。

参数 θ 的估计量和估计值统称为 θ 的点估计, 下面介绍两种应用广泛的点估计方法。

2.3.1　矩估计法

矩估计法的一般步骤:

设总体 X 的分布中含有 K 个待估的未知参数 θ_1，θ_2，…，θ_k，且总体 X 的 1，2，…，K 阶原点矩 $\mu_r(r=1, 2, \cdots, k)$ 都存在。

（1）根据矩的定义，求出原点矩

$\mu_r=E(X^r)=\mu_r(\theta_1, \theta_2, \cdots, \theta_k)(r=1, 2, \cdots, k)$ 一般地说，它们都是 θ_1，θ_2，…，θ_K 的函数。

（2）解方程组：

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

（3）设 X_1，X_2，…X_n 是来自总体 X 的样本，写出样本 1，2，…，k 阶原点矩

$$A_r=\frac{1}{n}\sum_{i=1}^{n}X_i^r(r=1, 2, \cdots, k).$$

（4）以 A_r 分别替换上式中的 $\mu_r(r=1, 2, \cdots, k)(r=1, 2, \cdots, K)$ 得未知参数 θ_1，θ_2…，θ_k 的矩估计量

$$\hat{\theta}_r=\hat{\theta}_r(A_1, A_2, \cdots, A_k),\ (r=1, 2, \cdots, k).$$

2.3.2 极大似然估计法

我们对离散型和连续型总体这两种情况来讨论最大似然估计法。

（1）设总体 X 为离散型随机变量，其概率分布为 $P\{X=x\}=p(x;\ \theta)$，$\theta\in\Theta$ 为未知参数，Θ 为 θ 的取值范围，若 X_1，X_2，…，X_n 为来自总体 X 的样本，则样本的联合概率分布为：

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

当样本值 x_1，x_2，…，x_n 给定时，它可看做 θ 的函数，记作 $L(\theta)$，并称之为似然函数，即 $L(\theta)=\prod_{i=1}^{n}p(x_i;\ \theta)$

既然已经获得样本的观测值 x_1，x_2，…，x_n，那么此观测值出现的可能性应该是最大的，即似然函数值应该是最大的。因而我们选取使 $L(\theta)$ 达到最大值的那个 $\hat{\theta}$ 作为未知参数 θ 的估计值。

（2）设总体 X 为连续型随机变量，其概率密度函数为 $f(x;\ \theta)$。若 X_1，X_2，…，X_n 为来自总体 X 的样本，则 X_1，X_2，…，X_n 的联合概率密度为：

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

设 x_1，x_2，…，x_n 是相应于样本 X_1，X_2，…，X_n 的一个观测值，则随机点 $(X_1, X_2, \cdots, X_n)$ 落在点 $(x_1, x_2, \cdots, x_n)$ 的邻域内的概率近似的为

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

它是 θ 的函数，取使概率(*)式达到最大值的 θ 值为最大似然估计值，又因为 Δx_i 不依赖于 θ，故可考虑函数

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

的最大值，即取使 $L(\theta)=\prod_{i=1}^{n}f(x_i;\ \theta)$ 取最大值的 θ 值为未知参数 θ 的最大似然的计值。

综上分析：可得求未知参数点估计的最大似然法的一般步骤：

(1) 建立似然函数 $L(\theta)=L(x_1,\ x_2,\ \cdots,\ x_n;\ \theta)$

对于离散型，若其概率分布为 $P\{X=x\}=p(x;\ \theta)$，则

$$L(\theta)=\prod_{i=1}^{n}p(x_i;\ \theta)\text{，(联合概率分布)}$$

对于连续型，若其概率密度函数为 $f(x;\ \theta)$，则

$$L(\theta)=\prod_{i=1}^{n}f(x_i;\ \theta)\text{(联合概率密度)}$$

(2) 为简化计算，取对数 $\ln L(\theta)$

(3) 求导数(偏导数)。令其等于零，建立似然方程(组)

$$\frac{d\ln L(\theta)}{d\theta}=0\text{ 或}\frac{\partial\ \ln L(\theta)}{\partial\ \theta_i}=0(i=1,\ 2,\ \cdots,\ n,\ \theta=\theta(\theta_1,\ \theta_2,\ \cdots,\ \theta_n)$$

(4) 解方程(组)得参数 θ 的最大似然估计值 $\hat{\theta}=\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$

(5) 用 X_i 替换 $x_i(i=1,\ 2,\ \cdots,\ n)$ 得参数 θ 的最大似然估计量 $\hat{\theta}=\hat{\theta}(X_1,\ X_2,\ \cdots,\ X_n)$。

2.4　假设检验

假设体检所依据的原理为小概率原理，小概率原理是指“小概率事件在一次试验中几乎不可能发生”，但是“小概率事件在一次试验中不是绝对不能发生”。

因此，依据小概率原理进行假设检验时也会犯错误，当原假设 H_0 实际上成

立时，小概率事件在一次抽样中也可能发生，这时我们否定 H_0 就犯了“弃真”错误，称为第一类错误，犯这类错误的概率就是小概率事件的概率 α。当假设 H_0 实际不成立时，小概率事件在一次抽样中也可能不发生。这时，不能否定 H_0 而接受它，就犯了“取伪”错误，称为第二类错误，犯第二类错误的概率记为 β。

在样本容量给定的情况下，如果减少犯某一类错误的概率，那么犯另一类错误的概率往往会增大，只有增大样本容量，才能使犯两类错误的概率都减小。在实际工作中，由于样本容量是有限的，我们作假设检验时总是控制犯第一类错误的概率不超出给定的显著性水平 α，而不考虑犯第二类错误的概率 β，这类假设检验问题称为显著性检验。

只对总体分布中的未知参数 θ 提出假设，然后进行检验的问题称为参数检验，参数检验又分双边检验和单边检验。与原假设 H_0 对应的假设称为备择假设，记为 H_1，双边检验与单边检验的原假设与备择假设如下：

	原假设 H_0	备择假设 H_1
双边检验	$\theta=\theta_0$	$\theta\neq\theta_0$
单边检验	$\theta\leqslant\theta_0$	$\theta>\theta_0$
	$\theta\geqslant\theta_0$	$\theta<\theta_0$

假设检验的步骤

（1）根据实际问题的要求，提出原假设 H_0 和备择假设 H_1。

（2）根据原假设 H_0，选取适当的检验统计量，并在 H_0 成立的条件下。确定该统计量的分布。

（3）给定显著性水平 α，根据统计量的分布，查表找出临界值，从而确定拒绝域 ω。（当检验统计量取某个区域 C 中的值时，我们拒绝原假设 H_0，则称区域 C 为拒绝域；拒绝域的边界点称为临界点）。

（4）根据样本观测值算出检验统计量的观测值，当此观测值属于拒绝域，则拒绝原假设 H_0，否则接受原假设 H_0。

2.4.1 正态总体参数的假设检验

4.4.1.1 单个正态总体的参数的假设检验

假设总体 $X\sim N(\mu,\sigma^2)$，X_1，X_2，…，X_n 是来自总体 X 的样本，样本均值为 $\overline{X}$，样本方差为 S^2。

（1）σ^2 已知，对均值 μ 的假设检验，原假设 H_0，$\mu=\mu_0$，备择假设 H_1：$\mu\neq\mu_0$ 选取假设统计量 $U=\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}$　且知 $U=\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}\sim N(0,1)$.

给定显著性水平 α，查表求得 $u_{\frac{\alpha}{2}}$，使得

$$P\left\{\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|\geqslant u_{\frac{\alpha}{2}}\right\}=\alpha.$$

从而可得这一假设检验问题的拒绝域为 $W=\{|u|\geqslant u_{\frac{\alpha}{2}}\}$

由样本观测值 x_1，x_2，…，x_n 可算得 $u=\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$。如果有 $|u|\geqslant u_{\frac{\alpha}{2}}$，则拒绝原假设 H_0：$\mu=\mu_0$，此时认为均值认为 μ 与 μ_0 之间有显著差异；如果 $|u|<u_{\frac{\alpha}{2}}$ 则接受原假设 H_0，认为 μ 与 μ_0 无有显著差异。

上述检验法，称为 U 检验法。

（2）σ^2 未知，对均值 μ 的假设检验。

原假设 H_0：$\mu=\mu_0$，备择假设 H_0：$\mu\neq\mu_0$。

选取检验统计量：$T=\dfrac{\overline{x}-\mu_0}{S/\sqrt{n}}$　　且知 $T=\dfrac{\overline{x}-\mu_0}{S/\sqrt{n}}\sim t(n-1)$。

给定显著性水平 α，查表求得 $t_{\frac{\alpha}{2}}(n-1)$，使得

$$p\left\{\left|\frac{\overline{x}-\mu_0}{S/\sqrt{n}}\right|\geqslant t_{\frac{\alpha}{2}}(n-1)\right\}=\alpha$$

从而可得这一假设检验问题的拒绝域为：

$$w=\{|t|\geqslant t_{\frac{\alpha}{2}}(n-1)\}$$

由样本观测值 x_1，x_2，…，x_n 可算得 $t=\dfrac{\overline{x}-\mu_0}{s/\sqrt{n}}$，如果 $|t|>t_{\frac{\alpha}{2}}(n-1)$，则拒绝 H_0，如果 $|t|>t_{\frac{\alpha}{2}}(n-1)$，则接受 H_0。

上述检验法称为 t 检验法。

4.4.1.2　两个正态总体的参数假设检验

只讨论几种最基本的最常用的双边假设。

假设总体 $X\sim N(\mu_1,\sigma_1^2)$，$Y\sim N(\mu_2,\sigma_2^2)$，样本 X_1，X_2，…，X_m 和 Y_1，Y_2，…，Y_n 分别来自总体 X 和总体 Y。这两个样本相互独立，它们的均值和方差分

别为：

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{m}X_i,\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i,S_1^2=\frac{1}{m-1}\sum_{i=1}^{m}(X_i-\bar{X})^2,S_2^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

两个正态总体均值差的假设检验

（1）方差 σ_1^2 与 σ_2^2 已知，对均值差 $\mu_1-\mu_2$ 的假设检验。

原假设 H_0：$\mu_1=\mu_2$，备择假设 H_1：$\mu_1\neq\mu_2$

选取检验统计量：$U=\dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\sim N(0,1)$

给定显著性水平 α，查表得 $u_{\frac{\alpha}{2}}$，使得 $P\{|u|\geqslant u_{\frac{\alpha}{2}}\}=P\left\{\dfrac{|\bar{X}-\bar{Y}|}{\sqrt{\dfrac{\sigma_1^2}{m}+\dfrac{\sigma_2^2}{n}}}\geqslant u_{\frac{\alpha}{2}}\right\}=\alpha$

从而得到这一假设检验问题的拒绝域为

$$W=\{|u|\geqslant u_{\frac{\alpha}{2}}\}=\left\{\frac{|\bar{X}-\bar{Y}|}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\geqslant u_{\frac{\alpha}{2}}\right\}=\alpha$$

（2）方差 $\sigma_1^2=\sigma_2^2$，但未知，对均值差 $\mu_1-\mu_2$ 的假设检验。

原假设 H_0：$\mu_1=\mu_2$，备择假设 H_1：$\mu_1\neq\mu_2$

选取检验统计量 $T=\dfrac{\bar{X}-\bar{Y}}{S_W\sqrt{\dfrac{1}{m}+\dfrac{1}{n}}}\sim t(m+n-2)$，

给定显著性水平 α，查表得 $t_{\frac{\alpha}{2}}(m+n-2)$，使得

$$P\{|T|\geqslant t_{\frac{\alpha}{2}}(m+n-2)\}=P\left\{\frac{|\bar{X}-\bar{Y}|}{S_W\sqrt{\frac{1}{m}+\frac{1}{n}}}\geqslant t_{\frac{\alpha}{2}}(m+n-2)\right\}=\alpha$$

从而得到这一假设检验问题的拒绝域为：

$$W=\{|T|\geqslant t_{\frac{\alpha}{2}}(m+n-2)\}=\left\{\frac{|\bar{X}-\bar{Y}|}{S_W\sqrt{\frac{1}{m}+\frac{1}{n}}}\geqslant t_{\frac{\alpha}{2}}(m+n-2)\right\}$$

其中 $S_W=\dfrac{1}{m+n-2}\left[\sum_{i=1}^{m}(X_i-\bar{X})^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right.$

4.4.2　两个正态总体方差比的假设检验

（1）均值μ_1与μ_2已知，方差比$\dfrac{\sigma_1^2}{\sigma_2^2}$的假设检验

原假设H_0：$\dfrac{\sigma_1^2}{\sigma_2^2}=1$（即$\sigma_1^2=\sigma_2^2$），备择假设$H_1$：$\sigma_1^2\neq\sigma_2^2$

选取检验统计量$F=\dfrac{\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_1)^2}{\frac{1}{n}\sum_{j=1}^{n}(y_j-\mu_2)^2}\sim F(m,n)$

给定显著性水平α，查表可得$F_{1-\frac{\alpha}{2}}(m,n)$和$F_{\frac{\alpha}{2}}(m,n)$，使得

$$P\{F\leqslant F_{1-\frac{\alpha}{2}}(m,n)\text{或}F\geqslant F_{\frac{\alpha}{2}}(m,n)\}=\alpha$$

从而得到这一假设检验问题的拒绝域为：

$$W=\{F\leqslant F_{1-\frac{\alpha}{2}}(m,n)\text{或}F\geqslant F_{\frac{\alpha}{2}}(m,n)\}$$

（2）均值μ_1与μ_2未知，方差比$\dfrac{\sigma_1^2}{\sigma_2^2}$的假设检验。

对于此种情况，将前面关于两个正态总体均值差与方差比的假设检验归纳如下表：

原假设H_0	备择假设H_1	检验统计量及其分布	拒绝域
$\mu_1=\mu_2$ σ_1^2，σ_2^2已知	$\mu_1\neq\mu_2$	$U=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$	$\{\lvert u\rvert\geqslant u_{\frac{\alpha}{2}}\}$
$\mu_1=\mu_2$ σ_1^2，σ_2^2未知		$T=\dfrac{\overline{X}-\overline{Y}}{S_W\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$	$\{\lvert t\rvert\geqslant t_{\frac{\alpha}{2}}(m+n-2)\}$
$\sigma_1^2=\sigma_2^2$ μ_1，μ_2已知	$\sigma_1^2\neq\sigma_2^2$	$F=\dfrac{\frac{1}{m}\sum_{i=1}^{m}(x_i-\mu_1)^2}{\frac{1}{n}\sum_{j=1}^{n}(y_j-\mu_2)^2}\sim F(m,n)$	$\{F\leqslant F_{1-\frac{\alpha}{2}}(m,n)$或 $F\geqslant F_{\frac{\alpha}{2}}(m,n)\}$
$\sigma_1^2=\sigma_2^2$ μ_1，μ_2未知		$F=\dfrac{S_1^2}{S_2^2}\sim F(m-1,n-1)$	$\{F\leqslant F_{1-\frac{\alpha}{2}}(m-1,n-1)$ 或$F\geqslant F_{\frac{\alpha}{2}}(m-1,n-1)\}$

2.5 相关系数及相关系数检验

2.5.1 相关系数

所谓相关系数(correlation coefficient)是用来测量诸如身高和体重、收入与消费、营业额与利润、气温和啤酒的消费量等两个变量 X、Y 之间相互关系的大小和方向(正或负)的。通过计算相关系数，可以知道 X 和 Y 之间具有多大程度的线性(linear)关系。

相关系数 R 的定义如下式：

$$R=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}}=\frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{[n\sum X^2-(\sum X)^2][n\sum Y^2-(\sum Y)^2]}}$$

相关系数 R 的取值范围为 $-1\leqslant R\leqslant 1$，R 的取值具有以下不同的含义，

(1) $R=1$→完全正相关；

(2) $R>0$→正相关；

(3) $R=0$→不相关；

(4) $R<0$→负相关；

(5) $R=-1$→完全负相关；

顺便提一下，在相关关系中，有时有因果关系，有时则没有，请读者注意区别。所谓因果关系，指的是原因明确的存在，并且由此产生了结果。但是，即使在没有因果关系的情况下，为了了解相关关系的大小，也需要进行相关分析。

2.5.2 相关系数的检验

计算出来的相关系数在多大程度上是值得信赖的，需要进行检验。计算出来的相关系数如下表 2-1，系数越大两个变量之间越有显著的相关关系。表 2-1 将显著性水平为 10%、5%。1%的相关系数，与不同的自由度(样本数-2=n-2)相对应的显示出来。显著性水平越小，检验越严格，这种选择需要由分析者根据研究内容自己决定。

表 2-1　相关系数检验

自由度 $n-2$	显著性水平 10%	显著性水平 5%	显著性水平 1%
1	0. 988	0. 997	1. 000
2	0. 900	0. 950	0. 990
3	0. 805	0. 878	0. 959
4	0. 729	0. 811	0. 917
5	0. 669	0. 754	0. 874
6	0. 622	0. 707	0. 834
7	0. 582	0. 666	0. 798
8	0. 549	0. 632	0. 765
9	0. 521	0. 602	0. 735
10	0. 497	0. 576	0. 708
20	0. 360	0. 423	0. 537
50	0. 231	0. 273	0. 354
100	0. 164	0. 159	0. 254
200	0. 116	0. 138	0. 181
500	0. 073	0. 088	0. 115

所谓显著性水平指的是很少会发生的概率，这里相当于相关系数为零$R=0$，即相当于不相关的概率。例如，计算出来的相关系数的绝对值，如果大于表 2-1中显著性水平为 1%的相关系数，那就意味着该相关系数为零的概率(即不相关的概率)小于 1%，因此存在着显著的相关。