第5章

统计推断

第4章介绍了概率论的一些基础知识，本章承接之前的话题，来探讨一些统计分析方面的内容。

5.1 随机采样

概率分布是对现实世界中客观规律的高度抽象和数学表达，在统计分析中它们无处不在。但又因为分布是一种抽象的数学表达，所以要设法从观察中找到一个合适的分布并非易事，甚至某些分布很难用常规的、现成的数学模型去描述。而在处理这类问题时，采样就变得非常重要。在统计学中，抽样（或称采样）是一种推论统计方法，它是指从目标总体（population）中抽取一部分个体作为样本（sample），通过观察样本的某些属性，依据所获得的数据对总体的数量特征得出具有一定可靠性的估计判断，从而达到对总体的认识。

在数理统计中，人们往往对有关对象的某一项数量指标感兴趣。为此，考虑开展与这一数量指标相联系的随机试验，并对这一数量指标进行试验或者观察。通常将试验的全部可能的观察值称为总体，并将每一个可能的观察值称为个体。总体中包含的个体数目称为总体的容量。容量有限的称为有限总体，容量无限的则称为无限总体。

总体中的每一个个体是随机试验的一个观察值，它对应于某一随机变量 X 的值。因此，一个总体对应于一个随机变量 X。于是对总体的研究就变成了对一个随机变量 X 的研究，X 的分布函数和数字特征就称为总体的分布函数和数字特征。这里将总体和相应的随机变量统一看待。

在实际中，总体的分布一般是未知的，或者只知道它具有某种形式而其中包含着未知参数。在数理统计中，人们都是通过从总体中抽取一部分个体，然后再根据获得的数据来对总体分布做出推断。被抽出的部分个体称为总体的一个样本。

所谓从总体抽取一个个体，就是对总体随机变量 X 进行一次观察并记录其结果。在相

同的条件下对总体随机变量 X 进行 n 次重复、独立的观察，并将 n 次观察结果按照试验的次序记为 $X_1, X_2, \cdots, X_n$。由于 $X_1, X_2, \cdots, X_n$ 是对随机变量 X 观察的结果，且各次观察是在相同的条件下独立完成的，所以认为 $X_1, X_2, \cdots, X_n$ 是相互独立的，且都是与 X 具有相同分布的随机变量。这样得到的 $X_1, X_2, \cdots, X_n$ 称为来自总体 X 的一个简单随机样本，n 称为这个样本的容量，如无特定说明文中所提到的样本都是指简单随机样本。当 n 次观察一经完成，便得到一组实数 $x_1, x_2, \cdots, x_n$，依次是随机变量 $X_1, X_2, \cdots, X_n$ 的观察值，称为样本值。

设 X 是具有分布函数 F 的随机变量，若 $X_1, X_2, \cdots, X_n$ 是具有同一分布函数 F 的且相互独立的随机变量，则称 $X_1, X_2, \cdots, X_n$ 为从分布函数 F（或总体 F、或总体 X）得到的容量为 n 的简单随机样本，简称样本。它们的观察值 $x_1, x_2, \cdots, x_n$ 称为样本值，又称为 X 的 n 个独立的观察值。也可将样本看成是一个随机向量，写成 $(X_1, X_2, \cdots, X_n)$，此时样本值相应地写成 $(x_1, x_2, \cdots, x_n)$。若 $(x_1, x_2, \cdots, x_n)$ 与 $(y_1, y_2, \cdots, y_n)$ 都是相应于样本 $(X_1, X_2, \cdots, X_n)$ 的样本值，一般来说它们是不相同的。

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，利用这些样本的函数进行统计推断。

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。

因为 $X_1, X_2, \cdots, X_n$ 都是随机变量，而统计量 $g(X_1, X_2, \cdots, X_n)$ 是随机变量的函数，因此统计量是一个随机变量。设 $x_1, x_2, \cdots, x_n$ 是相应于样本 $X_1, X_2, \cdots, X_n$ 的样本值，则称 $g(x_1, x_2, \cdots, x_n)$ 是 $g(X_1, X_2, \cdots, X_n)$ 的观察值。

样本均值和样本方差是两个最常用的统计量。假设 $X_1, X_2, \cdots, X_n$ 是来自总体 X 的一个样本，$x_1, x_2, \cdots, x_n$ 是这一样本的观察值。定义样本均值如下

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

样本方差为

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

标准差（也称均方差）就是方差的算术平方根，即

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

也许有读者会对上面的公式感到困惑，为什么样本方差计算公式里分母为 $n-1$？简单来说，这样做的目的是为了让方差的估计无偏，即无偏估计。无偏估计（unbiased estimator）的意思是指估计量的数学期望等于被估计参数的真实值，否则就是有偏估计（biased estimator）。之所以进行抽样，就是因为现实中总体的获取可能有困难或者代价太高。退而求其次，用样本的一些数量指标来对相应的总体指标做估计。例如，对于总体 X，样本均值就是总体 X 的数学期望的无偏估计，即

$$E(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

那为什么样本方差分母必须要是 $n-1$ 而不是 n 才能使得该估计无偏呢？这是令很多

人倍感困惑的地方。

首先，假定随机变量 X 的数学期望 μ 是已知的，然而方差 σ^2 未知。在这个条件下，根据方差的定义有

$$E[(X_i-\mu)^2]=\sigma^2,\quad \forall i=1,2,\cdots,n$$

由此可得

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]=\sigma^2$$

因此

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

是方差 σ^2 的一个无偏估计，式中的分母 n。这个结果符合直觉，并且在数学上也是显而易见的。

现在，考虑随机变量 X 的数学期望 μ 是未知的情形。这时，人们会倾向于直接用样本均值 $\overline{X}$ 替换掉上面式子中的 μ。这样做有什么后果呢？后果就是如果直接使用

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

作为估计，将会倾向于低估方差。这是因为

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2&=\frac{1}{n}\sum_{i=1}^{n}[(X_i-\mu)+(\mu-\overline{X})]^2\\&=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+\frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-\overline{X})+\frac{1}{n}\sum_{i=1}^{n}(\mu-\overline{X})^2\\&=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+2(\overline{X}-\mu)(\mu-\overline{X})+(\mu-\overline{X})^2\\&=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\overline{X})^2\end{aligned}$$

换言之，除非正好 $\overline{X}=\mu$，否则一定有

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

而不等式右边的才是对方差的无偏估计。这个不等式说明了为什么直接使用

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

会导致对方差的低估。那么，在不知道随机变量真实数学期望的前提下，如何正确的估计方差呢？答案是把上式中的分母 n 换成 $n-1$，通过这种方法把原来偏小的估计“放大”一点点，就能获得对方差的正确估计了，而且这个结论也是可以被证明的。

下面就来证明

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]=\sigma^2$$

记 $D(X_i)$，$E(X_i)$ 为 X_i 的方差和期望，显然有 $D(X_i)=\sigma^2$，$E(X_i)=\mu$。

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

$$E(\overline{X}^2)=D(\overline{X})+E^2(\overline{X})=\frac{\sigma^2}{n}+\mu^2$$

且有

$$E\Big[\sum_{i=1}^{n}X_i^2\Big]=\sum_{i=1}^{n}E[X_i^2]=\sum_{i=1}^{n}[D(X_i)+E^2(X_i)]=n(\sigma^2+\mu^2)$$

$$E\Big[\sum_{i=1}^{n}X_i\overline{X}\Big]=E\Big[\overline{X}\sum_{i=1}^{n}X_i\Big]=nE(\overline{X}^2)=n\Big(\frac{\sigma^2}{n}+\mu^2\Big)$$

由此可得

$$\begin{aligned}E\Big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\Big]&=\frac{1}{n-1}E\Big[\sum_{i=1}^{n}(X_i-\overline{X})^2\Big]\\&=\frac{1}{n-1}E\Big[\sum_{i=1}^{n}(X_i^2-2X_i\overline{X}+\overline{X}^2)\Big]\\&=\frac{1}{n-1}\Big[n(\sigma^2+\mu^2)-2n\Big(\frac{\sigma^2}{n}+\mu^2\Big)+n\Big(\frac{\sigma^2}{n}+\mu^2\Big)\Big]=\sigma^2\end{aligned}$$

结论得证。

既然已经知道样本方差的定义为

$$s^2=\frac{\sum_{i=1}^{n}[X_i-\overline{X}][X_i-\overline{X}]}{n-1}$$

那么也就可以因此给出样本协方差的定义如下

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}[X_i-\overline{X}][Y_i-\overline{Y}]}{n-1}$$

设总体 X(无论服从什么分布,只要均值和方差存在)的均值为 μ,方差为 σ^2,$X_1,X_2,\cdots X_n$是来自总体 X 的一个样本,$\overline{X}$ 和s^2 分别是样本均值和样本方差,则有

$$E(\overline{X})=\mu,\quad D(\overline{X})=\sigma^2/n$$

而

$$\begin{aligned}E(s^2)&=E\Big[\frac{1}{n-1}\sum_{i=1}^{n}(X_i^2-n\overline{X}^2)\Big]=\frac{1}{n-1}\sum_{i=1}^{n}[E(X_i^2)-nE(\overline{X}^2)]\\&=\frac{1}{n-1}\sum_{i=1}^{n}\Big[(\sigma^2+\mu^2)-n\Big(\frac{\sigma^2}{n}+\mu^2\Big)\Big]=\sigma^2\end{aligned}$$

即

$$E(s^2)=\sigma^2$$

回忆第 4 章中曾经给出的一个结论:设 $X_1,X_2,\cdots,X_n$是来自正态总体 $N(\mu,\sigma^2)$的一个样本,$\overline{X}$ 是样本的均值,则有

$$\overline{X}\sim N\Big(\mu,\frac{\sigma^2}{n}\Big)$$

如果将其转换为标准正态分布的形式,则得出

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

很多情况下，无法得知总体方差 σ^2，此时就需要使用样本方差 s^2 替代。但这样做的结果就是，上式将发生些许变化。最终的形式由下面这个定理给出，这也是本章后面将多次用到的一个重要结论。

定理 设 $X_1, X_2, \cdots X_n$ 是来自正态总体 $N(\mu, \sigma^2)$ 的一个样本，样本均值和样本方差分别是 $\overline{X}$ 和 s^2，则有

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

其中，$t(n-1)$ 表示自由度为 $n-1$ 的 t 分布。当 n 足够大时，t 分布近似于标准正态分布（此时即变成中央极限定理所描述的情况）。当对于较小的 n 而言，t 分布与标准正态分布有较大差别。

学生 t 分布，简称 t 分布，是类似正态分布的一种对称分布，但它通常要比正态分布平坦和分散。一个特定的 t 分布依赖于称之为自由度的参数，自由度越小，那么 t 分布的图形就越平坦，随着自由度的增大，t 分布也逐渐趋近于正态分布。图 5-1 为标准正态分布及两个自由度不同的 t 分布。

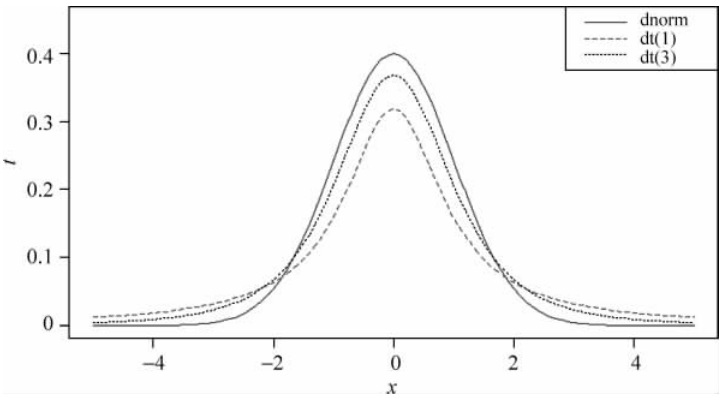

图 5-1 标准正态分布与 t 分布

这里谈到的 t 分布最初是由英国化学家和统计学家威廉·戈塞特（Willam Gosset）于 1908 年首先提出的，当时他还在爱尔兰都柏林的一家酿酒厂工作。酒厂虽然禁止员工发表一切与酿酒研究有关的成果，但还是允许他在不提到酿酒的前提下，以笔名发表 t 分布的发现，所以论文使用了“学生”（Student）这一笔名。后来，t 检验方法以及相关理论经由费希尔发扬光大，为了感谢戈塞特的功劳，费希尔将此分布命名为学生 t 分布（Student's t-distribution）。

5.2 参数估计

统计推断是以带有随机性的样本观测数据为基础，结合具体的问题条件和假定，而对未知事物做出的以概率形式表述的推断，它是数理统计的主要任务。总的来说，统计推断的基本问题可以分为两大类：一类是参数估计；另一类是假设检验。在参数估计部分，将着重关注点估计和区间估计这两类问题。

5.2.1　参数估计的基本原理

如果想知道某所中学高三年级全体男生的平均身高，其实只要测定每个人的身高然后再取均值即可。但是若想知道中国成年男性的平均身高似乎就不那么简单了，因为这个研究的对象群体过于庞大，要想获得全体中国成年男性的身高数据显然不切实际。这时一种可以想到的办法就是对这个庞大的总体进行采样，然后根据样本参数来推断总体参数，于是便引出了参数估计(parameter estimation)的概念。参数估计就是用样本统计量去估计总体参数的方法。例如，可以用样本均值估计总体均值，用样本方差估计总体方差。如果把总体参数(均值、方差等)笼统地用一个符号 θ 表示，而用于估计总体参数的统计量用 $\hat{\theta}$ 表示，那么参数估计也就是用 $\hat{\theta}$ 估计 θ 的过程，其中 $\hat{\theta}$ 也称为是估计量(estimator)，而根据具体样本计算得出的估计量数值就是估计值(estimated value)。

点估计(point estimate)就是用样本统计量 $\hat{\theta}$ 的某个取值直接作为总体参数 θ 的估计值。例如，可以用样本均值 $\bar{x}$ 直接作为总体均值 μ 的估计值，用样本比例 p 直接作为总体比例的估计值等。这种方式的点估计也称为矩估计，它的基本思路就是用样本矩估计总体矩，用样本矩的相应函数来估计总体矩的函数。由大数定理可知，如果总体 X 的 k 阶矩存在，那么样本的 k 阶矩以概率收敛到总体的 k 阶矩，样本矩的连续函数收敛到总体矩的连续函数，这就启发人们可以用样本矩作为总体矩的估计量，这种用相应的样本矩去估计总体矩的估计方法就称为矩估计法，这种方法最初是由英国统计学家卡尔·皮尔逊(Karl Pearson)提出的。

来看一个例子。2014 年 10 月 28 日，为了纪念美国实验医学家、病毒学家乔纳斯·爱德华·索尔克(Jonas Edward Salk)百年诞辰，谷歌特别在其主页上刊出了一幅如图 5-2 所示的纪念画。“二战”以后，由于缺乏有效的防控手段，脊髓灰质炎逐渐成为美国公共健康的最大威胁之一。1952 年的“大流行”是美国历史上最严重的爆发，那年报道的病例有 58 000 人，其中 3145 人死亡，另有 21 269 人致残，且多数受害者是儿童。直到索尔克研制出首例安全有效的“脊髓灰质炎疫苗”，曾经让人闻之色变的脊髓灰质炎才开始得到有效的控制。

图 5-2　索尔克纪念画

索尔克在验证他发明的疫苗效果时，设计了一个随机双盲对照试验，实验结果是在 200 745 名全部接种了疫苗的儿童中，最后患上脊髓灰质炎的一共有 57 例。那么采用点估计的办法就可以推断该疫苗的整体失效率大约为

$$\hat{p}=\frac{57}{200\,745}=0.0284\%$$

在重复抽样下，点估计的均值可以期望等于总体的均值，但由于样本是随机抽取的，由某一个具体样本算出的估计值可能并不等同于总体均值。在用矩估计法对总体参数进行估计时，还应该给出点估计值与总体参数真实值间的接近程度。通常围绕点估计值构造总体参数的一个区间，并用这个区间度量真实值与估计值之间的接近程度，这就是区间估计。

区间估计(interval estimate)是在点估计的基础上，给出总体参数估计的一个区间范围，而这个区间通常是由样本统计量加减估计误差得到的。与点估计不同，进行区间估计时，根据样本统计量的抽样分布可以对样本统计量与总体参数的接近程度给出一个概率度量。

例如，在以样本均值估计总体均值的过程中，由样本均值的抽样分布可知，在重复抽样或无限总体抽样的情况下，样本均值的数学期望等于总体均值，即 $E(\bar{x})=\mu$。还可以知道，样本均值的标准差 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$，其中 σ 是总体的标准差，n 是样本容量。根据中央极限定理可知样本均值的分布服从正态分布。这就意味着，样本均值 $\bar{x}$ 落在总体均值 μ 的两侧各一个抽样标准差范围内的概率为 0.6827；落在两个抽样标准差范围内的概率为 0.9545；落在三个抽样标准差范围内的概率是 0.9973。

事实上，完全可以求出样本均值落在总体均值两侧任何一个抽样标准差范围内的概率。但实际估计时，情况却恰恰相反。人们所知的仅是样本均值 $\bar{x}$，而总体均值 μ 未知，也正是需要估计的。由于 $\bar{x}$ 与 μ 之间的距离是对称的，如果某个样本均值落在 μ 的两个标准差范围之内，反过来 μ 也就被包括在以 $\bar{x}$ 为中心左右两个标准差的范围之内。因此，约有 95% 的样本均值会落在 μ 的两个标准差范围内。或者说，约有 95% 的样本均值所构造的两个标准差区间会包括 μ。图 5-3 给出了区间估计的示意图。

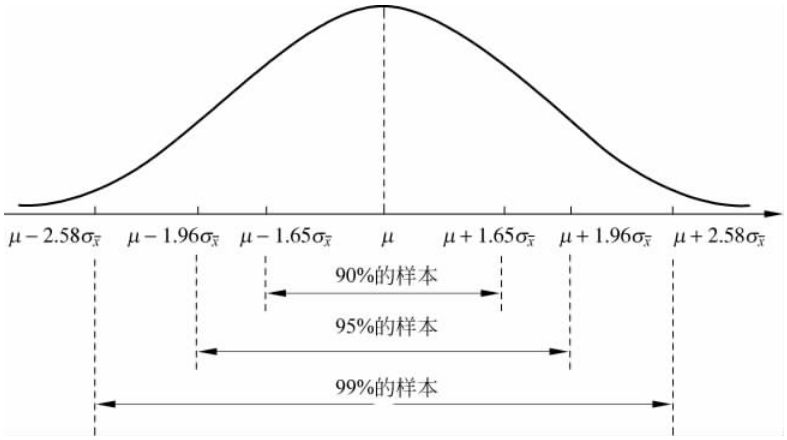

图 5-3　区间估计示意图

在区间估计中，由样本统计量所构造的总体参数的估计区间被称为**置信区间**(confidence interval)，而且如果将构造置信区间的步骤重复多次，置信区间中所包含的总体参数真实值的次数的占比称为置信水平，或置信度。在构造置信区间时，可以使用希望的任

意值作为置信水平。常用的置信水平和正态分布曲线下右侧面积为 $\alpha/2$ 时的临界值如表 5-1 所示。

表 5-1 常用置信水平临界值

置信水平	α	$\alpha/2$	临界值
90%	0.10	0.050	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

5.2.2 单总体参数区间估计

1. 总体比例的区间估计

比例问题可以看做是一项满足二项分布的试验。例如，在索尔克的随机双盲对照试验中，实验结果是在全部 200 745 名接种了疫苗的儿童中最后患上脊髓灰质炎的一共有 57 例。这就相当是做了 200 745 次独立的伯努利试验，而且每次试验的结果必为两种可能之一，即要么是患病，要么是不患病。本章前面也讲过，服从二项分布的随机变量 $X\sim B(n,p)$ 以 np 为期望，以 $np(1-p)$ 为方差。可以令样本比例 $\hat{p}=X/n$ 作为总体比例 p 的估计值，而且可以得知

$$E(\hat{p})=\frac{1}{n}E(x)=\frac{1}{n}\cdot np=p$$

同时还有

$$\mathrm{var}(\hat{p})=\frac{1}{n^2}\mathrm{var}(x)=\frac{1}{n^2}\cdot np(1-p)=\frac{p(1-P)}{n}$$

$$\mathrm{se}(\hat{p})=\sqrt{\frac{p(1-P)}{n}}$$

由此便已经具备了进行区间估计的必要素材。

第一种进行区间估计的方法被称为是 Wald 方法，它是一种近似方法。根据中央极限定理，当 n 足够大时，将会有

$$\hat{p}\sim N\left(p,\sqrt{\frac{p(1-P)}{n}}\right)$$

5.2.1 节中也给出了标准正态分布中 95%置信水平下的临界值，即 1.96，则

$$\Pr\left(-1.96<\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}<1.96\right)\approx 0.95$$

$$\Pr\left(\hat{p}-1.96\sqrt{\frac{p(1-p)}{n}}<p<\hat{p}+1.96\sqrt{\frac{p(1-p)}{n}}\right)\approx 0.95$$

Wald 方法对上述结果做了进一步的近似，即把根号下的 p 用 $\hat{p}$ 代替，于是总体比例 p 在 95%置信水平下的置信区间即为

$$\left(\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

以索尔克的随机双盲对照试验为例，可以算得总体比例估计的置信区间，保留小数点后

6 位有效数字的结果为(0.000 210,0.000 358)。

Wald 方法的基本原理是利用正态分布对二项分布进行近似，与之相对的另外一种方法是 Clopper-Pearson 方法。该方法完全是基于二项分布的，所以它是一种更加确切的区间估计方法。利用 Clopper-Pearson 方法，可以算得保留小数点后 6 位有效数字的 95%置信水平下的区间估计结果为(0.000 215,0.000 369)。可见，这一数值其实已经与 Wald 方法所得之结果非常相近了。

2. 总体均值的区间估计

在对总体均值进行区间估计时，需要分几种情况。首先，如果考虑的总体是正态分布且方差 σ^2 已知，或总体不满足正态分布但为大样本($n\geqslant 30$)时，样本均值$\bar{x}$的抽样分布均为正态分布，数学期望为总体均值 μ，方差为 σ^2/n。而样本均值经过标准化以后的随机变量服从标准正态分布，即

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

由此可知总体均值 μ 在 $1-\alpha$ 置信水平下的置信区间为

$$\left(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

其中，α 为显著水平，它是总体均值不包含在置信区间内的概率；$z_{\alpha/2}$为标准正态分布曲线与横轴围成的面积等于 $\alpha/2$ 时的 z 值。

如果总体服从正态分布但 σ^2 未知，或总体并不服从正态分布，只要是在大样本条件下，都可以用样本方差s^2 来代替总体方差 σ^2，此时总体均值在 $1-\alpha$ 置信水平下的置信区间为

$$\left(\bar{x}-z_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

其中需要注意的一点，也是本章前面着重讨论的一点，即如果设 $X_1,X_2,\cdots,X_n$是来自总体 X 的一个样本，那么作为总体方差 σ^2 的无偏估计的样本方差公式为

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2=\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-n\bar{X}^2$$

除此之外，考虑总体是正态分布，但方差 σ^2 未知且属于小样本($n<30$)的情况，仍需用样本方差s^2 替代总体方差 σ^2。但此时样本均值经过标准化以后的随机变量将服从自由度为($n-1$)的 t 分布，即

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

注意这也是本章前面给出的一个定理。于是，需要采用学生 t 分布建立总体均值 μ 的置信区间。根据 t 分布建立的总体均值 μ 在 $1-\alpha$ 置信水平下的置信区间为

$$\left(\bar{x}-t_{\alpha/2}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

其中，$t_{\alpha/2}$是自由度为 $n-1$ 时，t 分布中右侧面积为 $\alpha/2$ 的 t 值。

表 5-2 对本部分介绍的关于单总体均值的区间估计方法进行了总结，供有需要的读者参阅。

表 5-2　单总体均值的区间估计

总体分布	样本量	总体方差 σ^2 已知	总体方差 σ^2 未知
正态分布	大样本（$n\geqslant 30$）	$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
	小样本（$n<30$）	$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}$
非正态分布	大样本（$n\geqslant 30$）	$\bar{x}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm z_{\alpha/2}\frac{s}{\sqrt{n}}$

3. 总体方差的区间估计

此处仅讨论正态总体方差的估计问题。根据样本方差的抽样分布可知，样本方差服从自由度为 $n-1$ 的 χ^2 分布，所以考虑用 χ^2 分布构造总体方差的置信区间。给定一个显著水平 α，用 χ^2 分布建立总体方差 σ^2 的置信区间，其实就是要找到一个 χ^2 值，使得

$$\chi^2_{1-\alpha/2}\leqslant\chi^2\leqslant\chi^2_{\alpha/2}$$

由于

$$\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$$

所以可以用其来替代 χ^2，于是有

$$\chi^2_{1-\alpha/2}\leqslant\frac{(n-1)s^2}{\sigma^2}\leqslant\chi^2_{\alpha/2}$$

并根据上式推导出总体方差 σ^2 在 $1-\alpha$ 置信水平下的置信区间为

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}}\leqslant\sigma^2\leqslant\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

因此便可对总体方差的置信区间进行估计。

5.2.3　双总体均值差的估计

本章前面曾经指出，若 $X_i\sim N(\mu_i,\sigma_i^2)$，其中 $i=1,2,\cdots,n$ 且相互独立，则它们的线性组合为 $C_1X_1+C_2X_2+\cdots+C_nX_n$，仍服从正态分布，其中 $C_1,C_2,\cdots,C_n$ 是不全为 0 的常数，并由数学期望和方差的性质可知

$$C_1X_1+C_2X_2+\cdots+C_nX_n\sim N\left(\sum_{i=1}^{n}C_i\mu_i,\sum_{i=1}^{n}C_i^2\sigma_i^2\right)$$

所以假设随机变量的估计符合正态分布的一个好处就是它们的线性组合仍然可以满足正态分布的假设。如果有 $X_1\sim N(\mu_1,\sigma_1^2)$ 和 $X_2\sim N(\mu_2,\sigma_2^2)$，显然有

$$aX_1+bX_2\sim N(a\mu_1+b\mu_2,\sqrt{a^2\sigma_1^2+b^2\sigma_2^2})$$

当 $a=1,b=-1$ 时，有

$$X_1-X_2\sim N(\mu_1-\mu_2,\sqrt{\sigma_1^2+\sigma_2^2})$$

这其实给出了两个独立的正态分布的总体之差的分布。

从 X_1 和 X_2 这两个总体中分别抽取样本量为 n_1 和 n_2 的两个随机样本，样本均值分别

为$\bar{x}_1$和$\bar{x}_2$，则样本均值$\bar{x}_1$满足$\bar{x}_1 \sim (\mu_1, \sigma_1^2/n_1)$，样本均值$\bar{x}_2$满足$\bar{x}_2 \sim (\mu_2, \sigma_2^2/n_2)$。进而样本均值之差$\bar{x}_1 - \bar{x}_2$满足

$$(\bar{x}_1 - \bar{x}_2) \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

由此得到了进行双总体均值的差区间估计的所需素材。在具体讨论时将问题分成两类，即独立样本数据的双总体均值差估计问题，以及配对样本数据的双总体均值差估计问题。

1. 独立样本

如果两个样本是从两个总体中独立抽取的，即一个样本中的元素与另一个样本中的元素相互独立，则称为独立样本(independent samples)。

当两个总体的方差σ_1^2和σ_2^2已知的时候，根据前面推出的结论，类似于单个总体区间估计，可以得出$\mu_1 - \mu_2$的置信水平为$1-\alpha$的双尾置信区间为

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

如果两个总体的方差未知，可以用两个样本方差s_1^2和s_2^2代替，这时$\mu_1 - \mu_2$的置信水平为$1-\alpha$的双尾置信区间为

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

对于两个总体的方差未知的情况，将进一步划分为两种情况，首先当两个总体方差相同，即$\sigma_1^2 = \sigma_2^2$但未知时，可以得到

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

其中

$$s' = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

其中，s_1^2和s_2^2分别是样本方差。类似之前的做法，可以得到$\mu_1 - \mu_2$的置信水平为$1-\alpha$的双尾置信区间为

$$\left(\bar{x}_1 - \bar{x}_2 - t_{\alpha/2}(n_1 + n_2 - 2)s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2}(n_1 + n_2 - 2)s'\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

看一个例子。假设有编号为1和2的两种饲料，现在分别用它们喂养两组肉鸡，然后记录每只鸡的增重情况，数据如表5-3所示。

表5-3 喂食不同饲料的肉鸡增重情况

饲料	增重
1	42，68，85
2	42，97，81，95，61，103

首先分别计算两组数据的均值和方差，均值分别为65和79.83，方差分别为21.66和23.87。两组样本观察值的标准差是非常相近的，因此假设两个总体的方差是相等的。

根据上面给出的公式，首先来计算s'的值，计算过程如下

$$s' = \sqrt{\frac{2\times 21.66^2 + 5\times 23.87^2}{3+6-2}} = 23.26$$

因此，$\mu_1 - \mu_2$ 在 95% 置信水平下的置信区间为

$$65 - 79.83 \pm c_{0.975}(t_7)\times 23.26\sqrt{\frac{1}{6}+\frac{1}{3}}$$

$$= -14.83 \pm 38.90 = (-53.72, 24.06)$$

此外，当两个总体的方差未知，且 $\sigma_1^2 \neq \sigma_2^2$ 时，可以证明

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t(\nu)$$

近似成立，其中

$$\nu = \left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^2 \Big/ \left[\frac{(\sigma_1^2)^2}{n_1^2(n_1-1)}+\frac{(\sigma_2^2)^2}{n_2^2(n_2-2)}\right]$$

但由于 σ_1^2 和 σ_2^2 未知，所以用样本方差 s_1^2 和 s_2^2 近似，即

$$\hat{\nu} = \left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2 \Big/ \left[\frac{(s_1^2)^2}{n_1^2(n_1-1)}+\frac{(s_2^2)^2}{n_2^2(n_2-2)}\right]$$

可以近似地认为 $t \sim t(\hat{\nu})$。并由此得到 $\mu_1 - \mu_2$ 的置信水平为 $1-\alpha$ 的双尾置信区间为

$$\left(\bar{x}_1 - \bar{x}_2 - t_{\alpha/2}(\hat{\nu})\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{\alpha/2}(\hat{\nu})\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)$$

仍以饲料和肉鸡增重的数据为例，可以得到

$$\frac{s_1^2}{n_1} = \frac{21.66^2}{3} \approx 156.3852, \quad \frac{s_2^2}{n_2} = \frac{23.87^2}{6} \approx 94.9628$$

进而有

$$\hat{\nu} = \frac{(156.3852+94.9628)^2}{(156.3852^2/2)+(94.9628^2/5)} \approx 4.503$$

因此，$\mu_1 - \mu_2$ 在 95% 置信水平下的置信区间为

$$65 - 79.83 \pm c_{0.975}(t_{4.503})\times\sqrt{\frac{23.87^2}{6}+\frac{21.66^2}{3}}$$

$$= -14.83 \pm 2.6585\times 15.85 = (-56.97, 27.30)$$

2. 配对样本

在前面的例子中，为了讨论两种饲料的差异，从两个独立的总体中进行了抽样，但使用独立样本估计两个总体均值之差也潜藏着一些弊端。试想一下，如果喂食饲料 1 的肉鸡和喂食饲料 2 的肉鸡体质上本来就存在差异，可能其中一种吸收更好而另一组则略差，显然试验结果的说服力将大打折扣。这种“有失公平”的独立抽样往往会掩盖一些真正的差异。

在实验设计中，为了控制其他”有失公平”的因素，尽量降低不利影响，使用配对样本(paired sample)就是一种值得推荐的做法。所谓配对样本就是指一个样本中的数据与另一个样本中的数据是相互对应的。例如，在验证饲料差异的试验中，可以选用同一窝诞下的一对小鸡作为一个配对组，因为人们认为同一窝诞下的小鸡之间差异最小。按照这种思路，如表 5-4 所示，一共有 6 个配对组参与实验，然后从每组中随机选取一只小鸡喂食饲料 1，然后

向另外一只喂食饲料 2，并记录肉鸡体重增加的数据。

表 5-4 配对试验数据

饲料	配对 1 组	配对 2 组	配对 3 组	配对 4 组	配对 5 组	配对 6 组
1	44	55	68	85	90	97
2	42	61	81	95	97	103

使用配对样本进行估计时，在大样本条件下，两个总体均值之差 $\mu_1-\mu_2$ 在 $1-\alpha$ 置信水平下的置信区间为

$$\left(\bar{d}-z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}},\bar{d}+z_{\alpha/2}\frac{\sigma_d}{\sqrt{n}}\right)$$

其中，d 表示一组配对样本之间的差值，$\bar{d}$ 表示各差值的均值，σ_d 表示各差值的标准差。当总体 σ_d 未知时，可用样本差值的标准差 s_d 来代替。

在小样本情况下，假定两个总体观察值的配对差值服从正态分布。那么两个总体均值之差 $\mu_1-\mu_2$ 在 $1-\alpha$ 置信水平下的置信区间为

$$\left(\bar{d}-t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}},\bar{d}+t_{\alpha/2}(n-1)\frac{s_d}{\sqrt{n}}\right)$$

例如，根据表 5-4 中的数据可以算得各配对组之差分别为 -2、6、13、10、7 和 6，以及 $\bar{d}=6.667$，$s_d=5.046$。因此，总体均值之差 $\mu_1-\mu_2$ 在 95%置信水平下的置信区间为

$$6.667\pm c_{0.975}(t_5)\times\frac{5.046}{\sqrt{6}}\approx(1.37,11.96)$$

5.2.4 双总体比例差的估计

由样本比例的抽样分布可知，从两个满足二项分布的总体中抽出两个独立的样本，那么两个样本比例之差的抽样服从正态分布，即

$$(\hat{p}_1-\hat{p}_2)\sim N\left(p_1-p_2,\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}\right)$$

再对两个样本比例之差进行标准化，即

$$z=\frac{(\hat{p}_1-\hat{p}_2)-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}}\sim N(0,1)$$

当两个总体的比例 p_1 和 p_2 未知时，可用样本比例 $\hat{p}_1$ 和 $\hat{p}_2$ 代替。所以，根据正态分布建立的两个总体比例之差 p_1-p_2 在 $1-\alpha$ 置信水平下的置信区间为

$$(\hat{p}_1-\hat{p}_2)\pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

下面来看一个例子。在某电视节目的收视率调查中，从农村随机调查了 400 人，其中有 128 人表示收看了该节目；从城市随机调查了 500 人，其中 225 人表示收看了该节目。请以 95%的置信水平来估计城市与农村收视率差距的置信区间。利用上述公式，不难算出置信

区间为(6.68%,19.32%),即城市与农村收视率差值的95%的置信区间为6.68%～19.32%。如果使用连续性修正,为6.46%～19.54%。

5.3　假设检验

假设检验是除参数估计之外的另一类重要的统计推断问题。它的基本思想可以用小概率原理来解释。所谓小概率原理,就是认为小概率事件在一次试验中是几乎不可能发生的。也就是说,对总体的某个假设是真实的,那么不利于或者不能支持这一假设的事件在一次试验中是几乎不可能发生的;要是在一次试验中该事件竟然发生了,人们就有理由怀疑这一假设的真实性,进而拒绝这一假设。

5.3.1　基本概念

大卫·萨尔斯伯格(David Salsburg)在《女士品茶:20世纪统计怎样变革了科学》一书中,以英国剑桥一群科学家及其夫人们在一个慵懒的午后所做的一个小小的实验为开篇,为读者展开了一个关于20世纪统计革命的别样世界。而开篇这个品茶故事大约是这样的,当时一位女士表示向一杯茶中加入牛奶和向一杯奶中加入茶水,两者的味道品尝起来是不同的。她的这一表述立刻引起了当时在场的众多睿智头脑的争论。其中一位科学家决定用科学的方法来测试一下这位女士的假设。这个人就是大名鼎鼎的英国统计学家,现代统计科学的奠基人罗纳德·费希尔(Ronald Fisher)。费希尔给这位女士提供了8杯兑了牛奶的茶,其中一些是先放的牛奶,另一些则是先放的茶水,然后费希尔让这位女士品尝后判断每一杯茶的情况。

现在问题来了,这位女士能够成功猜对多少杯茶的情况才足以证明她的理论是正确的,8杯?7杯?还是6杯?解决该问题的一个有效方法是计算一个P值,然后由此推断假设是否成立。P值(P-value)就是当原假设为真时所得到的样本观察结果或更极端结果出现的概率。如果P值很小,说明原假设情况的发生的概率很小,而如果确实出现了P值很小的情况,根据小概率原理,人们就有理由拒绝原假设。P值越小,拒绝原假设的理由就越充分。就好比说种瓜得瓜,种豆得豆。在原假设"种下去的是瓜"这个条件下,正常得出来的也应该是瓜。相反,如果得出来的是瓜这件事越不可能发生,人们否定原假设的把握就越大。如果得出来的是豆,也就表明得出来的是瓜这件事的可能性小到了零,这时就有足够的理由推翻原假设,也就可以确定种下去的根本就不是瓜。

假定总共的8杯兑了牛奶的茶中,有6杯的情况都被猜中了。现在就来计算一下这个P值。不过在此之前,还需要先建立原假设和备择假设。原假设通常是指那些单纯由随机因素导致的采样观察结果,通常用H_0表示。而备择假设,则是指受某些非随机原因影响而得到的采样观察结果,通常用H_1表示。如果从假设检验具体操作的角度来说,常常把一个被检验的假设称为原假设,当原假设被拒绝时而接收的假设称为备择假设,原假设和备择假设往往成对出现。此外,原假设往往是研究者想收集证据予以反对的假设,当然也是有把握且不能轻易被否定的命题,而备择假设则是研究者想收集证据予以支持的假设,同时也是无

把握且不能轻易肯定的命题作。

就当前所讨论的饮茶问题而言，显然在不受非随机因素影响的情况下，那个常识性的，似乎很难被否定的命题应该是“无论是先放茶水还是先放牛奶是没有区别的”。如果将该命题作为 H_0，其实也就等同于那位女士对茶的判断完全是随机的，因此她猜中的概率应该是0.5。这时随机变量 $X\sim B(8,0.5)$，即满足 $n=8$，$p=0.5$ 的二项分布。相应的备择假设 H_1 为该女士能够以大于 0.5 的概率猜对茶的情况。

直观上，如果 8 杯兑了牛奶的茶中，有 6 杯的情况都被猜中了，可以算出 $\hat{p}=6/8=0.75$，这个值大于 0.5，但这是否大到可以令人们相信先放茶水还是先放牛奶确有不同这个结论。所以需要来计算一下 P 值，即 $Pr(X\geqslant 6)$。可以算得 P 值是 0.144 531 2。可见，P 值并不是很显著。通常都需要 P 值小于 0.05，才能有足够的把握拒绝原假设。而本题所得结果则表明没有足够的证据支持拒绝原假设。所以如果那位女士猜对了 8 杯中的 6 杯，也没有足够的证据表明先加牛奶或者先加茶水会有何不同。

还应该注意到以上所讨论的是一个单尾的问题。因为备择假设是说该女士能够以大于0.5 的概率猜对茶的情况。日常遇到的很多问题也有可能是双尾的，例如原假设是概率等于某个值，而备择假设则是不等于该值，即大于或者小于该值。在这种情况下，通常需要将算得的 P 值翻倍，除非已经求得的 P 值大于 0.5，此时令 P 值为 1。另外，当 n 较大的时候，还可以用正态分布来近似二项分布。

1965 年，美国联邦最高法院对斯文诉亚拉巴马州一案作出了裁定。该案也是法学界在研究预断排除原则时常常被提及的著名案例。本案的主角斯文是一个非洲裔美国人，他被控于亚拉巴马州的塔拉迪加地区对一名白人妇女实施了强奸犯罪，并因此被判处死刑。

最终该项案件被上诉至最高法院，理由是陪审团中没有黑人成员，斯文据此认为自己受到了不公正的审判。最高法院驳回了上述请求。根据亚拉巴马州法律，陪审团成员是从一个 100 人的名单中抽选的，而当时的 100 名备选成员中有 8 名是黑人。根据诉讼过程中的无因回避原则，这 8 名黑人被排除在了此处审判的陪审团之外，而无因回避原则本身是受宪法保护的。最高法院在裁决书中也指出：“无因回避的功能不仅在于消除双方的极端不公正，也要确保陪审员仅仅依赖于呈现在他们面前的证据做出裁决，而不能依赖于其他因素……无因回避可允许辩护方通过预先审核程序中的调查提问以确定偏见的可能，消除陪审员的敌意。”此外最高法院还认为，在陪审团备选名单上有 8 名黑人成员，表明整体比例上的差异很小，所以也就不存在刻意引入或者排除一定数量的黑人成员的意图。

亚拉巴马州当时规定只要超过 21 岁就符合陪审团成员的资格，而在塔拉迪加地区满足这个条件的大约有 16 000 人，其中 26％是非洲裔美国人。现在的问题是，如果这 100 名备选的陪审团成员确实是从符合条件的人群中随机选取的，那么其中黑人成员的数量会否是 8 人或者更少？可以算得这个概率是 0.000 004 7，也就相当于二十万分之一的机会。

对于假设检验而言，也可以使用正态分布的近似参数计算置信区间。唯一的不同在于此时是在原假设 $H_0:p=p_0$ 的前提下计算概率值，所以原来在计算置信区间时所采用的近似

$$\frac{p(1-p)}{n}\approx\frac{\hat{p}(1-\hat{p})}{n}$$

现在就不再需要了。取而代之的是在计算标准误差和 P 值时直接使用 p_0 即可。

如果估计值用 $\hat{p}$ 表示，其(估计的)标准误差是

$$\sqrt{p_0(1-p_0)/n}$$

检验统计量为

$$Z=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$$

是当 n 比较大时，在原假设前提下，通过对标准正态分布的近似得到。

继续前面的例子，现在原假设可以表述为 $H_0: p=0.26$，相对应的备择假设为 $H_1: p<0.26$。在 100 人的备选陪审团名单中有 8 名黑人成员，此时 P 值可由下式给出

$$\Pr\left(Z\leqslant\frac{0.08-0.26}{\sqrt{0.26\times0.74/100}}\right)=\Pr(Z\leqslant-4.104)=0.000\,020$$

由此便可以拒绝原假设，从而认为法院的裁定在很大程度上是错误的。

需要说明的是，当使用正态分布(连续的)作为二项分布(离散的)的近似时，要对二项分布中的离散整数 x 进行连续性修正，将数值 x 用从 $x-0.5$ 到 $x+0.5$ 的区间代替(即加上与减去 0.5)。就本题而言，为了得到一个更好的近似，连续性修正就是令 $\Pr(X\leqslant8)\approx\Pr(X^*<8.5)$。所以有

$$\Pr\left(Z\leqslant\frac{0.085-0.26}{\sqrt{0.26\times0.74/100}}\right)=\Pr(Z\leqslant-3.989\,657)=0.000\,033$$

此处无须对连续性修正做过多的解释，但请记住，若不使用连续性修正，那么所得的 P 值将总是偏小，相应的置信区间也偏窄。

5.3.2　两类错误

对原假设提出的命题，要根据样本数据提供的信息进行判断，并得出“原假设正确”或者“原假设错误”的结论。而这个判断有可能正确，也有可能错误。前面在假设检验的基本思想中已经指出，假设检验所依据的基本原理是小概率原理，由此原理对原假设做出判断，而在整个推理过程中所运用的是一种反证法的思路。由于小概率事件，无论其概率多么小，仍然还是有可能发生的，所以利用前面方法进行假设检验时，有可能作出错误的判断。这种错误的判断有两种情形。

一方面，当原假设 H_0 成立时，由于样本的随机性，结果拒绝了 H_0，犯了“弃真”错误，又称为第一类错误，也就是当应该接受原假设 H_0 而拒绝这个假设时，称为犯了第一类错误。当小概率事件确实发生时，就会导致拒绝 H_0 而犯第一类错误，因此犯第一类错误的概率为 α，即假设检验的显著性水平。

另一方面，当原假设 H_0 不成立时，因样本的随机性，结果接受了 H_0，便犯了“存伪”错误，又称为第二类错误，即当应该拒绝原假设 H_0 而接受了这个假设时，称为犯了第二类错误. 犯第二类错误的概率为 β。

当原假设 H_0 为真，人们却将其拒绝，如果犯这种错误的概率用 α 表示，那么当 H_0 为真时，人们没有拒绝它，就表示做出来正确的决策，其概率显然就应该是 $1-\alpha$；当原假设 H_0 为假，人们却没有拒绝它，犯这种错误的概率用 β 表示。那么当 H_0 为假，且正确地拒绝了它，

其概率自然为 $1-\beta$。正确决策和错误决策的概率可以归纳为表 5-5。

表 5-5 假设检验中各种可能结果及其概率

	接受 H_0	拒绝 H_0
H_0 为真	决策正确($1-\alpha$)	弃真错误(α)
H_1 为真	取伪错误(β)	决策正确($1-\beta$)

人们总是希望两类错误发生的概率 α 和 β 都越小越好，然而实际上却很难做到。当样本容量 n 确定后，如果 α 变小，则检验的拒绝域变小，相应的接受域就会变大，因此 β 值也就随之变大；相反，若 β 变小，则不难想到 α 又会变大。人们有时不得不在两类错误之间做权衡。通常来说，哪一类错误所带来的后果更严重、危害更大，在假设检验中就应该把哪一类错误作为首选的控制目标。但实际检验时，通常所遵循的原则都是控制犯第一类错误的概率 α，而不考虑犯第二类错误的概率 β，这样的检验称为显著性检验。这里所讨论的检验，都是显著性检验。又由于显著性水平 α 是预先给定的，因而犯第一类错误的概率是可以控制的，而犯第二类错误的概率通常是不可控的。

5.3.3 均值检验

根据假设检验的不同内容和进行检验的不同条件，需要采用不同的检验统计量，其中 z 统计量和 t 统计量是两个最主要也最常用的统计量，它们常常用于均值和比例的假设检验。具体选择哪个统计量往往要考虑样本量的大小以及总体标准差 σ 是否已知。事实上，因为统计实验往往是针对来自某一总体的一组样本而进行的，所以更多情况下，人们都认为总体标准差 σ 是未知的。在参数估计部分，已经学习了对单总体样本的均值估计以及双总体样本的均值差估计，本节的内容大致上都是基于前面这些已经得到的结果而进行的。

样本量大小是决定选择哪种统计量的一个重要考虑因素。因为大样本条件下，如果总体是正态分布，样本统计量将也服从正态分布，即使总体是非正态的，样本统计量也趋近于正态分布。所以，大样本下的统计量将都被看成是正态分布的，此时即需要使用 z 统计量。z 统计量是以标准正态分布为基础的一种统计量，当总体标准差 σ 已知时，它的计算公式如下

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

正如前面刚刚说过的，实际中总体标准差 σ 往往很难获取，这时一般用样本标准差 s 来代替，如此一来上式便可改写为

$$z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

在样本量较小的情况下，且总体标准差未知，由于检验所依赖的信息量不足，只能用样本标准差来代替总体标准差，此时样本统计量就服从 t 分布，故应使用 t 统计量，其计算公式为

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

这里 t 统计量的自由度为 $n-1$。

例如现在为了测定一块土地的 pH，随机抽取了 17 块土壤样本，相应的 pH 检测结果如表 5-6 所示。现在想问该区域的土壤是否是中性的(即 pH=7)?

表 5-6　土壤 pH 检测数据

6.0	5.7	6.2	6.3	6.5	6.4
6.9	6.6	6.8	6.7	6.8	7.1
6.8	7.1	7.1	7.5	7.0	

首先提出原假设和备择假设如下

$$H_0:\text{pH}=7, H_1:\text{pH}\neq 7$$

该题目显然属于小样本且总体方差未知的情况，此时可以计算其 t 统计量如下

$$t=\frac{6.676\,47-7}{0.454\,88/\sqrt{17}}\approx -2.9326$$

因为这是一个双尾检验，所以计算出其 P 值为 0.009 757 353。

下面分析这个结果。首先可以查表或者使用数学软件求出双尾检验的两个临界值分别为−2.1199 和 2.1199。由于原假设是 pH=7，那么它不成立的情况就有两种，要么 pH>7，要么 pH<7，所以它是一个双尾检验。如图 5-4 所示，其中两部分阴影的面积之和占总图形面积的 5%，即两边各 2.5%。已经算得的 t 统计量要小于临界值−2.1199，对称地，t 统计量的相反数也大于另外一个临界值 2.1199，即样本数据的统计量落入了拒绝域中。样本数据的统计量对应的 P 值也小于 0.05 的显著水平，所以应该拒绝原假设。因此认为该区域的土壤不是中性的。

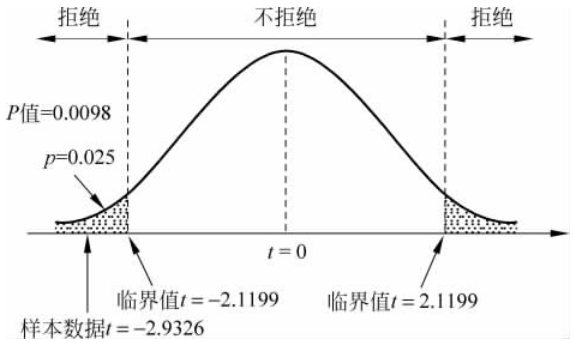

图 5-4　双尾检测的拒绝域与接受域

除了进行双尾检验以外，当然还可执行一个单尾检验。例如现在问该区域的土壤是否呈酸性(即 pH<7)，那么便可提出如下的原假设与备择假设

$$H_0:\text{pH}=7, H_1:\text{pH}<7$$

此时所得之 t 统计量并未发生变化，但是 P 值却不同了，可以算得 P 值为 0.004 878 676。

如图 5-5 所示，t 统计量小于临界值−1.7459，即样本数据的统计量落入了拒绝域中。样本数据的统计量对应的 P 值也小于 0.05 的显著水平，所以应该拒绝原假设。因此认为该区域的土壤是酸性的。

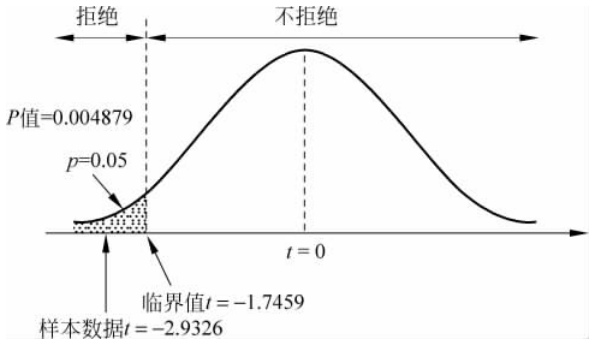

图 5-5　单尾检测的拒绝域与接受域

相比之下，讨论双总体均值之差的假设检验其实更有意义。因为在统计实践中，最常被问到的问题就是两个总体是否有差别。例如，医药公司研发了一种新药，在进行双盲对照实验时，新药常常被用来与安慰剂做比较。如果新药在统计上不能表现出与安慰剂的显著差别，显然这种药就是无效的。再比如前面讨论过的饲料问题，当对比两种饲料的效果时，必然要问及它们之间是否有差别。

同在研究双总体均值差的区间估计问题时所遵循的思路一致，此时仍然分独立样本数据和配对样本数据两种情况来讨论。

对于独立样本数据而言，如果两个总体的方差 σ_1^2 和 σ_2^2 未知，但是可以确定 $\sigma_1^2=\sigma_2^2$，那么在此情况下检验统计量的计算公式为

$$t=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

其中，s'的表达式本章前面曾经给出过，这里不再重复。另外，t 分布的自由度为n_1+n_2-2。

对于独立样本数据，若两个总体的方差 σ_1^2 和 σ_2^2 未知，且 $\sigma_1^2\neq\sigma_2^2$，那么在此情况下检验统计量的计算公式为

$$t=\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$$

此时检验统计量近似服从一个自由度为 $\hat{\nu}$ 的 t 分布，$\hat{\nu}$ 前面已经给出，这里不再重复。

仍然以饲料与肉鸡增重的数据为例，并假设两个总体的方差不相等，同样提出原假设和备择假设如下

$$H_0:\mu_1=\mu_2,H_1:\mu_1\neq\mu_2$$

在原假设前提下，可以计算检验统计量的数值为

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{s_1^2/n_1+s_2^2/n_2}}=\frac{65-79.83}{\sqrt{\frac{21.66^2}{3}+\frac{23.87^2}{6}}}=\frac{-14.83}{15.854}\approx-0.9357$$

这仍然是一个双尾检测，所以可以求得检验临界值为－2.658 和 2.658。因为－2.658≤－0.9357≤2.658，所以检验统计量落在了接受域中。更进一步还可以算得与检验统计量相对应的 P 值等于 0.3968、大于 0.05 的显著水平，所以无法拒绝原假设，即不能认为两种饲料之间存在差异。

最后来研究双总体均值差的假设检验中，样本数据属于配对样本的情况。此时的假设检验其实与单总体均值的假设检验基本相同，即把配对样本之间的差值看成是从单一总体中抽取的一组样本。在大样本条件下，两个总体间各差值的标准差 σ_d 未知，所以用样本差值的标准差 s_d 来代替，此时统计量的计算公式为

$$z = \frac{\bar{d} - \mu}{s_d / \sqrt{n}}$$

其中，d 是一组配对样本之间的差值，$\bar{d}$ 表示各差值的均值，μ 表示两个总体中配对数据差的均值。

在样本量较小的情况下，样本统计量就服从 t 分布，故应使用 t 统计量，其计算公式为

$$t = \frac{\bar{d} - \mu}{s_d / \sqrt{n}}$$

其中，t 统计量的自由度为 $n-1$。

继续前面关于双总体均值差中配对样本的讨论，欲检验喂食了两组不同饲料的肉鸡在增重数据方面是否具有相同的均值，现提出下列原假设和备择假设

$$H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2$$

在原假设前提下，很容易得出配对差的均值 μ 也为零的结论，于是可以计算检验统计量如下

$$t = \frac{6.67}{5.05\sqrt{6}} = \frac{6.67}{2.062} \approx 3.235$$

这仍然是一个双尾检测，所以可以求得检验临界值 -2.571 和 2.571。因为 $3.235 \geqslant 2.571$，所以检验统计量落在了拒绝域中。更进一步还可以算得与检验统计量相对应的 P 值等于 0.023 05、小于 0.05 的显著水平，所以应该拒绝原假设，即认为两种饲料之间存在差异。

5.4　极大似然估计

正如本章前面所讲的，统计推断的基本问题可以分为两大类：一类是参数估计；另一类是假设检验。其中，假设检验又分为参数假设检验和非参数假设检验两大类。本章所讲的假设检验都属于是参数假设检验的范畴。参数估计也分为两大类，即参数的点估计和区间估计。用于点估计的方法一般有矩方法和最大似然估计法（Maximum Likelihood Estimate，MLE）两种。

5.4.1　极大似然法的基本原理

最大似然这个思想最初是由高斯提出的，但真正将其发扬光大的则是费希尔。费希尔在其 1922 年发表的一篇论文中再次提出了最大似然估计这个思想，并且首先探讨了这种方法的一些性质。而且，费希尔当年正是凭借这一方法彻底撼动了皮尔逊在统计学界的统治地位。从此开始，统计学研究正式进入了费希尔时代。

为了引入最大似然估计法的思想，先来看一个例子。设一个口袋中有黑白两种颜色的小球，并且知道这两种球的数量比为 3∶1，但不知道具体哪种球占 3/4，哪种球占 1/4。现在从袋子中有返回地任取 3 个球，其中有一个是黑球，那么试问袋子中哪种球占 3/4，哪种球占 1/4。

设 X 是抽取 3 个球中黑球的个数，又设 p 是袋子中黑球所占的比例，则有 $X\sim B(3,p)$，即

$$P(X=k)=\left(\frac{3}{k}\right)p^k(1-p)^{3-k},k=0,1,2,3$$

当 $X=1$ 时，不同的 p 值对应的概率分别为

$$P\left(X=1;p=\frac{3}{4}\right)=3\times\frac{3}{4}\times\left(\frac{1}{4}\right)^2=\frac{9}{64}$$

$$P\left(X=1;p=\frac{1}{4}\right)=3\times\frac{1}{4}\times\left(\frac{3}{4}\right)^2=\frac{27}{64}$$

由于第一个概率小于第二个概率，所以判断黑球的占比应该是 1/4。

在上面的例子中，p 是分布中的参数，它只能取 3/4 或者 1/4。需要通过抽样结果来决定分布中参数究竟是多少。在给定了样本观察值以后再去计算该样本的出现概率，而这一概率依赖于 p 值。所以就需要用 p 的可能取值分别去计算最终的概率，在相对比较之下，最终所取的 p 值应该是使得最终概率最大的那个 p 值。

极大似然估计的基本思想就是根据上述想法引申出来的。设总体含有待估参数 θ，它可以取很多值，所以就要在 θ 的一切可能取值之中选出一个使样本观测值出现概率为最大的 θ 值，记为 $\hat{\theta}$，并将此作为 θ 的估计，并称 $\hat{\theta}$ 为 θ 的极大似然估计。

首先来考虑 X 属于离散型概率分布的情况。假设在 X 的分布中含有未知参数 θ，记为

$$P(X=a_i)=p(a_i;\theta),i=1,2,\cdots,\theta\in\Theta$$

现从总体中抽取容量为 n 的样本，其观测值为 $x_1,x_2,\cdots,x_n$，这里每个 x_i 为 $a_1,a_2,\cdots$ 中的某个值，该样本的联合分布为

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

由于这一概率依赖于未知参数 θ，故可将它看成是 θ 的函数，并称其为似然函数，记为

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

对不同的 θ，同一组样本观察值 $x_1,x_2,\cdots,x_n$ 出现的概率 $L(\theta)$ 也不一样。当 $P(A)>P(B)$ 时，事件 A 出现的可能性比事件 B 出现的可能性大，如果样本观察值 $x_1,x_2,\cdots,x_n$ 出现了，当然就要求对应的似然函数 $L(\theta)$ 的值达到最大，所以应该选取这样的 $\hat{\theta}$ 作为 θ 的估计，使得

$$L(\hat{\theta})=\max_{\theta\in\Theta}L(\theta)$$

如果 $\hat{\theta}$ 存在的话，则称 $\hat{\theta}$ 为 θ 的极大似然估计。

此外，当 X 是连续分布时，其概率密度函数为 $p(x;\theta)$，θ 为未知参数，且 $\theta\in\Theta$，这里的 Θ 表示一个参数空间。现从该总体中获得容量为 n 的样本观测值 $x_1,x_2,\cdots,x_n$，那么在 $X_1=x_1,X_2=x_2,\cdots,X_n=x_n$ 时联合密度函数值为

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

它也是 θ 的函数，也称为似然函数，记为

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

对不同的 θ，同一组样本观察值 $x_1,x_2,\cdots,x_n$ 的联合密度函数值也是不同的，因此应该选择 θ 的极大似然估计 $\hat{\theta}$，从而使下式得到满足

$$L(\hat{\theta})=\max_{\theta\in\Theta} L(\theta)$$

5.4.2　求极大似然估计的方法

当函数关于参数可导时，可以通过求导方法来获得似然函数极大值对应的参数值。在求极大似然估计时，为求导方便，常对似然函数 $L(\theta)$ 取对数，称 $l(\theta)=\ln L(\theta)$ 为对数似然函数，它与 $L(\theta)$ 在同一点上达到最大。根据微积分中的费马定理，当 $l(\theta)$ 对 θ 的每一分量可微时，可通过 $l(\theta)$ 对 θ 的每一分量求偏导并令其为 0 求得，称

$$\frac{\partial l(\theta)}{\partial \theta_j}=0, j=1,2,\cdots,k$$

为似然方程，其中 k 是 θ 的维数。

下面就结合一个例子来演示这个过程。假设随机变量 $X\sim B(n,p)$，又知 $x_1,x_2,\cdots,x_n$ 是来自 X 的一组样本观察值，现在求 $P(X=T)$ 时，参数 p 的极大似然估计。首先写出似然函数

$$L(p)=\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

然后，对上式左右两边取对数，可得

$$l(p)=\sum_{i=1}^{n}[x_i\ln p+(1-x_i)\ln(1-p)]=n\ln(1-p)+\sum_{i=1}^{n}x_i[\ln p-\ln(1-p)]$$

将 $l(p)$ 对 p 求导，并令其导数等于 0，得似然方程

$$\frac{\mathrm{d}l(p)}{\mathrm{d}p}=-\frac{n}{1-p}+\sum_{i=1}^{n}x_i\left(\frac{1}{p}+\frac{1}{1-p}\right)$$

$$=-\frac{n}{1-p}+\frac{1}{p(1-p)}\sum_{i=1}^{n}x_i=0$$

解似然方程得

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

可以验证，当 $\hat{p}=\bar{x}$ 时，$\partial^2 l(p)/\partial p^2<0$，这就表明 $\hat{p}=\bar{x}$ 可以使函数取得极大值。最后将题目中已知的条件代入，可得 p 的极大似然估计为 $\hat{p}=\bar{x}=T/n$。

再来看一个连续分布的例子。假设有随机变量 $X\sim N(\mu,\sigma^2)$，μ 和 σ^2 都是未知参数，$x_1,x_2,\cdots,x_n$ 是来自 X 的一组样本观察值，试求 μ 和 σ^2 的极大似然估计值。首先写出似然函数

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}\cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

然后，对上式左右两边取对数，可得

$$l(\mu,\sigma^2)=-\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

将 $l(\mu,\sigma^2)$ 分别对 μ 和 σ^2 求偏导数，并令它们的导数等于 0，于是可得似然方程

$$\begin{cases}\dfrac{\partial l(\mu,\sigma^2)}{\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0\\[2ex]\dfrac{\partial l(\mu,\sigma^2)}{\sigma^2}=-\dfrac{n}{2\sigma^2}+\dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases}$$

求解似然方程可得

$$\hat{\mu}=\bar{x},\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=0$$

而且还可以验证 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 可以使得 $l(\mu,\sigma^2)$ 达到最大。用样本观察值替代后便得出 μ 和 σ^2 的极大似然估计分别为

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

因为 $\hat{\mu}=\overline{X}$ 是 μ 的无偏估计，但 $\hat{\sigma}^2=S_n^2$ 并不是 σ^2 的无偏估计，可见参数的极大似然估计并不能确保无偏性。

最后给出一个被称为“不变原则”的定理：设 $\hat{\theta}$ 是 θ 的极大似然估计，$g(\theta)$ 是 θ 的连续函数，则 $g(\theta)$ 的极大似然估计为 $g(\hat{\theta})$。

这里并不打算对该定理进行详细证明。下面将通过一个例子来说明它的应用。假设随机变量 X 服从参数为 λ 的指数分布，$x_1,x_2,\cdots,x_n$ 是来自 X 的一组样本观察值，试求 λ 和 $E(X)$ 的极大似然估计值。首先写出似然函数

$$L(\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda\sum_{i=1}^{n}x_i}$$

然后，对上式左右两边取对数，可得

$$l(\lambda)=n\ln\lambda-\lambda\sum_{i=1}^{n}x_i$$

将 $l(\lambda)$ 对 λ 求导得似然方程为

$$\frac{dl(\lambda)}{d\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}x_i=0$$

解似然方程得

$$\hat{\lambda}=n\Big/\sum_{i=1}^{n}x_i=\frac{1}{\bar{x}}$$

可以验证它使 $l(\lambda)$ 达到最大，而且上述过程对一切样本观察值都成立，所以 λ 的极大似然估计值为 $\hat{\lambda}=1/\overline{X}$。此外，$E(x)=1/\lambda$，它是 λ 的函数，其极大似然估计可用不变原则进行求解，即用 $\hat{\lambda}$ 代入 $E(x)$，可得 $E(x)$ 的最大似然估计为 $\overline{X}$，这与矩法估计的结果一致。

本章参考文献

[1] 贾俊平，何晓群，金勇进. 统计学[M]. 4 版. 北京：中国人民大学出版社，2009.
[2] 奥特，朗格内克. 统计学方法与数据分析引论[M]. 5 版. 张忠占，等译. 北京：科学出版社，2003.
[3] 萨尔斯伯格. 女士品茶：20 世纪统计怎样变革了科学[M]. 邱东，等译. 北京：中国统计出版社，2004.
[4] Dawen Griffiths. 深入浅出统计学. 北京：电子工业出版社，2012.
[5] Mario F. Triola. 初级统计学[M]. 8 版. 刘新立，译. 北京：清华大学出版社，2004.