第 3 章

统 计 推 断

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。统计推断要面对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，本章将着重关注点估计和区间估计这两类问题。

3.1 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。但又因为分布仅仅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、现成的数学模型描述。在处理这类问题时，采样就变得非常重要。在统计学中，采样（或称抽样）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），通过观察样本的某一或某些属性，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，我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，容量为无限的则称为无限总体。

总体中的每个个体是随机试验的一个观察值（observation），因此它与某一随机变量 X 的值相对应。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对总体

分布作出推断。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 X 进行一次观察并记录其结果。在相同的条件下对总体 X 进行 n 次重复的、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，…，X_n。由于 X_1，X_2，…，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，所以可以认为 X_1，X_2，…，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，…，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，我们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，…，x_n，它们依次是随机变量 X_1，X_2，…，X_n 的观察值，称为样本值。

设 X 是具有某种分布函数 F 的随机变量，若 X_1，X_2，…，X_n 是具有同一分布函数 F 的、相互独立的随机变量，则称 X_1，X_2，…，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，…，x_n 称为样本值，又称为 X 的 n 个独立的观察值。亦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 $(y_1, y_2, \cdots, y_n)$ 都是对应于样本 $(X_1, X_2, \cdots, X_n)$ 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，…，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，…，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，…，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，…，x_n 是相应于样本 X_1，X_2，…，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，…，X_n 是来自总体 X 的一个样本，x_1，x_2，…，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

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

很多人会对上面的公式感到困惑，疑问之处就在于为什么样本方差计算公式里分母为 $n-1$？简单地说，这样做的目的是为了使方差的估计无偏。无偏估计（Unbiased Estimator）的意思是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）。之所以进行采样，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，我们用样本的一些数量指标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 之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？这是令很多人倍感迷惑的地方。

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

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

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，注意式中的分母 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，我们会倾向于直接用样本均值 $\overline{X}$ 替换上面式子中的 μ。这样做的后果是什么呢？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

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

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，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”地估计方差呢？答案是把上式中的分母 n 换成 $n-1$，通过这种方法把原来的偏小的估计“放大”一点点，就能获得对方差的正确估计了。这个结论是可以证明的。

下面证明

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

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

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

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

而且有

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

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

所以可得

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

结论得证。

既然已经知道样本方差的定义式为

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

那么也就可以据此给样本协方差定义如下：

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

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

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

即 $E(s^2)=\sigma^2$。

回忆前面给出的一个结论：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

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

如果将其转换为标准正态分布的形式，就会得出

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

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（此时即变成中心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（或简称 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 分布依赖于称为自由度的参数。自由度越小，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。下面的代码绘制了标准正态分布以及两个自由度不同的 t 分布，结果如图 3-1 所示。

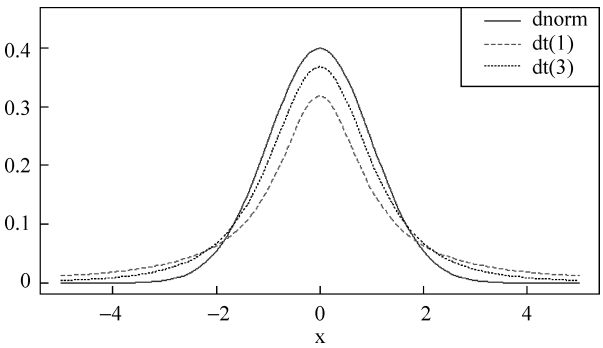

图 3-1　标准正态分布与 t 分布

```
> curve(dnorm(x), from =- 5, to = 5, ylim = c(0, 0.45),
 +   ylab = "", col = "blue")
> par(new = TRUE)
> curve(dt(x, 1), from =- 5, to = 5, ylim = c(0, 0.45),
 +   ylab = "", lty = 2, col = "red")
> par(new = TRUE)
> curve(dt(x, 3), from =- 5, to = 5, ylim = c(0, 0.45),
 +   ylab = "", lty = 3)
> text.legend = c("dnorm","dt(1)", "dt(3)")
> legend("topright", legend = text.legend, lty = c(1,2,3),
 +   col = c("blue", "red", "black"))
```

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(Fisher)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)。

3.2 参数估计

如果想知道某所中学高三年级全体男生的平均身高,其实只要测定他们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实在过于庞大,要想获得全体中国成年男性的身高数据显然有点不切实际。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)。

3.2.1 参数估计的基本原理

点估计(Point Estimate)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等。这种方式的点估计也称为矩估计,其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,若总体 X 的 k 阶矩存在,则样本的 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,这就启发我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。

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)诞辰百年,谷歌特别在其主页上刊出了一幅纪念画,如图3-2所示。第二次世界大战以后,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。其中,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,3145人死亡,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。

索尔克在验证他发明的疫苗的效果时,设计了一个随机双盲对照试验,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。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

图 3-2　索尔克纪念画

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

或者在 R 语言中执行下面的代码计算结果。

```
> 57/200745
[1] 0.0002839423
```

虽然在重复采样下，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。通常我们会围绕点估计值来构造总体参数的一个区间，并用这个区间来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，根据样本统计量的采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，由样本均值的采样分布知，在重复采样或无限总体采样的情况下，样本均值的数学期望等于总体均值，即 $E(\bar{x})=\mu$。回忆前面曾经给出过的一些结论，还可以知道，样本均值的标准差等于 $\sigma_{\bar{x}}=\sigma/\sqrt{n}$，其中 σ 是总体的标准差，n 是样本容量。根据中心极限定理可知，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 的两侧各一个采样标准差范围内的概率为 0.6827；落在两个采样标准差范围内的概率为 0.9545；落在 3 个采样标准差范围内的概率是 0.9973……下面是 R 中用于计算的代码。

```
> pnorm(1) - pnorm( - 1)
[1] 0.6826895
> pnorm(2) - pnorm( - 2)
[1] 0.9544997
> pnorm(3) - pnorm( - 3)
[1] 0.9973002
```

事实上，我们完全可以求出样本均值落在总体均值两侧任何一个采样标准差范围内的概率。但实际估计时，情况却是相反的。我们所知道的仅仅是样本均值 $\bar{x}$，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，大约有 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 3-3 给出了区间估计的示意图。

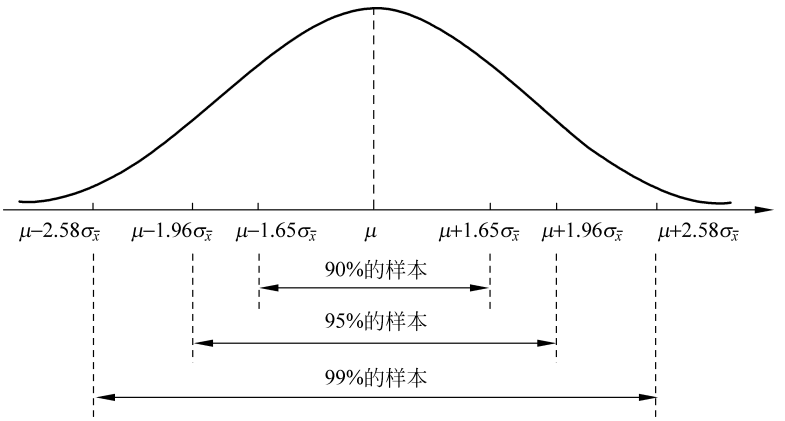

图 3-3 区间估计示意图

在区间估计中，由样本统计量所构造的总体参数之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之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 3-1 所示。

表 3-1 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

3.2.2 单总体参数区间估计

1. 总体比例的区间估计

比例问题可以看作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：要么是患病，要么是不患病。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

$$\mathrm{var}(\hat{p})=\frac{1}{n^2}\mathrm{var}(X)=\frac{1}{n^2}\cdot np(1-p)=\frac{p(1-P)}{n};\quad se(\hat{p})=\sqrt{\frac{p(1-P)}{n}}$$

由此便已经具备了进行区间估计的必备素材。

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

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

3.2.1 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，即

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

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

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，可以在 R 中使用下面的代码算得总体比例估计的置信区间。由输出结果可知，保留小数点后 6 位有效数字的置信区间为(0.000 210，0.000 358)。

```
> n <- 200745
> (p.hat <- 57/n)
[1] 0.0002839423
> p.hat + c(-1.96, 1.96) * sqrt(p.hat * (1 - p.hat)/n)
[1] 0.0002102390 0.0003576456
```

Wald 方法的基本原理是利用正态分布来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，所以它是一种更加精确的区间估计方法。在 R 中可以使用 binom.test()函数执行 Clopper-Pearson 方法，下面给出示例代码。

```
> binom.test(57,200745)

        Exact binomial test

data:  57 and 200745
number of successes = 57, number of trials = 200745, p-value <
2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.0002150620 0.0003678648
sample estimates:
probability of success
          0.0002839423
```

从以上代码的输出中可以得到，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)。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。首先若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 $\bar{x}$ 的采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 是显著水平,它是总体均值不包含在置信区间内的概率;$z_{\alpha/2}$ 是标准正态分布曲线与横轴围成的面积等于 $\alpha/2$ 时的 z 值。

假设现在有一家生产袋装食品的食品厂。按照规定每袋食品的重量应为 100 克。为对产品质量进行监测,质检部门从当天生产的一批食品中随机抽取了 25 袋,并测得每袋的重量数据如表 3-2 所示。已知产品重量的分布服从正态分布,且总体标准差为 10 克。请计算该天每袋食品平均重量的置信区间,置信水平为 95%。

表 3-2 食品重量抽检数据 (单位:克)

112.5	101.0	103.0	102.0	100.5
102.6	107.5	95.0	108.8	115.6
100.0	123.5	102.0	101.6	102.2
116.6	95.4	97.8	108.6	105.0
136.8	102.8	101.5	98.4	93.3

由于 R 中并没有提供方差已知时置信区间的计算函数,所以需要手动编写函数,代码如下。

```
> conf.int <- function(x,n,sigma,alpha){
    options(digits = 5)
    mean <- mean(x)
    c(mean - sigma * qnorm(1 - alpha/2,mean = 0, sd = 1,
    lower.tail = TRUE)/sqrt(n),
    mean + sigma * qnorm(1 - alpha/2,mean = 0, sd = 1,
    lower.tail = TRUE)/sqrt(n))
}
```

然后调用上述函数来计算置信区间,代码如下。

```
> x <- c(112.5, 101.0, 103.0, 102.0, 100.5,
+    102.6, 107.5, 95.00, 108.8, 115.6,
+    100.0, 123.5, 102.0, 101.6, 102.2,
+    116.6, 95.40, 97.80, 108.6, 105.0,
+    136.8, 102.8, 101.5, 98.40, 93.30)
```

```
> n <- 25
> alpha <- 0.05
> sigma <- 10
> result <- conf.int(x, n, sigma, alpha)
> result
[1] 101.44 109.28
```

结果表明，该批食品平均重量95%的置信区间为(101.44，109.28)。

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

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

需要注意的是，也是本书前面着重讨论的一点，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之无偏估计的样本方差公式为

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

此外，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

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

于是，就需要采用学生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

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

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

例如现在为了测定一块土地的pH值，随机抽取了17块土壤样本，相应的pH值检测结果如表3-3所示。由于样本容量仅为17，所以属于小样本的情况，于是采用上述方法对这块土地的pH值均值进行区间估计。

表3-3　土壤pH值检测数据

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	

根据已经给出的公式可以在R语言中下面的代码进行区间估计，其估计结果用方框来进行标识。

```
> pH <- c(6, 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)
> mean(pH); sd(pH)
[1] 6.676471
```

```
[1] 0.4548755
> mean(pH) + qt(c(0.025,0.975),length(pH) - 1) * sd(pH)/sqrt(length(pH))
[1] 6.442595 6.910346
```

或者更简单地，可以直接使用 R 中的 t.test()函数计算，示例代码如下。结果同样用方框进行标识，可见与前面得到的结果是一致的。

```
> t.test(pH, mu = 7)

        One Sample t - test

data:  pH
t = - 2.9326, df = 16, p - value = 0.009758
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
6.442595 6.910346
sample estimates:
mean of x
 6.676471
```

表 3-4 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 3-4 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。根据样本方差的采样分布可知，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

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

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

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

据此便可对总体方差的置信区间进行估计。由于 R 中并没有提供直接用于方差区间估计的函数，我们便自行编写了下面这个函数用以执行相应计算。

```
> chisq.var.test <- function(x, alpha){
    options(digits = 4)
    result <- list( )
    n <- length(x)
    v <- var(x)
    result$conf.int.var <- c(
        (n-1) * v/qchisq(alpha/2, df = n-1, lower.tail = F),
        (n-1) * v/qchisq(alpha/2, df = n-1, lower.tail = T))
    result$conf.int.se <- sqrt(result$conf.int.var)
    result
    }
```

以食品厂抽检产品重量的数据为例，调用以上函数可以算得 $56.83 \leqslant \sigma^2 \leqslant 180.39$。相应地，总体标准差的置信区间则为 $7.538 \leqslant \sigma \leqslant 13.431$，即该食品厂生成的食品总体重量标准差 95%的置信区间为 7.538～13.431g。

```
> chisq.var.test(x, 0.05)
 $conf.int.var
[1]  56.83 180.39
 $conf.int.se
[1]  7.538 13.431
```

3.2.3　双总体均值差的估计

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：$C_1X_1+C_2X_2+\cdots+C_nX_n$，仍服从正态分布，其中 $C_1, C_2, \cdots, C_n$ 是不全为 0 的常数。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潜在的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

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

由此即得到了进行双总体均值之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，则称之为独立样本(Independent Samples)。

当两总体的方差 σ_1^2 和 σ_2^2 已知时，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，

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

此处的 s_1^2 和 s_2^2 分别是样本方差。与之前的做法类似，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 3-5 所示。

表 3-5 喂食不同饲料的肉鸡增重情况

饲料	增重/g
1	42,68,85
2	42,97,81,95,61,103

首先在 R 中录入数据，并分别计算两组数据的均值和方差，示例代码如下。

```
> chicks <- data.frame(feed = rep(c(1,2), times = c(3,6)),
 +                     weight_gain = c(
```

```
+                   42, 68, 85,
+                   42, 97, 81, 95, 61, 103))

> tapply(chicks$weight_gain, chicks$feed, mean)
       1        2
65.00000 79.83333
> tapply(chicks$weight_gain, chicks$feed, sd)
       1        2
21.65641 23.86979
```

从输出结果看，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

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

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

或者在 R 中使用 t.test()函数执行上述计算过程，示例代码如下。区间估计的结果已经用方框标识。这个输出中的其他指标结果将在假设检验部分继续讨论。

```
> t.test(weight_gain ~ feed, data = chicks, var.equal = TRUE)

        Two Sample t-test
data:  weight_gain by feed
t = -0.9019, df = 7, p-value = 0.3971
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -53.72318   24.05651
sample estimates:
mean in group 1 mean in group 2
       65.00000        79.83333
```

通过设置函数 t.test()中的参数可以修改它的一些执行细节，具体参数列表读者可以参阅 R 的帮助文档。这里仅提其中几个比较重要的参数。首先，参数 paired 的默认值为 FALSE，表示执行的是独立样本的情况。若将其置为 TRUE，则表示要处理的是配对样本。参数 conf.level 的默认值为 0.95，即在 95% 的置信水平下进行区间估计，调整它便可以改变置信水平。参数 var.equal 默认为 FALSE，如果将其置为 TRUE，则表示两个总体具有相同的方差。

此外，当两总体的方差未知，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，

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

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

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

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，可以算得

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

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

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

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

同样，上述计算过程可以在 R 中使用 t.test()函数完成，示例代码如下。输出中的其他指标结果在假设检验部分还会有更为详细的讨论。

```
> t.test(weight_gain ~ feed, data = chicks)

        Welch Two Sample t-test

data:  weight_gain by feed
t = -0.9357, df = 4.503, p-value = 0.3968
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -56.97338  27.30671
sample estimates:
mean in group 1 mean in group 2
       65.00000        79.83333
```

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采样，但使用独立样本来估计两个总体均值之差也潜在一些弊端。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大折扣。这种“有失公平”的独立采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，可选用同一窝诞下的一对小鸡作为一个配对组，因为我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 3-6 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如下。

表 3-6　配对试验数据　（单位：克）

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 是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 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 3-6 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

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

同样可以在 R 中使用 t.test()函数完成以上计算过程，此时需要将参数 paired 置为 TRUE。示例代码如下，输出结果中的置信区间估计已经用方框标出。这个区间估计不包含零，其实也就意味二者是存在差异的，即饲料 1 和饲料 2 的喂食结果不同。

```
> Feed.1 <- c(44, 55, 68, 85, 90, 97)
> Feed.2 <- c(42, 61, 81, 95, 97, 103)
> t.test(Feed.2, Feed.1, paired = T)

        Paired t-test

data:  Feed.2 and Feed.1
t = 3.2359, df = 5, p-value = 0.02305
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.370741 11.962592
sample estimates:
mean of the differences
                6.666667
```

当然，如果先计算配对组之差，然后再执行 t.test()，所得结果将是一样的。读者可以自行尝试下面的代码，并观察结果。

```
> diff = Feed.2 - Feed.1
> t.test(diff)
```

最后需要说明的是，如果仅是执行普通的 t.test()，而非是做配对数据的 t.test()，那么将得到一个宽泛得多的区间估计。如下代码所示，而且最终估计的置信区间还包含了零，这使得我们无法确定饲料 1 和饲料 2 的喂食结果是否有不同。

```
> Feed <- c(Feed.1, Feed.2)
> group <- c(rep(1, 6), rep(2, 6))
> t.test(Feed ~ group)

        Welch Two Sample t-test

data:  Feed by group
t = -0.514, df = 9.837, p-value = 0.6186
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35.63370  22.30037
sample estimates:
mean in group 1 mean in group 2
       73.16667        79.83333
```

3.2.4 双总体比例差的估计

由样本比例的采样分布可知，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采样服从正态分布，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，即得

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

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 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%的置信水平估计城市与农村收视率差距的置信区间。

在 R 中可以使用 prop.test()函数执行双总体比例差的区间估计，示例代码如下。输出结果中的置信区间估计已经用方框标出。参数 correct 的默认值为 TRUE，表示计算过程中需要使用连续性修正。如果将其置为 FALSE，则所得结果将同依据上述公式所得结果完全一致。

```
> prop.test(x = c(225,128),n = c(500,400), correct = F)

        2 - sample test for equality of proportions without continuity
        correction

data:  c(225, 128) out of c(500, 400)
X - squared = 15.7542, df = 1, p - value = 7.213e - 05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06682346 0.19317654
sample estimates:
prop 1 prop 2
  0.45   0.32
```

从输出结果中可以看出，估计的置信区间为(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，所得结果如下。

```
> prop.test(x = c(225,128),n = c(500,400))

        2 - sample test for equality of proportions with continuity
        correction

data:  c(225, 128) out of c(500, 400)
X - squared = 15.2136, df = 1, p - value = 9.601e - 05
alternative hypothesis: two.sided
95 percent confidence interval:
 0.06457346 0.19542654
sample estimates:
prop 1 prop 2
  0.45   0.32
```

3.3　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。换言之，对总体的某个假设是真实的，那么不利于或不能支持这一假设的事件在一次试验中几乎不可能发生；如果在一次试验中该事件竟然发生了，我们就有理由怀疑这一假设的真

实性，进而拒绝这一假设。

3.3.1 基本概念

戴维·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。而开篇的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，二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与遗传学家、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，则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，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，我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为零，这时就有足够的理由推翻原假设。也就是说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 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。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受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，当然也是有把握的、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，同时也是无把握的、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”。如果将这个命题作为 H_0，其实也就等同于那个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，则可以算出 $\hat{P}=6/8=0.75$，这个值大于0.5，但这是否大到可以令我们相信先放茶水还是先放牛奶确有不同这个结论。所以需要计算一下 P 值，即 $\Pr(X\geqslant 6)$。使用下面这段代码可以算得 P 值是0.144 531 2。

```
> 1 - pbinom(5, size = 8, prob = 0.5)
[1] 0.1445312
```

可见，P 值并不是很显著。通常都需要 P 值小于 0.05，才能令我们有足够的把握拒绝原假设。而本题所得之结果则表明没有足够的证据支持我们拒绝原假设。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。我们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，此时就令 P 值为 1。另外，当 n 较大的时候，还可以用正态分布近似二项分布。

1965 年，美国联邦最高法院对斯文诉阿拉巴马州一案做出了裁定。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，他被控于阿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。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。

最终，最高法院驳回了上述请求。根据阿拉巴马州法律，陪审团成员是从一个 100 人的名单中抽选的，而当时的 100 个备选成员中有 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法律保护的。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。

阿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人大约有 16 000 人，其中 26％是非洲裔美国人。我们现在的问题就是，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？可以在 R 中用下列命令计算得到我们想要的答案。

```
> pbinom(8, 100, 0.26)
[1] 4.734795e-06
```

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来计算置信区间。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 值是直接使用 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的。

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 $H_1:p<0.26$。在一个 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

$$\Pr\left(Z\leqslant\frac{0.08-0.26}{\sqrt{0.26\times0.74/100}}\right)=\Pr(Z\leqslant-4.104)=0.0000\ 20$$

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（它是连续的）作为二项分布（它是离散的）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

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

此处无意要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，那么所得之 P 值将总是偏小，相应的置信区间也偏窄。

上述计算过程在 R 中可以使用 prop.test()实现，示例代码如下。

```
> prop.test(8,100,p = 0.26,alternative = "less")

        1 - sample proportions test with continuity correction

data:  8 out of 100, null probability 0.26
X - squared = 15.9174, df = 1, P - value = 3.308e - 05
alternative hypothesis: true p is less than 0.26
95 percent confidence interval:
0.0000000 0.1424974
sample estimates:
   p
0.08
```

如同前面所分析的那样，如果不使用正态分布对二项分布做近似，仅仅基于二项分布来进行检验也是可行的。此时需要用到 binom.test()函数，示例代码如下。

```
> binom.test(8,100,p = 0.26,alternative = "less")

        Exact binomial test

data:  8 and 100
number of successes = 8, number of trials = 100, p - value = 4.735e - 06
alternative hypothesis: true probability of success is less than 0.26
95 percent confidence interval:
```

```
0.0000000 0.1397171
sample estimates:
probability of success
                0.08
```

3.3.2　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。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，而在整个推理判断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，所以利用前面的方法进行假设检验时，有可能做出错误的判断。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，我们却将其拒绝，如果犯这种错误的概率用 α 表示，那么当 H_0 为真时，我们没有拒绝它，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，我们却没有拒绝它，犯这种错误的概率用 β 表示。那么，当 H_0 为假，我们也正确地拒绝了它，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 3-7。

表 3-7　假设检验中各种可能结果及其概率

情　形	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，β 值也就随之变大；相反，若 β 变小，则不难想到 α 又会变大。有时不得不在两类错误之间权衡。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。

3.3.3 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需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，所以在更多的情况下，我们都认为总体标准差 σ 是未知的。3.2 节已经介绍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；即使总体是非正态分布的，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 代替，如此一来，上述式子便可改写为

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

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 $n-1$。

仍以土壤 pH 值检验的数据为例，现在我们想问该区域的土壤是否是中性的(即 pH=7)？为此首先提出原假设和备择假设如下：

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

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，所以可在 R 中计算其 P 值如下：

```
> 2 * pt( - 2.9326, 16, lower.tail = T)
[1] 0.009757353
```

注意：以上结果与先前使用 t.test()函数算得之结果是一致的，下面分析这个结果意味着什么。首先可以在 R 中使用下面的代码求出双尾检验的两个临界值。

```
> qt(0.025, 16); qt(0.975, 16)
[1] -2.119905
[1] 2.119905
```

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 3-4 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%。一方面已经算得的 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。由此认为该区域的土壤不是中性的。

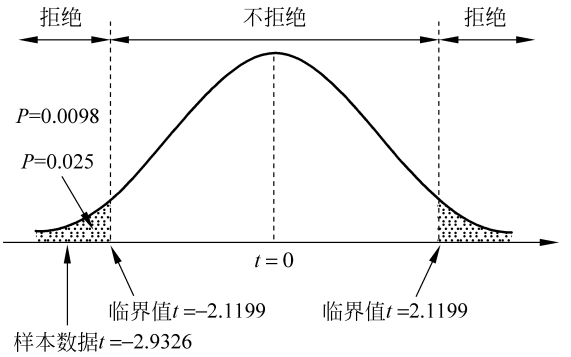

图 3-4　双尾检测的拒绝域与接受域

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），那么可提出如下的原假设与备择假设：

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

此时所得之 t 统计量并未发生变化，但是 P 值却不同了，可以在 R 中算得 P 值如下：

```
> pt( -2.9326, 16)
[1] 0.004878676
```

如图 3-5 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。由此认为该区域的土壤是酸性的。

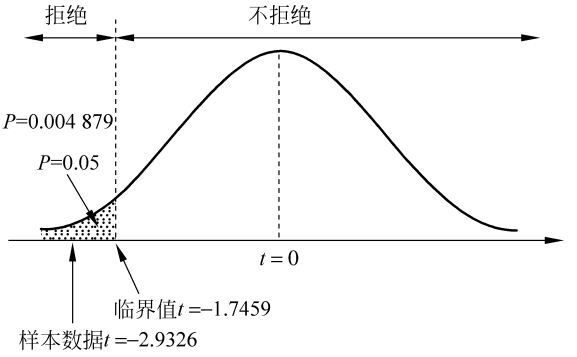

图 3-5　单尾检测的拒绝域与接受域

以上单尾检验过程也可以使用 t.test()函数完成，只需将其中的参数 alternative 的值置为"less"。下面给出示例代码。

```
> t.test(pH, mu = 7, alternative = "less")

        One Sample t-test

data:  pH
t = -2.9326, df = 16, p-value = 0.004879
alternative hypothesis: true mean is less than 7
95 percent confidence interval:
     -Inf 6.869083
sample estimates:
mean of x
6.676471
```

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，那么这种药就是无效的。再例如前面讨论过的饲料问题，当我们对比两种饲料的效果时，必然了解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，仍然分独立样本数据和配对样本数据两种情况讨论。

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，这里不再重复。另外，t 分布的自由度为 n_1+n_2-2。

仍然以饲料与肉鸡增重的数据为例，现在想知道两种饲料在统计上是否有差异，为此提出原假设和备择假设如下：

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

在原假设前提下，可以计算检验统计量的数值为

$$t=\frac{\bar{x}_1-\bar{x}_2}{s'\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{-14.83}{16.447}\approx-0.9019$$

这仍然是一个双尾检测，所以可以使用如下所示的 R 代码求得检验临界值。

```
> qt(0.025, 7); qt(0.975, 7)
[1] -2.364624
[1] 2.364624
```

因为 $-2.365\leqslant-0.9019\leqslant2.365$，所以检验统计量落在了接受域中。更进一步，还可以

在 R 中使用下面的代码算得与检验统计量相对应的 P 值：

```
> pt(-0.9019, 7, lower.tail = T) * 2
[1] 0.3970802
```

因为 P 值=0.397，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。以上计算结果与本章前面由 t.test()函数所得之结果是完全一致的。

对于独立样本数据而言，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：

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

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，可以使用如下的 R 代码求得检验临界值：

```
> qt(0.025, 4.503); qt(0.975, 4.503)
[1] -2.658308
[1] 2.658308
```

因为 $-2.658 \leqslant -0.9357 \leqslant 2.658$，所以检验统计量落在接受域中。更进一步，还可以在 R 中使用下面的代码算得与检验统计量相对应的 P 值：

```
> pt(-0.9357, 4.503, lower.tail = T) * 2
[1] 0.3968415
```

因为 P 值=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。以上计算结果与本章前面由 t.test()函数所得结果是完全一致的。

最后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 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

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

这里 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：

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

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，可以使用如下所示的 R 代码求得检验临界值：

```
> qt(0.025, 5); qt(0.975, 5)
[1] -2.570582
[1] 2.570582
```

因为 3.235≥2.571，所以检验统计量落在了拒绝域中。更进一步，还可以在 R 中使用下面的代码算得与检验统计量相对应的 P 值：

```
> 2 * (pt(3.2359, 5, lower.tail = F))
[1] 0.02305406
```

因为 P 值=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。以上计算结果与本章前面由 t.test()函数所得之结果是完全一致的。

3.4 最大似然估计

正如本章最初所讲的，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。

3.4.1 最大似然法的基本原理

最大似然这个思想最初是由德国著名数学家卡尔·高斯（Carl Gauss）提出的，但真正将其发扬光大的则是英国的统计学家罗纳德·费希尔（Ronald Fisher）。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。假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

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

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。我们需要通过采样结果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 的值。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。

最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 的一切可能取值中选出一个使样本观测值出现的概率为最大的 θ 值，记为 $\hat{\theta}$，并将此作为 θ 的估计，并称 $\hat{\theta}$ 为 θ 的最大似然估计。

首先考虑 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

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

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

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

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 $\mathcal{L}(\theta)$ 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 $\mathcal{L}(\theta)$ 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

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

如果 $\hat{\theta}$ 存在，则称 $\hat{\theta}$ 为 θ 的最大似然估计。

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

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

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 的最大似然估计 $\hat{\theta}$，从而满足

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

3.4.2 求最大似然估计的方法

当函数关于参数可导时，可以通过求导方法获得似然函数极大值对应的参数值。在求最大似然估计时，为求导方便，常对似然函数$\mathcal{L}(\theta)$取对数，称 $l(\theta)=\ln\mathcal{L}(\theta)$为对数似然函数，它与$\mathcal{L}(\theta)$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

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

为似然方程，其中 k 是θ 的维数。

下面结合一个例子演示这个过程。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 的最大似然估计。首先写出似然函数

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

然后对上式左右两边取对数，可得

$$\begin{aligned} l(p)&=\sum_{i=1}^{n}\left[x_i\ln p+(1-x_i)\ln(1-p)\right]\\ &=n\ln(1-p)+\sum_{i=1}^{n}x_i\left[\ln p-\ln(1-p)\right]\end{aligned}$$

将 $l(p)$对 p 求导，并令其导数等于 0，得似然方程

$$\begin{aligned}\frac{\mathrm{d}l(p)}{\mathrm{d}p}&=-\frac{n}{1-p}+\sum_{i=1}^{n}x_i\left(\frac{1}{p}+\frac{1}{1-p}\right)\\ &=-\frac{n}{1-p}+\frac{1}{p(1-p)}\sum_{i=1}^{n}x_i\\ &=0\end{aligned}$$

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 的最大似然估计为 $\hat{p}=\bar{x}=T/n$。

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 的最大似然估计值。首先写出似然函数

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

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，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

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

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 的最大似然估计分别为

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

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，可见参数的最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 的最大似然估计，$g(\theta)$ 是 θ 的连续函数，则 $g(\theta)$ 的最大似然估计为 $g(\hat{\theta})$。

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 $E(X)$ 的最大似然估计值。首先写出似然函数

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

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

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

解似然方程得

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

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 的最大似然估计值为 $\hat{\lambda}=1/\overline{X}$。此外，$E(X)=1/\lambda$，它是 λ 的函数，其最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。

3.4.3 最大似然估计应用举例

3.4.2 节演示了通过解方程 $\partial l(\theta)/\partial\theta_j=0$ 从而求得参数 θ 的最大似然估计值的基本方法。显而易见的是，这个求解过程非常复杂，本节将通过几个实例演示在 R 中进行最大似然估计的方法。

对于不同的分布而言，其似然函数的形式也是各式各样的，所以最后得到的似然方程解（也即是参数的最大似然估计值）的表达式也很难统一。所以很难找到一种通用的方法对所有情况下的参数做最大似然估计。因此，使用 R 语言进行最大似然估计，往往是先要确定似然函数的表达式，然后再借助 R 中的极值求解函数完成。

在单参数情况下，可以使用 R 中的函数 optimize()求最大似然估计值，它的调用格式如下：

```
optimize(f, interval, lower = min(interval),
        upper = max(interval), maximum = FALSE)
```

函数 optimize()的作用是在由参数 interval 指定的区间内搜索函数 f 的极值。这个区间也可以由参数 lower（即区间的下界）和 upper（即区间的上界）控制。默认情况下，参数 maximum 为 FALSE 表示求极小值；如果将其置为 TRUE，则表示求极大值。

例如，现在已知某批电子元件的使用寿命服从参数为 λ 的指数分布，λ 未知且有 $\lambda>0$。现在，随机抽取一组样本并测得其使用寿命如下（单位：小时）：

518　612　713　388　434

试尝试用最大似然估计其这批产品的平均寿命。

前面已经求出了指数函数的对数似然函数形式，可以用 R 语言代码将似然函数写为

```
> f <- function(lamda){
    logL = n * log(lamda) - lamda * sum(x)
    return (logL)
}
```

然后用 optimize()求使得似然函数取得极值时的参数 λ 的估计值，结果如下：

```
> x = c(518,612,713,388,434)
> n = length(x)
> duration <- optimize(f, c(0,1), maximum = TRUE)
> duration
 $ maximum
```

```
[1] 0.001878689
 $ objective
[1] -36.39261
```

由此便求出了参数λ的估计值为0.001 878 689，再根据3.4.2节最后得出的结论，可知这批电子元件的平均使用寿命$E(X)=1/\lambda$，即有

```
> 1/duration $ maximum
[1] 532.2862
```

与之前推导的结论相比，当X服从指数分布时，$E(X)$的最大似然估计为$\overline{X}$，因此算得的结果是一致的。

再来看一个稍微复杂的例子，这次要估计的参数将有多个。首先在R中导入程序包MASS中的数据geyser，示例代码如下：

```
> library(MASS)
> attach(geyser)
```

该数据集是地质学家记录的美国黄石公园内“忠实泉”(Old Faithful)一年内的喷发数据，数据有两个变量，分别是泉水持续涌出的时间(eruptions)和喷发相隔的时间(waiting)，在这个例子中我们将仅会用到后者。“忠实泉”(如图3-6所示)是黄石公园内的著名景观，也是世界上最容易被预测的地理特征。

现在打算对变量waiting的分布进行拟合，于是首先通过直方图大致了解数据的分布形态。执行下面的代码，其结果如图3-7所示。

图3-6　黄石公园中的“忠实泉”

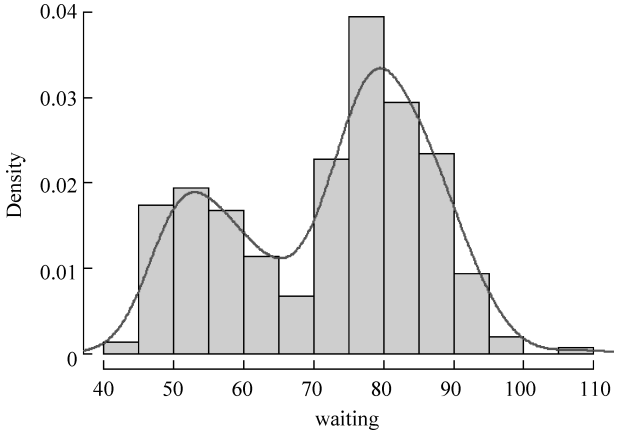

图3-7　数据分布的直方图

```
> hist(waiting, freq = FALSE, col = "wheat")
> lines(density(waiting), col = 'red', lwd = 2)
```

从绘制的结果看，图形中有两个峰，很像是两个分布叠加在一起而成的结果，于是可以推断分布是两个正态分布的混合，故用下面的函数描述

$$p(x)=\alpha N(x;\mu_1,\sigma_1)+(1-\alpha)N(x;\mu_2,\sigma_2)$$

所以在构建的模型中，需要估计的参数有 5 个，即 α、μ_1、σ_1、μ_2 和 σ_2。上述分布函数的对数最大似然函数为

$$l=\sum_{i=1}^{n}\log p(x)$$

接下来，在 R 中定义对数似然函数，示例代码如下。由于在后面将要使用的极值求解法会在迭代过程中产生一些似然函数不能处理的无效值，尽管这并不会影响到最终的求解结果，但是为了避免出现不必要的警告信息，此处使用了 suppressWarnings()函数忽略那些警告信息。

```
> LL <- function(params,data){
        t1 <- suppressWarnings(dnorm(data,params[2],params[3]))
        t2 <- suppressWarnings(dnorm(data,params[4],params[5]))
        ll <- sum(log(params[1] * t1 + (1 - params[1]) * t2))
        return(ll)
}
```

为了进行最大似然估计，下面将调用 R 语言中的程序包 maxLik，该包为进行最大似然估计提供了诸多便利，在多参数估计时可以考虑使用它。在进行极值求解时，可以通过修改 maxLik()函数中的参数 method 选择不同的数值求解方法。可选的值有 NR、BHHH、BFGS、NM 和 SANN 5 种，默认情况下函数将使用默认值 NR，即采用 Newton-Raphson 算法。

```
> library("maxLik")
> mle <- maxLik(logLik = LL, start = c(0.5,50,10,80,10), data = waiting)
> mle
Maximum Likelihood estimation
Newton - Raphson maximisation, 8 iterations
Return code 2: successive function values within tolerance limit
Log - Likelihood: - 1157.542 (5 free parameter(s))
Estimate(s): 0.3075935 54.20265 4.951998 80.36031 7.507638
```

最后，通过图形评估采用最大似然法所估计出来的参数拟合效果，示例代码如下：

```
> a <- mle $ estimate[1]
> mu1 <- mle $ estimate[2]; s1 <- mle $ estimate[3]
> mu2 <- mle $ estimate[4]; s2 <- mle $ estimate[5]
> X <- seq(40,120,length = 100)
> f <- a * dnorm(X,mu1,s1) + (1 - a) * dnorm(X,mu2,s2)

> hist(waiting, freq = FALSE, col = "wheat")
> lines(density(waiting), col = 'red', lty = 2)
```

```
> lines(X, f, col = "blue")

> text.legend = c("Density Line","Max Likelihood")
> legend("topright", legend = text.legend, lty = c(2,1),
 + col = c("red","blue"))
```

执行以上代码，结果如图 3-8 所示，其中实线是基于估计参数绘制的数据分布曲线，虚线是系统自动生成的密度曲线，可见拟合效果还是比较理想的。

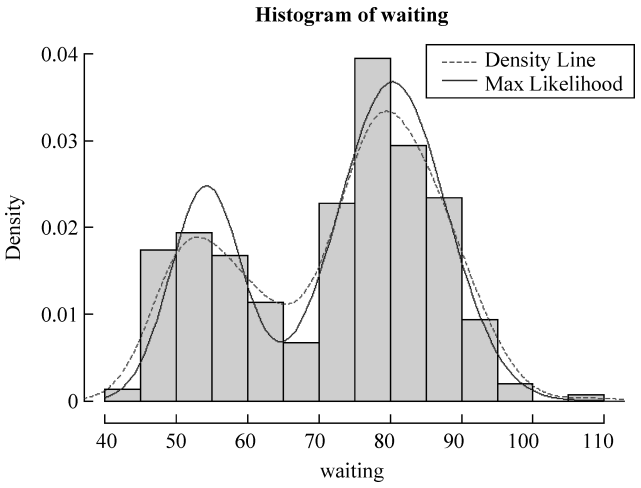

图 3-8　最大似然法拟合效果