第五章 参数估计与假设检验

章节引入

这一章我们将介绍统计分析中的两个重要方法：参数估计与假设检验。在实际业务场景中，我们在希望得到关于某个业务问题的结论时，往往需要收集一些数据进行分析，并期望得到尽可能精确的结论。为什么这里说是“尽可能精确”呢？这是由于我们采集到的数据往往带有随机性，抑或存在采样的偏差等。例如，在研究灯泡的平均寿命时，我们不可能采集到世界上所有灯泡的数据，这时候只能随机抽取几十个灯泡，观察它们的损耗情况。如何依据所采集到的灯泡样本数据对灯泡总体的寿命做出估计、推断，是本章主要探讨的问题。

下面举一个简单的例子帮助我们理解本章所要研究的具体问题。假设 2050 年，某手机公司发布新一代手机 BearPhone。该手机声称续航能力比上一版本（20 小时）有较大提升。为调查市面上售卖的手机是否满足这样的续航能力，我们收集并测试了 50 部手机的续航时间。假设手机的续航时间服从指数分布，那么我们希望得知：

（1） 该手机的平均续航时间是多久？

（2） 这批手机是否续航时间超过 20 小时？

以上两个问题是本章希望回答的两个典型问题。首先，“手机的续航时间服从指数分布”是本问题的一个核心的模型假设。对于参数为 λ 的指数分布，可知其均值为 $1/\lambda$。因此，如果得知 λ 的数值，则问题（1）即可迎刃而解。然而，现实中我们往往无法得知λ的真实值。这就需要我们通过样本（例子中的 50 部手机）的取值对总体（所有手机）的平均值进行一个估计。这个估计的统计性质如何，估计产生的误差有多大？这是**参数估计**（parameter estimation）部分需要回答的问题。

接下来，问题（2）的回答则属于假设检验的范畴。我们可以采取一个非常严格的准则：如果这批收集到的样本均值 $\bar{x} \geq 20$，则认为这批手机达到了所声称的标准，否则认为没有达到。这个标准十分严格，但不一定是对问题（2）的最合适的解答。正如前面参数估计中提到，通过样本均值估计总体均值存在一定的

误差。在这种误差存在的前提下，可以对以上准则进行调整：当 $\bar{x} \geq l$ 时，认为达到了公司声称的续航时间。这又产生了新问题：l应该取什么值比较合适？这是**假设检验**需要回答的问题。

案例引入

本章主要包括三个简短的案例，以下对每个案例及数据进行简要介绍。

案例一：手机续航时间研究

背景介绍

手机续航能力是指手机在正常工作时的待机时间。它与手机本身的耗电量和所配电池容量的大小有关。在电池容量相同的情况下，耗电量越小，则续航能力越强；在耗电量相同的情况下，电池容量越大，待机时间越长。近年来，各个品牌厂家都致力于打造续航能力更强的新款手机，从而满足现代人们对手机设备的长时间使用需求，提高自身的市场竞争力。

数据介绍

该案例使用的数据是 50 部手机 BearPhone 的续航时间，单位为小时。

```
# 显示续航时间样本数据
print(battery)
```

```
##  [1] 20.2 19.4 18.6 18.2 19.8 21.4 18.2 19.8 17.4 19.4 21.4 19.0 19.0 17.0
## [15] 23.0 22.6 21.4 20.2 22.6 20.2 21.4 20.2 18.2 20.2 19.0 19.8 19.8 20.6
## [29] 20.6 18.6 19.0 21.4 20.6 21.8 19.4 19.8 21.8 19.4 17.4 20.2 19.8 17.8
## [43] 17.4 19.8 20.2 19.0 20.2 18.2 21.0 19.8
```

本案例研究的问题是判断这一批手机的续航时间是否达到了声称的 20 小时。

案例 2：减肥药疗效研究

背景介绍

随着生活水平的提高，油腻、高热量的食物成为了很多年轻人的便捷之选，肥胖患者日剧增多。然而由此引发的肥胖问题会极大危害到身体健康，仅在 2015

年，全球因肥胖造成死亡的人数就已超过 400 万，世界卫生组织已将超重、肥胖定义为一种慢性病。因此，减肥不仅是一个时尚与外貌相关的话题，更关乎身体的健康。为了应对肥胖带来的身体健康问题，制药公司纷纷推出减肥特效药。减肥药药效如何，需要大量试验验证，得出结论。

数据介绍

在减肥的案例中，将会使用以下减肥药药效的数据。一个合规的减肥药要在美国市场上合法上市，必须有美国食品药品管理局（FDA）的批准。为此，FDA 对使用该药物的受试者做了一组试验，记录了其在使用减肥药前后的体重（单位：kg）。试验的数据如下所示：

受试者编号	1	2	3	4	5	6	7	8	9	10
服用药物前的体重x	50	59	55	60	58	54	56	53	61	51
服用药物后的体重y	43	55	49	62	59	49	57	54	55	48
差 $d = x - y$	7	4	6	-2	-1	5	-1	-1	6	3

受试者在服用减肥药后是否体重发生了显著下降，这是本章所要探究的问题。

案例 3：红楼梦作者争议

背景介绍

《红楼梦》是一部章回体长篇小说，被列为中国古代四大名著之首。《红楼梦》文学价值极高，围绕作品本身的争议也很多。其中一大争议便是作者归属问题。

红楼梦共 120 章回，前 80 回比较公认的作者是曹雪芹。他自述“批阅十载，增删五次”，方成此书。有学者认为，红楼梦整个故事的发展，正是曹雪芹家族的镜像。红楼梦后 40 回原作散失，至今作者归属仍是谜团，各学派争论不一。1920 年，胡适先生“大胆假设”，认为后四十回并非曹雪芹所著，而是高鹗续书。周汝昌认为《红楼梦》共 108 回，现存 80 回，后 28 回遗失。白先勇认为，没有人能续作红楼梦，后四十回中作者笔触细腻，前后呼应，一百二十回应全系曹雪芹所做。

众多大家各执一词，学术界仍无定论。本章从文本分析的角度，分析作者用语习惯的改变，探索《红楼梦》作者归属的问题。

数据介绍

本案例收集了《红楼梦》中有代表性的若干个文言虚词，并希望探究这些虚词在《红楼梦》前后的使用差异。在本章中，我们选择具有代表性的“之”和“亦”两个虚词的章回词频，通过对比词频在整本书前后出现的差异，可以了解作者用语习惯的改变。具体而言，统计每个虚词在每个章回的总词频，将全书 1-40 回、41-80 回、81-120 回切分为三个总体。全书 1-40 回、41-80 回、81-120 回中两个字的平均字频如图 5.1 所示。

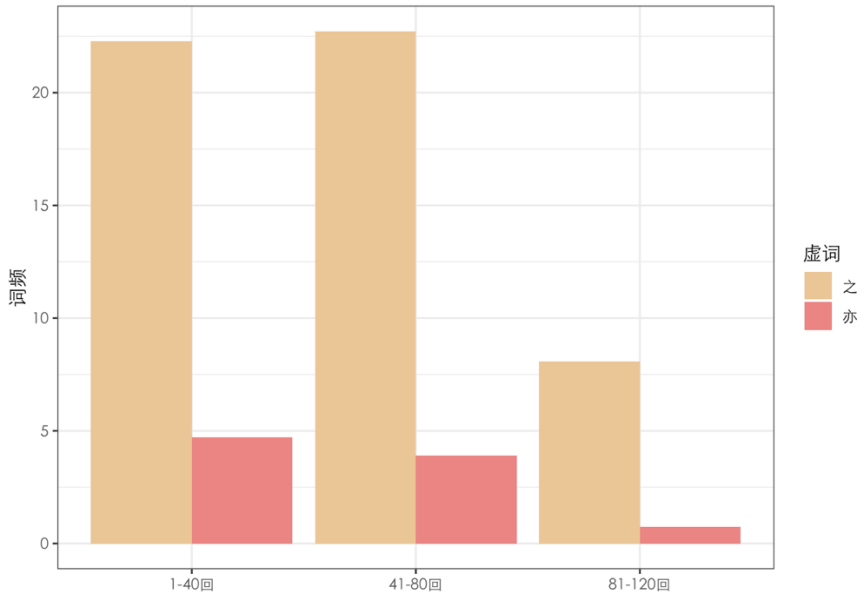

图 5.1：《红楼梦》中虚词“之”、“亦”使用频数柱状图

从柱状图可以看到，这两个文言虚词在前 80 回变化不大，而在后 40 回使用频率大大降低。这说明作者的语言风格在后 40 回更加偏向白话。那么这种差异是否是显著的呢？这是我们要探索的问题。

本章难点

（1）理解并掌握不同的参数估计方法，包括矩估计、极大似然估计和区间估计；熟练掌握常见的参数估计方法在实际案例中的应用，并使用 R 语言实现。

（2）理解假设检验的基本思想、步骤、p 值的概念，熟练掌握假设检验问题的

基本类型，并使用 R 语言实现。

（3）了解单因素方差分析的基本思想和解读。

5.1 总体、样本和样本量

在学习具体方法之前，我们先介绍一些需要用到的概念。理解这些概念有助于理解本章所学习的主要内容的原理。

5.1.1 总体

总体是指所研究问题的对象的全体。注意到，这里“总体”随着所研究问题的不同而不同。例如，案例一中希望研究手机 BearPhone 的续航时间，此时，所有型号为 BearPhone 的手机就构成了总体。假如希望研究中国区域内手机 BearPhone 的续航时间，那么，总体就变成了“所有中国区内 BearPhone 手机”。注意到这里限定了手机的某种特性：续航时间，而不是手机的重量、大小等。因此，这里所说的总体 “手机 BearPhone 的续航时间”构成了一组数值。

这些数值在分布上可能存在一定的规律特征。在对总体进行研究时，往往赋予其一定概率分布，使得可以通过概率论等工具对其进行研究。对于总体的数目有两种认知方式，一种是认为总体数目有限，此时可以采用一些离散分布描述总体的分布情况。另外一种认为总体是无限的。比如上面例子提到的 BearPhone 手机的续航时间，可以认为总体既包含目前市面上已经生产售卖的手机，也包括正在生产和即将生产的手机。“无限总体”这个概念由著名统计学家 Ronald Fisher 引进。这种认知方式可以使得我们用连续分布刻画总体特征。比如，可以认为手机的续航时间服从指数分布。从现实看，这种近似的分布形式往往能很好地表述数据的规律，相关的近似误差在实际应用中也可以被忽略。

5.1.2 样本

样本是按照一定规律在总体中抽出的一部分个体。在案例一中，如果我们要研究所有市面上 BearPhone 手机的续航时间，则可能需要收集所有的 BearPhone 手机。需要投入大量的人力、物力，这显然是不可行的。因此，我们常常通过样本来推断总体的特征。在案例一中，我们抽出的 50 部手机的续航时间就是总体

的一组样本，这里的 50 就是样本量（或称作样本容量）。

从总体中抽取样本的方式不能是随意的。比如希望了解所有 BearPhone 手机的续航情况，但抽取的样本仅来自于中国地区，则这些样本并不能很好地代表全部总体的特征。又如在总统大选中，希望通过问卷的形式调查民意，但如果调查在互联网上进行，那么收集到的信息只能代表一部分愿意使用互联网的年轻选民的，从而导致对整体选票情况认知的偏差。因此，一般需要从总体中以简单随机抽样的形式采集样本，这种抽样方式要求每个个体有同等机会被选入样本，样本之间相互独立，分布相同。因此，这样抽取的一组样本可以较好的代表总体的性质。

5.1.3 统计量

统计量是样本的函数，其取值完全依赖于样本。什么叫完全依赖于样本呢？在案例一中，我们计算了样本均值：$\bar{x} = 19.824$ 。这里的样本均值就是关于样本的一个函数，同时其取值不依赖于其他未知数值，因此是一个统计量。但是，如果假设总体的均值为 μ，那么 $\bar{x} - \mu$ 就不是一个统计量。这是因为这里的总体均值 μ 未知，所以不能说“取值完全依赖于样本”。

为什么要有统计量这样一个概念呢？统计量可以看作是对样本的加工，目的是刻画总体的特征。简单来说，每个样本抽样来自于总体，因此也携带了总体的“基因”。例如，可以认为每个样本都含有总体均值 μ 的信息，通过样本均值的形式进行汇总后，可以更加明确地反映出总体均值的性质。第三章中介绍的样本方差、标准差等都是统计量。

5.2 参数估计

假设总体的分布由分布函数 $F(x;\theta)$ 刻画，其中包含未知参数 θ。例如，正态总体 $N(\mu,\sigma^2)$ 包含两个未知参数 μ 和 σ^2。应该如何估计这些未知参数的值呢？这就是参数估计所要解决的问题。不过，参数估计所讨论的范畴不仅限于对未知参数的估计，还包括总体分布的各种特征的估计，比如总体均值、方差、高阶矩等。参数估计的主要形式有两种：点估计与区间估计。点估计要构造一个统计量 $\hat{\theta} = \hat{\theta}(x_1,\cdots,x_n)$。将样本取值代入后则得到一个点估计值。可以认为这里

的未知参数是一个点，通过另外一个点 $\hat{\theta}$ 去估计它，则称为点估计。区间估计与点估计不同，它通过构造一个区间 $[\hat{\theta}_L, \hat{\theta}_U]$ 给出未知参数 θ 的估计。上述区间以较高的概率（例如，95%）包含真实参数 θ，这个概率被称为置信水平。

5.2.1 矩估计

1. 定义

矩估计（method of moments）是点估计的重要方式之一。矩估计的思想由著名统计学家 K. Pearson 提出，其基本想法是通过样本矩去估计总体矩。

具体而言，设 $x_1, \cdots, x_n$ 是来自总体 X 的一个样本。总体的 k 阶矩为：$\mu_k = E(X^k)$。样本的 k 阶原点矩可以定义为：

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

那么矩估计的想法就是通过样本 k 阶原点矩去估计总体 k 阶矩。

2. 实例分析

设 $x_1, \cdots, x_n$ 是从正态总体 $N(\mu, \sigma^2)$ 中抽出的一个样本。其中，μ 是总体均值，也就是总体一阶原点矩。σ^2 是总体方差，也就是总体二阶原点矩。根据矩估计的方法，可以用样本一阶矩（样本均值）以及样本二阶矩分别对其进行估计。其中，在估计总体方差时，往往做一个小的修正，即采用样本方差对其进行估计。我们通过一个随机模拟实验模拟这种估计效果。

```
set.seed(123)  # 设置随机数种子
data1 <- rnorm(n = 1000, mean = 5, sd = 2)  # 产生服从正态分布的随机数
mean(data1)  # 样本均值
## [1] 5.032256
var(data1)  # 样本方差
## [1] 3.933836
```

以 1000 个服从$N(5,4)$的随机数为样本，求它的一阶原点矩（均值）和二阶中心矩（方差），作为总体均值和方差的估计。结果显示，对总体均值的估计为 $\hat{\mu} =$

$\frac{\Sigma_{i=1}^{n} x_i}{n} = 5.03$，总体方差的估计为 $\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 3.93$。

上述示例给出了一个简单的例子，其中未知参数 μ 和 σ^2 的值都可以通过样本一阶矩、二阶矩获得。但并不是所有分布的未知参数都可以以这种方式获得。一般来说，假设总体 X 的分布函数含有 k 个未知参数 $\theta_1, \cdots, \theta_k$，且假设总体分布的前 k 阶矩存在，此时可以通过联立方程组的形式求得 θ_j 的估计。设 $\mu_j = g_j(\theta_1, \cdots, \theta_k)$，则可以得到 k 个方程 $(j = 1, \cdots, k)$。在方程可以求解的前提下，设求解得 $\theta_j = h_j(\mu_1, \cdots, \mu_k)$，带入样本矩的估计值 $\hat{\mu}_1, \cdots, \hat{\mu}_k$，则可以求得未知参数的估计值：$\hat{\theta}_j = h_j(\hat{\mu}_1, \cdots, \hat{\mu}_k)$，$j = 1, \cdots, k$。

5.2.2 极大似然估计

极大似然估计（maximum likelihood estimation，MLE）是本章将要介绍的第二种重要的点估计方式。当总体参数类型已知（例如，已知是正态分布）时，则可以选用极大似然估计。

1. 定义

以下我们简述极大似然估计的原理。假设总体X的分布是连续的，其概率密度函数为$f(x, \theta)$（当总体X分布离散时，可以替换为概率函数）。当样本 $x_1, \cdots, x_n$ 满足独立同分布假设时，可以写出样本的联合密度函数：

$$L(x_1, \cdots, x_n; \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta)$$

以上联合密度函数可以作为一组样本$(x_1, \cdots, x_n)$出现的可能性。假设已经观察到$(x_1, \cdots, x_n)$的取值，则应寻找θ的值，使得这组观测到的样本出现的可能性尽量大。当$x_1, \cdots, x_n$代入样本的观察值时，以上联合密度可以看成是θ的函数，称为似然函数，记为L(θ)。极大似然估计就是使得似然函数最大时的估计值 $\hat{\theta} = \arg\max L(\theta)$。极大似然估计法自提出以来，其性质被广泛研究，一般来说，当模型参数形式已知时，极大似然估计具备优良的性质。

2. 求解

如何求解最大似然估计呢？当函数$f(\cdot)$对θ可导时，可以通过求导法求解似然函数的最大值。由于在似然函数中$f(x_i,\theta)$是相乘的形式，并不易于直接求导，一般可以将似然函数转换为对数似然函数：$\log L(\theta) = \sum_{i=1}^{n}\log(f(x_i;\theta))$。注意到这里的log (·)变换是一个单调变换，因此求解极大似然估计可以转换为求解对数极大似然函数的最大值。同时，log (·)变换将密度函数相乘形式转换为相加形式，更加易于求导求解。具体地，通过求解以下方程可以求得最大似然的估计值。

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n}\frac{\partial \log(f(x_i;\theta))}{\partial \theta} = 0$$

从最大似然估计的定义可以看出，若$L(\theta)$与联合密度函数相差一个与θ无关的比例因子，不会影响最大似然估计，因此，可以在$L(\theta)$中剔去与θ无关的因子。

但是，如果函数$f(\cdot)$对θ不可导，甚至$f(\cdot)$本身也不连续时，无法使用求导进行求解。此时仍需要通过最大化似然函数求解。

最大似然估计有一个简单而有用的性质：如果 $\hat{\theta}$是θ的最大似然估计，则对任一函数$g(\theta)$，其最大似然估计为$g(\hat{\theta})$。这个性质称为最大似然估计的**不变性**，这一性质使得一些复杂结构的参数的最大似然估计更易求解。

3. 实例分析

以下介绍两个在 R 中求解 MLE 的具体步骤。

（1）正态分布的极大似然估计

假设新一代手机 BearPhone 续航时间是服从正态分布$N(\mu,\sigma^2)$，其中，μ 是总体均值，σ^2 是总体方差。请根据案例数据中的 50 个续航时间样本使用最大似然估计的方法，估计总体的均值和方差。

正态分布的似然函数为：

$$L(\mu,\sigma^2) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(x_i-\mu)^2\right\}$$

求解对数似然函数并求导，令导数等于 0，可得：

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

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

$$\frac{\partial lnL(\mu,\sigma^2)}{\partial\sigma^2}=\frac{1}{2\sigma^4}\Sigma_{i=1}^{n}(x_i-\mu)^2-\frac{n}{2\sigma^2}=0$$

因此μ的极大似然估计是$\bar{x}$。进一步代入$\hat{\mu}=\bar{x}$得到：

$$\hat{\sigma}^2=\frac{1}{n}\Sigma_{i=1}^{n}(x_i-\bar{x})^2=s^2$$

由推导结果可知，正态总体均值和方差的最大似然估计和矩估计相同，R 代码实现如下：

```
cat('手机电池续航数据为：', battery, '\n')
```

```
## 手机电池续航数据为： 20.2 19.4 18.6 18.2 19.8 21.4 18.2 19.8 17.4 19.4 21.4 19
19 17 23 22.6 21.4 20.2 22.6 20.2 21.4 20.2 18.2 20.2 19 19.8 19.8 20.6 20.6 1
8.6 19 21.4 20.6 21.8 19.4 19.8 21.8 19.4 17.4 20.2 19.8 17.8 17.4 19.8 20.2 19
 20.2 18.2 21 19.8
```

```
# 续航时间均值的极大似然估计

mu_mle <- mean(battery)

# 续航时间方差的极大似然估计

sigma_mle <- mean((battery - mean(battery)) ** 2)

cat('均值的最大似然估计值是：', mu_mle,

    '方差的最大似然估计值是：', sigma_mle, '\n')
```

```
## 均值的最大似然估计值是： 19.824 方差的最大似然估计值是： 1.948224
```

（2）均匀分布的极大似然估计

虽然求导函数是求最大似然估计最常用的方法，但并不是在所有场合求导都是有效的，比如似然函数不可导的情况下需要采取其他方式求解最大似然估计，下面的例子说明了这个问题。

设 $x_1,\cdots,x_n$ 是从均匀分布总体 $U(0,\theta)$ 中抽出的一个样本，观测值如下：6.0627 9.3764 2.6435 3.8009 8.0748 9.7808 9.5793 7.6273 5.0965 0.6448 6.4357 9.1591 0.9523 2.9537 7.6993 2.5589 5.1790 6.7785 1.4723 7.0053 试求解θ的最大似然估计。

似然函数：

$$L(\theta)=\frac{1}{\theta^n}\prod_{i=1}^{n}I[0<x_i\le\theta]=\frac{1}{\theta^n}I[x_{(n)}\le\theta]$$

要使得$L(\theta)$达到最大，首先一点是示性函数的取值应该为 1，其次是$1/\theta^n$尽可能大。由于$1/\theta^n$是θ的单调减函数，所以θ的取值应尽可能小，同时示性函数为 1 决定了$\theta\ge x_{(n)}$，这里$x_{(n)}$代表样本观测的第n个次序统计量。由此给出θ的最大似然估计 $\hat{\theta}=x_{(n)}$。

R 代码实现如下：

```
# theta 的最大似然估计值是

theta_mle = max(samples)

cat('theta 的最大似然估计值是：', theta_mle, '\n')
```

```
## theta 的最大似然估计值是： 9.7808
```

5.2.3 区间估计

1. 基本概念

点估计给出了未知参数的一个“点”的估计，但是未能概括估计的精度。在案例一中，尽管可以通过样本均值的方式得到总体 BearPhone 手机的续航时间均值的估计，但并不知道这种估计的误差是多少，此时区间估计很好地弥补了这种不足。假如我们通过估计得到，手机的续航时间以极大的可能性位于区间$[18.5,21]$内，则这个区间便可以用来作为一种描述误差的形式。

设θ是总体的一个参数，$x_1,\cdots,x_n$是样本，所谓区间估计就是要找到两个统计量 $\hat{\theta}_L(x_1,\cdots,x_n)$和$\hat{\theta}_U(x_1,\cdots,x_n)$，使得 $\hat{\theta}_L<\hat{\theta}_U$。在得到样本的观测值后，则可以得到区间估计$[\hat{\theta}_L,\hat{\theta}_U]$。区间估计的长度越小，则越精确。由于样本的随机性，区间$[\hat{\theta}_L,\hat{\theta}_U]$盖住未知参数$\theta$的可能性并不确定，人们通常要求区间$[\hat{\theta}_L,\hat{\theta}_U]$盖住$\theta$的概率$P(\hat{\theta}_L\le\theta\le\hat{\theta}_U)$尽可能大，但这必然会导致区间长度增大。为了平衡这种矛盾，Neyman 建议采取一种折中的方案：把区间$[\hat{\theta}_L,\hat{\theta}_U]$盖住$\theta$的概率（也称为

置信水平）事先给定，寻找区间长度尽量小的区间估计。以下给出置信区间的概念。

【定义 5.2.1】设θ是总体的一个参数，其参数空间为Θ，$x_1,\cdots,x_n$是来自该总体的样本，对给定的一个置信水平$\alpha\ (0<\alpha<1)$，假设有两个统计量$\hat{\theta}_L=\hat{\theta}_L(x_1,\cdots,x_n)$和$\hat{\theta}_U=\hat{\theta}_U(x_1,\cdots,x_n)$，若对于任意的$\theta\in\Theta$，有$P\left(\hat{\theta}_L\le\theta\le\hat{\theta}_U\right)\ge 1-\alpha$，则称随机区间$[\hat{\theta}_L,\hat{\theta}_U]$为$\theta$的置信水平为$1-\alpha$的置信区间（confidence interval），或简称$[\hat{\theta}_L,\hat{\theta}_U]$是$\theta$的$1-\alpha$置信区间，$\hat{\theta}_L$和$\hat{\theta}_U$分别称为$\theta$的（双侧）置信下限和置信上限。

在实际数据分析中，一般取$\alpha=0.05$。当然，你也可以选择α为其它较小的数值。可以这样理解$1-\alpha$置信区间：通过样本设法构造出置信区间$[\theta_L,\theta_U]$，使得该区间覆盖真实的未知参数θ的概率为$1-\alpha$。可以设想，当样本取值不同时，构造出的区间也有所不同。如果进行 100 次抽样，对每次抽样的样本以相同的方式进行置信区间的估计，则约有 $100(1-\alpha)$次构造的区间包含真实参数，而有约100α次区间估计没有包含真实参数。那么，给定一组数据，应该如何构造置信区间呢？下面将给出具体估计方法。

2. 枢轴量法

构造未知参数θ的置信区间的一种常用方法是枢轴量法，它的具体步骤是：

（1）从θ的一个点估计$\hat{\theta}$出发，构造$\hat{\theta}$与θ的一个函数$G\left(\hat{\theta},\ \theta\right)$，使得$G$的分布是已知的，而且与$\theta$无关。通常称这种函数$G\left(\hat{\theta},\ \theta\right)$为**枢轴量**。一般选择常用的已知分布作为$G$的分布，例如标准正态分布$N(0,1)$、$t$分布等。

（2）适当选取两个数 c 和 d，使得对于给定的α有：

$$P\left(c\le G\left(\hat{\theta},\theta\right)\le d\right)\ge 1-\alpha \qquad (5.2.1)$$

这里概率的大于等于号是专门为离散分布而设置的，当$G\left(\hat{\theta},\theta\right)$的分布是连续分布时，应选 c 和 d 使得式（5.2.1）中的概率等于$1-\alpha$，这样就能充分地使用置信水平$1-\alpha$，并获得同等置信区间。

（3）利用不等式运算，将不等式$c \le G(\hat{\theta},\theta) \le d$等价变形，最后得到形如$\hat{\theta}_L \le \theta \le \hat{\theta}_U$的不等式。完成以上步骤后，$[\hat{\theta}_L,\hat{\theta}_U]$就是$\theta$的$1-\alpha$置信区间。因为此时有：

$$P(\hat{\theta}_L \le \theta \le \hat{\theta}_U) = P(c \le G(\hat{\theta},\theta) \le d) \ge 1-\alpha \qquad (5.2.2)$$

满足上述条件的c和d可以有很多，选择的目的是希望（5.2.1）中的区间平均长度$E_\theta(\hat{\theta}_U - \hat{\theta}_L)$尽可能短。假如可以找到这样的$c$和$d$当然是最好的，但在不少场合很难做到这一点。因此常选择c和d，使得两个尾部概率各为$\alpha/2$，即：

$$P_\theta(G(\hat{\theta},\theta) < c) = P_\theta(G(\hat{\theta},\theta) > d) = \alpha/2 \qquad (5.2.3)$$

这样得到的置信区间称为**等尾置信区间**，实用的置信区间多为等尾置信区间。

3. 正态总体均值的置信区间

（1）单个正态总体均值的置信区间

正态总体$N(\mu,\sigma^2)$是最常见的分布，我们首先来讨论σ已知时μ的置信区间。在σ已知的情况下，由于μ的点估计为$\bar{x}$，其分布为$N(\mu,\frac{\sigma^2}{n})$，因此枢轴量可以选为$G = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$，c 和 d 应满足$P(c \le G \le d) = \Phi(d) - \Phi(c) = 1-\alpha$，经过不等式变形可得：

$$P_\mu\left(\bar{x} - \frac{d\sigma}{\sqrt{n}} \le \mu \le \bar{x} - \frac{c\sigma}{\sqrt{n}}\right) = 1-\alpha$$

该区间长度为$\frac{(d-c)\sigma}{\sqrt{n}}$，由于标准正态分布为单峰对称的，由图 5.2 可以看出，在$\Phi(d) - \Phi(c) = 1-\alpha$的条件下，当$d = -c = z_{1-\frac{\alpha}{2}}$时，$d-c$达到最小，其中$z$服从标准正态分布，其中$\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$，由此给出了$\mu$的$1-\alpha$同等置信区间为：

$$\left[\bar{x} - \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{x} + \frac{z_{1-\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$$

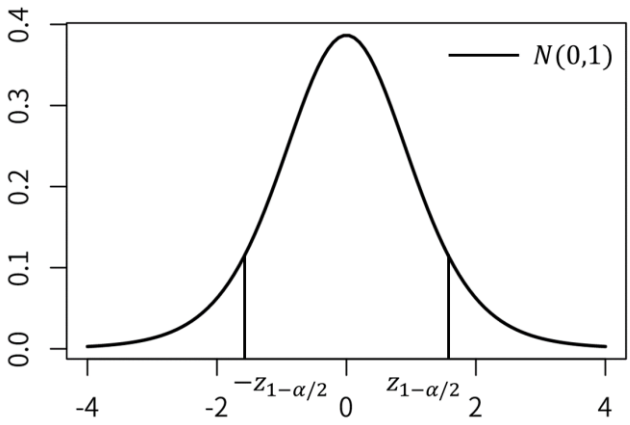

图 5.2：标准正态分布示意图

实例分析

已知案例中手机的续航时间服从正态分布，其标准差为 2 小时，试求该型号手机续航时间的 0.95 置信区间。

使用 R 实现如下：

```
# 续航时间区间估计

# 方法一：按照公式计算

mu <- mean(battery)

sigma <- 2

n <- 50

spread <- qnorm(1-0.05/2, 0, 1) * sigma / sqrt(n)

cat(sprintf('续航时间的置信区间是：[%s, %s]',

            round(mu - spread, 4),

            round(mu + spread, 4), '\n'))
```

```
## 续航时间的置信区间是：[19.2696, 20.3784]
```

```
# 方法二：使用 z.test() 计算

library(BSDA)

interval_1 <- z.test(battery, sigma.x = sigma, conf.level = 0.95)$conf.int

cat(sprintf('续航时间的置信区间是：[%s, %s]',

            round(interval_1[1], 4),

            round(interval_1[2], 4), '\n'))
```

```
## 续航时间的置信区间是：[19.2696, 20.3784]
```

当σ未知时，可以用样本方差s^2估计总体方差σ^2。此时可以使用T统计量：$T=\frac{\sqrt{n}(\bar{x}-\mu)}{s}$，该统计量满足自由度为$n-1$的 t 分布$t(n-1)$。因此 T 可以用来作为枢轴量。和上一种情况的推导类似，可以得到μ的$1-\alpha$置信区间为：

$$\bar{x}\pm t_{1-\frac{\alpha}{2}}(n-1)s/\sqrt{n}$$

其中，$P\left(t(n-1)\le t_{1-\frac{\alpha}{2}}(n-1)\right)=1-\frac{\alpha}{2}$，此处$s^2$是总体方差$\sigma^2$的无偏估计。

（2）两个正态总体均值差的置信区间

设$x_1,x_2,\dots,x_m$是来自$N(\mu_1,\sigma_1^2)$的样本，$y_1,y_2,\dots,y_n$是来自$N(\mu_2,\sigma_2^2)$的样本，且两个样本相互独立，$\bar{x}$和$\bar{y}$分别是他们的样本均值。下面来讨论两个均值差的置信区间。

（2.1）当σ_1^2和σ_2^2已知时，有$\bar{x}-\bar{y}\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}\right)$，取枢轴量为

$$z=\frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$$

得到$\mu_1-\mu_2$的$1-\alpha$置信区间为$\bar{x}-\bar{y}\pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}$。

（2.2）当$\sigma_1^2=\sigma_2^2=\sigma^2$未知时，同样需通过样本方差估计总体方差值。这时有

$$\bar{x}-\bar{y}\sim N\left(\mu_1-\mu_2,\left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2\right)$$

$$\frac{(m-1)s_x^2+(n-1)s_y^2}{\sigma^2}\sim\chi^2(m+n-2),$$

其中s_x^2和s_y^2分别代表两样本各自的样本方差值。由于我们假设$\sigma_1^2=\sigma_2^2=\sigma^2$，因此可以将两样本合并估计总体方差：$s_w^2=\frac{(m-1)s_x^2+(n-1)s_y^2}{m+n-2}$。则此时$\mu_1-\mu_2$的$1-\alpha$置信区间为

$$\bar{x}-\bar{y}\pm\sqrt{\frac{m+n}{mn}}s_w t_{1-\frac{\alpha}{2}}(m+n-2)$$

（2.3）当$\sigma_1^2\neq\sigma_2^2$未知时。

在H_0成立的情况下，检验统计量$T=(\bar{x}-\bar{y})\sqrt{\left(s_x^2/m+s_y^2/n\right)}$服从自由度为

$$v = \frac{\left(s_x^2/m + s_y^2/n\right)^2}{\frac{(s_x^2/m)^2}{m-1} + \frac{\left(s_y^2/n\right)^2}{n-1}}$$

的t分布。则此时$\mu_1 - \mu_2$的$1-\alpha$置信区间为

$$(\bar{x} - \bar{y}) \pm T_{a/2}\sqrt{\left(s_x^2/\mathrm{m} + s_y^2/\mathrm{n}\right)}$$

实例分析

为了了解红楼梦在前、中、后三个部分对虚词使用的习惯，拟对不同章节的“之”字词频进行分析。假设小说对文言虚词的使用频率服从正态分布，试求第 41-80 回和第 81-120 回虚词“之”词频差的 0.95 置信区间。

使用 R 实现如下：

```
wenyan <- read.csv("红楼梦虚词词频统计.csv", stringsAsFactors = F)

freqx <- wenyan$`之`[41:80]

freqy <- wenyan$`之`[81:120]

m = length(freqx); n = length(freqy)

u_x = mean(freqx); u_y = mean(freqy)

s_x = sd(freqx); s_y = sd(freqy)


## (1) 如果两部分数据的方差相等

## 方法一：使用公式计算

s_w = sqrt(((m-1)*s_x^2 + (n-1)*s_y^2) / (m+n-2))

spread = qt(1 - 0.05/2, m+n-2) * s_w * sqrt(1/m + 1/n)

cat(sprintf('如果两部分数据方差相等，第 41-80 回和第 81-120 回“之”词频差的置信区间是：[%s, %s]',

           round(u_x - u_y - spread, 4),

           round(u_x - u_y + spread, 4), '\n'))
```

```
## 如果两部分数据方差相等，第 41-80 回和第 81-120 回“之”词频差的置信区间是：[8.9824, 20.3176]
```

```
## 方法二：使用 t.test() 计算

interval_3 <- t.test(freqx, freqy, var.equal = T, conf.level = 0.95)$conf.int

cat(sprintf('如果两部分数据方差相等，第 41-80 回和第 81-120 回“之”词频差的置信区间是：[%s, %s]',
```

```
          round(interval_3[1], 4),
          round(interval_3[2], 4), '\n'))
## 如果两部分数据方差相等，第 41-80 回和第 81-120 回"之"词频差的置信区间是：[8.9824, 20.3176]
## (2) 如果两部分数据方差不等
## 一般直接使用 t.test()计算
interval_4 <- t.test(freqx, freqy, var.equal = F, conf.level = 0.95)$conf.int
cat(sprintf('如果两部分数据方差不等，第 41-80 回和第 81-120 回"之"词频差的置信区间是：[%s, %s]',
          round(interval_4[1], 4),
          round(interval_4[2], 4), '\n'))
## 如果两部分数据方差不等，第 41-80 回和第 81-120 回"之"词频差的置信区间是：[8.928, 20.372]
```

5.3 假设检验

假设检验（hypothesis test）是统计推断关注的另外一个重要问题。从引入案例可以看出，参数估计侧重参数的估计方式、估计误差等。与参数估计不同，假设检验则是在做判断题：手机的续航时间是否达到了标准？答案只有两个：是，或者否。为了更好地理解假设检验问题，我们再举一个例子。

在减肥药的案例中，FDA 对该减肥药做了一组试验，观察受试者在使用减肥药前后的体重是否有明显的变化。摆在 FDA 面前只有两个结果：(a) 受试者在使用制药公司的减肥产品后体重发生变化；(b) 受试者体重在用药前后没有明显差异。因此，假设检验可以理解为一个判断题。

这样类似的判断题还有哪些？你会发现身边充满了这样的场景：法官判决嫌疑人是否有罪？签证官判断你是否有移民倾向？消费者判断是否应作出购买决策？

如何得到判断题的答案呢？这个答案是对是错？错误的可能性是多大？这是假设检验需要回答的重要问题。为此，以下叙述假设检验的重要步骤。

5.3.1 假设检验的基本步骤

1. 提出假设

假设检验的首要元素是“假设”。首先要提出一个假设，才能针对这个假设的命题进行判断（也就是检验）。这个假设一般被称为“原假设”（null hypothesis），一般用H_0表示。

在案例一中，原假设是：H_0：“市面上的手机续航时间未超过为 20 小时”；在案例二中，原假设是：H_0：“受试者在使用制药公司的减肥药前后体重没有差异”。

如果认为原假设不正确，那么可以选择另外一个备选的假设，也称为“备择假设”（alternative hypothesis）。备择假设一般用H_1表示。

在案例一中，备择假设是：H_1：“市面上的 BearPhone 手机续航时间超过 20 小时”；在案例二中，备择假设是：H_1：“受试者在使用减肥药后体重有明显下降”。

备择选择的确定与问题本身有关。比如，在案例二中，希望得知减肥药是否有显著的减肥效果，则将体重下降确定为备择假设。

2. 选择检验统计量

我们可以通过数学语言对原假设及备择假设进行描述。在案例一中，假设总体 BearPhone 手机的续航时间均值为μ，那么假设可以描述为：

$$H_0: \mu \leq 20; \quad H_1: \mu > 20$$

在案例二中，假设在使用减肥产品体重均值为μ_1，使用后体重均值为μ_2。那么假设问题可以描述为：

$$H_0: \mu_1 = \mu_2; \quad H_1: \mu_2 < \mu_1$$

假设检验的首要任务是确认原假设H_0是否成立。注意到H_0是使用总体的参数描述的。而在现实中，我们只能收集到样本数据，因此需要通过样本数据对总体参数的假设进行判断。在 5.2 节中我们介绍到，对总体均值可以使用样本均值进行估计。因此，在构建检验统计量时，可以将样本均值考虑进来。在案例一中，可以选择样本均值作为检验统计量，并与 20 进行比较；在案例二中，可以将使用减肥药前后的受试者体重均值的差作为检验统计量，并比较其与 0 的差异。

3. 确定拒绝域的形式

在构造完检验统计量之后，我们如何借助这个工具来做出决策呢？这里为了展示基本想法，我们以案例一为例进行说明。首先，我们最关心的是H_0的真伪。有了具体的样本之后，我们可以把样本空间划分为两个互不相交的部分W和$\overline{W}$，当样本属于W时，拒绝H_0；否则接受H_0。于是，我们称W为该检验的拒绝域。而$\overline{W}$称为接受域。

通常我们将注意力放在拒绝域上：正如在数学上我们不能用一个例子去证明一个结论一样，但可以用一个反例来推翻一个命题。因此，从逻辑上来看，注重拒绝域是合适的。事实上，在“拒绝原假设”和“拒绝备择假设”之间还有一个模糊域，如今把它并入接受域，仍然称为接受域。因此，接受域$\overline{W}$中有两类样本点：

（1）一类样本点使得原假设H_0为真，是应该接受的；

（2）另一类样本点多提供的信息不足以拒绝原假设H_0，不宜列入W，只能保留在$\overline{W}$内，待有新的样本信息之后再议。这一点是今后在接受H_0时需要引起关注的。

那么如何表示拒绝域呢？在案例一中，样本均值$\bar{x}$是一个很好的检验统计量，这是因为，样本均值是对正态总体均值的一个无偏的点估计。在案例一中，样本均值$\bar{x}$越大，意味着总体均值θ可能越大；样本均值$\bar{x}$越小，意味着总体均值θ可能越小。所以拒绝域形如$W = \{(x_1, x_2, \dots, x_n): \bar{x} \le c\} = \{\bar{x} \le c\}$是合理的，其中临界值$c$待定。

当拒绝域确定了，检验的判断准则跟着也确定了：如果$(x_1, x_2, \dots, x_n) \in W$，则拒绝$H_0$，如果$(x_1, x_2, \dots, x_n) \in \overline{W}$，则接受$H_0$。由此可见，一个拒绝域$W$唯一确定一个检验法则，反之，一个检验法则也唯一确定一个拒绝域。

4. 给出显著性水平

当我们使用某种检验做判断时，可能做出正确或错误的判断。因此我们可能犯如下两种错误：

第一类错误（type I error），当$\theta \in \Theta$时，但检验拒绝了原假设H_0（$(x_1, x_2, \dots, x_n) \in W$）也称为“拒真”错误；

第二类错误（type II error），当$\theta \in \Theta_1$时，但检验接受了原假设H_0

（$(x_1, x_2, \dots, x_n) \in \overline{W}$），也称为“取伪”错误；

由于样本采样的随机性，以上两种错误无法避免。具体总结如表 5.3.1 所示。

表 5.3.1 检验的两类错误

观测数据情况	总体情况	
	H_0为真	H_1为真
$(x_1, x_2, \dots, x_n) \in W$	第一类错误	正确
$(x_1, x_2, \dots, x_n) \in \overline{W}$	正确	第二类错误

由于检验结果受到样本随机性的影响，所以我们可以用总体分布定义第一类、第二类错误的概率如下：

犯第一类错误概率：$\alpha = P_\theta\{X \in W\}, \theta \in \Theta$，也记为$P\{X \in W|H_0\}$；

犯第二类错误概率：$\beta = P_\theta\{X \in \overline{W}\}, \theta \in \Theta_1$，也记为$P\{X \in \overline{W}|H_1\}$；

理论研究表明，在固定样本量 n 的前提下，要减小α必导致β增大，反之亦然。如果想要同时减小α, β，则需要增加样本量。如何处理α和β之间不易调和的矛盾呢？统计学家根据实际使用情况提出了如下的建议：在样本量n已固定的场合，主要控制第一类错误的概率，并构造出“水平为α的检验”，它的具体定义如下：

【定义 5.3.1】在一个假设检验问题中，先选定一个数α，若一个检验犯第一类错误的概率不超过α，即

$$P(\text{type I error}) \leq \alpha$$

则称该检验是水平为α的检验，其中α称为显著性水平。若在检验中拒绝了原假设，则这个检验是显著的。

由于α过小会导致β过大，因此在一个检验中显著性水平α不宜定得过小。在实际中常选择$\alpha = 0.05$，有时也用$\alpha = 0.1$或$\alpha = 0.01$。

综上所述，进行假设检验都需要经过上述四个步骤，即：

（1）建立假设：原假设H_0和备择假设H_1；

（2）选择合适的检验统计量；

（3）确定拒绝域W的形式；

（4）给出显著性水平α，确定临界值，做出判断。

讲解完假设检验的基本步骤，我们再回到前面提到的一个问题：如何确定哪

个是H_0？这个问题没有统一的答案，但一般的做法是，选择一个保守的选项作为H_0。比如，在制药公司减肥药的例子中，零假设设定$\mu_1 = \mu_2$。这里 FDA 默认是什么呢？可以看出，FDA 默认减肥药没有效果，这是一个保守的选择。在假设检验的过程，则要小心地拒绝零假设。为什么要小心地拒绝零假设呢？如果减肥药没有效果，但是错误判定为有效果，那么可能给 FDA 带来消费者的无限投诉甚至是法律诉讼。因此，FDA 在拒绝零假设时，必须要控制好显著性水平。相反，如果减肥药有效果，但是接受了零假设，这可能仅是错过了一个潜在的好产品，但不会带来更严重的问题。因此，这样设定零假设对 FDA 更为谨慎。

5.3.2 假设检验的 p 值

提出和使用假设检验的四个基本步骤是强调正确进行假设检验的方法，当熟悉了这个方法之后，有些步骤并不总是需要的——现实场景中，我们一般使用 p 值（即 p-value）直接给出假设检验问题的判断。那么 p 值具体指的是什么呢？我们将在这一小节给出详细介绍。

假设检验的结论通常是简单的，在给定显著性水平下，只能选择拒绝原假设或者接受原假设。然而有时会出现这样的情况：在一个较大的显著性水平（比如$\alpha = 0.05$）下得到拒绝原假设的结论，而在一个较小的显著性水平（比如$\alpha = 0.01$）下却会接受原假设。这种情况在理论上很容易解释：显著性水平变小后会导致检验的拒绝域变小，于是原来落在拒绝域中的观测值可能就落入了接受域。但这种情况在实际应用中会带来一些麻烦：假如这时一个人主张选择显著性水平$\alpha = 0.05$，而另一个人主张选择$\alpha = 0.01$，则两个人就会得到截然相反的结论。

我们用一个例子更直观地说明这个问题：

假设手机 BearPhone 的续航时间服从正态分布$N(\theta, 0.8^2)$，其中θ的设计值不低于 20 小时。根据案例引入中给出的 50 次抽样检测数据，判断该手机的续航时间是否满足出厂设计的要求？

对这个实际问题，建立一对假设：

$$H_0: \theta \geq 20 \ \ vs \ \ H_1: \theta < 20$$

由于总体方差已知，且总体均值 $\bar{X} \sim N\left(\theta, \frac{0.8^2}{50}\right)$，则可构造检验统计量：

$$Z=\frac{\sqrt{50}\times(\bar{X}-\theta)}{0.8}\sim N(0,1)$$

拒绝域形式为$W=\left\{\frac{\sqrt{50}(\bar{X}-\theta)}{0.8}\leq\frac{\sqrt{50}(c-\theta)}{0.8}\right\}$，即对于给定的显著性水平$\alpha$，要求对于任意的$\theta\geq 20$（$H_0$成立），第一错误的概率$P\left(\frac{\sqrt{50}(\bar{X}-\theta)}{0.8}\leq\frac{\sqrt{50}(c-\theta)}{0.8}\right)=\Phi\left(\frac{\sqrt{50}(c-\theta)}{0.8}\right)\leq\alpha$。由于不等号左侧是$\theta$的减函数，因此只需要

$$\Phi\left(\frac{\sqrt{50}(c-\theta)}{0.8}\right)=\alpha$$

成立即可。求解以上方程式可得：$c=\theta+\frac{0.8}{\sqrt{50}}z_\alpha=20+\frac{0.8}{\sqrt{50}}z_\alpha$，其中$z_\alpha$代表标准正态分布的$\alpha$分位数。因此，检验的拒绝域为$W=\{\bar{X}\leq 20+\frac{0.8}{\sqrt{50}}z_\alpha\}$。由于样本观测值$\bar{x}=19.824$，可以计算得检验统计量$z=\frac{\sqrt{50}\times(\bar{x}-\theta)}{0.8}=-1.56$，我们选择不同的显著性水平，比较该检验问题的结论：

表 5.3.2 案例一中的拒绝域

显著性水平	拒绝域	对应的结论
$\alpha=0.1$	$z\leq-1.282$	拒绝H_0
$\alpha=0.05$	$z\leq-1.645$	接受H_0
$\alpha=0.025$	$z\leq-1.96$	接受H_0
$\alpha=0.01$	$z\leq-2.326$	接受H_0

现在换一个角度来看，在$\theta=20$时，检验统计量$Z\sim N(0,1)$，此时由样本可以算得$z=-1.56$，据此可以算得一个概率$p=P(Z\leq-1.56)=\Phi(-1.56)=0.0599$，若以此为基准看待上述检验问题，同样可以做出判断：

- 当$\alpha<0.0599$时，$z_\alpha<-1.56$，由于拒绝域为$W=\{Z\leq z_\alpha\}$，于是观测值$z=-1.56$不在拒绝域里，应接受原假设；
- 当$\alpha\geq 0.0599$时，$z_\alpha\geq-1.56$，由于拒绝域为$W=\{Z\leq z_\alpha\}$，于是观测值$z=-1.56$落在拒绝域里，应拒绝原假设；

由此可以看出，0.0599 是能用样本计算得到的检验统计量$z=-1.56$做出“拒绝H_0”的最小的显著性水平，这就是p值。

【定义 5.3.2】在一个假设检验问题中，利用样本观测值能够做出拒绝原假设的最小显著性水平称为检验的 p 值。

将检验的 p 值与事先给定的显著性水平α进行比较可以得出假设检验的结论：

- 如果$p \leq \alpha$，则在显著性水平α下拒绝H_0；
- 如果$p > \alpha$，则在显著性水平α下接受H_0；

后面章节的检验问题可以从两方面进行，一方面是建立拒绝域，考察样本观测值是否落入拒绝域而加以判断；另一方面是根据样本观测值计算检验的 p 值，通过将 p 值与事先设定的显著性水平α比较大小而做出判断。两个角度是等价的，因此选择较为方便的一种方法即可。

在 R 中计算单边检验 p 值的程序如下：

```
a <- 20

s <- 0.8

n <- 50

xbar <- mean(battery)

p_value <- pnorm(xbar, mean = a, sd = s/sqrt(n))

print(paste0("检验的 p 值为", round(p_value, 4)))
```

```
## [1] "检验的 p 值为 0.0599"
```

5.3.3 假设检验问题的基本类型

常见的假设检验分为单样本检验和双样本检验两种情况，其中，每一种检验又可以分为单侧检验和双侧检验。这里以正态总体$N(\mu, \sigma^2)$的均值μ为例，分别介绍基本的四种检验问题。

1. 单样本检验与两样本检验

设$x_1, \ldots, x_n$是来自$N(\mu, \sigma^2)$的样本，考虑如下三种关于μ的检验问题：

$$H_0: \mu \leq \mu_0 \quad vs \quad H_1: \mu > \mu_0$$

$$H_0: \mu \geq \mu_0 \quad vs \quad H_1: \mu < \mu_0$$

$$H_0: \mu = \mu_0 \quad vs \quad H_1: \mu \neq \mu_0$$

其中μ_0是已知常数。这种只对一组样本所属的总体参数进行检验的问题为单样本检验问题。在前文给出的例题中，案例一为单样本检验问题。

设$x_1,\dots,x_n$是来自正态总体$N(\mu_1,\sigma_1^2)$的样本，$y_1,\dots,y_n$是来自另一个正态总体$N(\mu_2,\sigma_2^2)$的样本，两个样本相互独立，考虑如下三类检验问题：

$$H_0:\mu_1-\mu_2\le 0 \quad vs \quad H_1:\mu_1-\mu_2>0$$

$$H_0:\mu_1-\mu_2\ge 0 \quad vs \quad H_1:\mu_1-\mu_2<0$$

$$H_0:\mu_1-\mu_2=0 \quad vs \quad H_1:\mu_1-\mu_2\ne 0$$

这种对两组来自不同总体的样本所属的总体参数进行检验的问题统称为两样本检验问题，如上一节的减肥药案例。

在 5.3.4 节中，将给出更多实例分析来协助大家更好地理解与区分单样本和双样本检验问题。

2. 单边检验与双边检验

顾名思义，单边检验为具有方向性的检验，即判断总体参数是否大于或小于某个已知的值，如案例一中的检验问题；双边检验不具有方向性，结论只能得出总体参数是否等于某个值，或两个总体的参数是否有显著差异，但是无法判断孰高孰低，如案例二中判断受试者在使用减肥产品后体重是否发生显著变化。

在 5.3.4 节中，将给出更多实例分析来协助大家更好地理解与区分单边检验和双边检验问题。

5.3.4 正态总体的假设检验

在上一节对假设检验的基本问题类型介绍之后，本节对正态总体参数μ的各种检验分别进行讨论。

1. 单个正态总体均值的检验

设$x_1,\dots,x_n$是来自$N(\mu,\sigma^2)$的样本，考虑如下三种关于μ的检验问题：

$$I:H_0:\mu\le\mu_0 \quad vs \quad H_1:\mu>\mu_0$$

$$II:H_0:\mu\ge\mu_0 \quad vs \quad H_1:\mu<\mu_0$$

$$III:H_0:\mu=\mu_0 \quad vs \quad H_1:\mu\ne\mu_0$$

其中μ_0是已知常数。由于正态总体含有两个参数，总体方差σ^2已知与否对检验有影响，因此我们分σ已知和未知两种情况叙述。

情况一：σ已知时的z检验

（1）单边假设检验

对于I所示的单边检验问题，由于μ的点估计是 $\bar{x}$，且 $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$，故选择$z$检验统计量：

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

是恰当的。根据假设的符号设定，可以初步判断当样本均值 $\bar{x}$不超过设定均值μ_0时，应当倾向于接受原假设；当样本均值 $\bar{x}$超过设定均值μ_0时，应当倾向于拒绝原假设。然而，观测是具有随机性的，在这种情况下，当$\bar{x}$比μ_0大到一定程度时，才有足够的信心拒绝原假设。因此，存在一个临界值c，拒绝域为：

$$W_1 = \{(x_1, \dots, x_n): z \geq c\}$$

常简记为$\{z \geq c\}$，若要求检验的显著性水平为α，则c需要满足：

$$P_{\mu_0}(z \geq c) = \alpha$$

其中$\mathrm{P}_{\mu_0}(\cdot)$表示在$\mu = \mu_0$的情况下的概率值。由于在$\mu = \mu_0$时，$z \sim N(0,1)$，因此$c = z_{1-\alpha}$（见图 5.3），得到拒绝域为图 5.3（a）中的阴影部分：

$$W_1 = \{z \geq z_{1-\alpha}\}$$

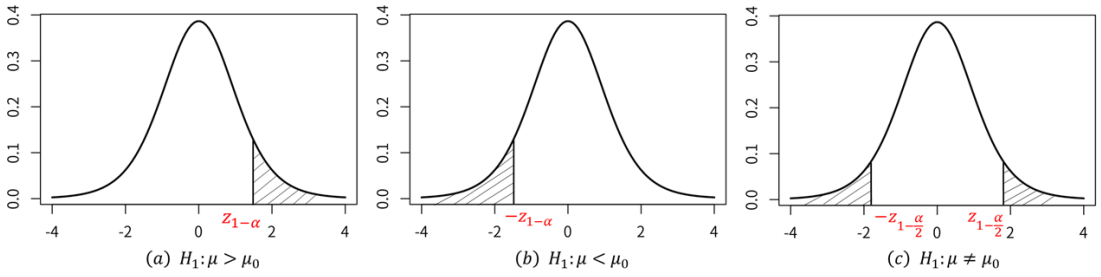

图 5.3：三种检验问题的拒绝域

图 5.3 展示了三种不同检验问题中，在显著性水平为α的情况下对应的拒绝域（阴影部分），其中横轴表示统计量z的取值，曲线为标准正态分布统计量z的密度曲线。

同样地，使用 p 值也可以进行等价的检验。此时可以计算 p 值为：$p_I = 1 - \Phi(z)$，其中$\Phi(\cdot)$为标准正态分布的分布累积函数。

对检验问题II所示的单边检验问题的讨论是类似的，其拒绝域（见图 5.3（b））

为：

$$W_{II} = \{z \leq z_{1-\alpha}\}$$

而检验的 p 值为

$$p_{II} = \Phi(z)$$

其中z的含义同上。

（2）双边假设检验

对于III所示的双边检验问题，检验的 p 值稍有不同。考虑到检验问题III的备择假设H_1分散在两侧，故其拒绝域应当在两侧，即拒绝域如下所示：

$$W_{III} = \{|z| \geq c\}$$

对于给定的显著性水平$\alpha(0 < \alpha < 1)$，由$P_{\mu_0}(|z| \geq c) = \alpha$可以得出$c = z_{1-\frac{\alpha}{2}}$（见图 5.3（c）），得到的拒绝域为：

$$W_{III} = \{|z| \geq z_{1-\alpha/2}\}$$

对于检验统计量分布是对称的情况，双边检验的 p 值的计算与单边检验是相似的。此时 p 值的计算方式为：

$$p_{III} = 2(1 - \Phi(|z|)) \qquad (5.3.2)$$

（3）实例分析

以案例一中提到的手机数据为例，对其续航时间进行双边假设检验。仍然使用案例一中给出的 50 部手机续航时间观测值，并且已知续航时间是一个服从正态分布$N(\mu, 0.8^2)$的随机变量。猜测该型号的手机续航时间为 20 小时，是否可以接受这个猜测呢？

首先，这是一个双边假设检验问题，总体$X \sim N(\mu, 0.8^2)$，待检验的原假设H_0和备择假设H_1分别为：

$$H_0: \mu = 20 \quad vs \quad H_1: \mu \neq 20$$

检验的拒绝域为$\{|z| \geq z_{1-\frac{\alpha}{2}}\}$。取显著性水平$\alpha = 0.05$，则对应的分位数为$z_{0.975} = 1.96$。根据该例中的观测值可以计算得出：

$$\bar{x} = 19.824, |z| = \frac{\sqrt{50}|19.824 - 20|}{0.8} = 1.56$$

z统计量的值未落入拒绝域$\{|z| \geq 1.96\}$内，因此不能拒绝原假设。

我们也可以采用 p 值完成此次检验，可以计算得 p 值为：

$$p = 2\big(1 - \Phi(1.56)\big) = 0.1188$$

由于 p 值大于实现给定的水平 0.05，因此不能拒绝原假设，结论是相同的。

进一步地，我们从 p 值还可以看到，只要事先给定的显著性水平不高于 0.1188，则不能拒绝原假设；而如果事先给定的显著性水平高于 0.1188，则拒绝原假设。

在 R 中使用 z.test()进行检验如下：

```
# 单样本均值的双边检验
z.test(battery, mu = 20, sigma.x = 0.8, alternative = "two.sided", conf.level
= 0.95)
```

```
##
##  One-sample z-Test
##
## data:  battery
## z = -1.5556, p-value = 0.1198
## alternative hypothesis: true mean is not equal to 20
## 95 percent confidence interval:
##  19.60226 20.04574
## sample estimates:
## mean of x
##    19.824
```

情况二：σ未知时的t检验

（1） 理论分析

对于检验问题 I，当σ未知时，z 统计量则会因为含有未知参数σ而无法计算，此时需要选择其他的检验统计量。将 z 统计量表达式中未知的σ替换成样本标准差 s，这就形成t检验统计量：

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

在$\mu = \mu_0$时，t服从自由度是 n-1 的 t 分布$t(n-1)$，从而检验问题I的拒绝域为：

$$W_I = \{t \geq t_{1-\alpha}(n-1)\}$$

p 值的计算是类似的，对给定的观测样本值，可以计算出相应的检验统计量

t的值，记为$t=\frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$，其中 $\bar{x},s$是样本均值及标准差。此时 p 值可以计算为：

$$p_I = 1 - F(t)$$

其中F(·)表示自由度为 n-1 的 t 分布的累积分布函数。对于另外两组检验问题的讨论类似于上一节，这里只列出结果。其中，检验问题 II 的拒绝域及 p 值为：

$$W_{II} = \{t \le t_{1-\alpha}(n-1)\},\ p_{II} = F(t)$$

检验问题 III 的拒绝域及 p 值为：

$$W_{III} = \{|t| \ge t_{1-\frac{\alpha}{2}}(n-1)\},\ p_{III} = 2*(1-F(|t|))$$

（2） 实例分析

如果只知道案例一给出的手机续航时间服从正态分布，但不知道总体方差是多少，如果还是猜测该型号的手机续航时间为 20 小时，此时是否可以接受这个猜测呢？

此时原假设仍然是$H_0:\mu=20$，备择假设是$H_1:\mu\neq 20$。由于σ未知，故采用t检验，其拒绝域为$\{|t| \ge t_{1-\alpha/2}(n-1)\}$，若取$\alpha=0.05$，则查表可得$t_{0.975}(49)=2.010$。现由样本计算得到 $\bar{x}=19.824, s=1.41$，故

$$t=\sqrt{50}\times\frac{|19.824-20|}{1.41}=0.883$$

由于 0.883<2.010，因此不能拒绝原假设，认为该型号的手机续航时间为 20 小时。

R 中使用 t.test()检验如下：

```
# 单样本均值的双边检验（方差未知）
t.test(battery, mu = 20, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  battery
## t = -0.88266, df = 49, p-value = 0.3817
## alternative hypothesis: true mean is not equal to 20
## 95 percent confidence interval:
##  19.42329 20.22471
## sample estimates:
```

```
## mean of x

##   19.824
```

2. 两个正态总体均值差的检验

（1）理论分析

设$x_1,\dots,x_n$是来自正态总体$N(\mu_1,\sigma_1^2)$的样本，$y_1,\dots,y_n$是来自另一个正态总体$N(\mu_2,\sigma_2^2)$的样本，两个样本相互独立，考虑如下三类检验问题：

$$I\quad H_0:\mu_1-\mu_2\le 0\quad vs\quad H_1:\mu_1-\mu_2>0$$

$$II\quad H_0:\mu_1-\mu_2\ge 0\quad vs\quad H_1:\mu_1-\mu_2<0$$

$$III\quad H_0:\mu_1-\mu_2=0\quad vs\quad H_1:\mu_1-\mu_2\ne 0$$

这种对两组来自不同总体的样本所属的总体参数进行检验的问题统称为两样本检验问题。实际中一般假设两样本方差相同但未知。 此时可首先构造检验统计量$\bar{x}-\bar{y}\sim N\left(\mu_1-\mu_2,\left(\frac{1}{m}+\frac{1}{n}\right)\sigma^2\right)$。由于方差相同且未知，可以将两样本合并用于估计方差：

$$s_w^2=\frac{1}{m+n-2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(y_i-\overline{y})^2\right]$$

由于

$$\frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i-\overline{x})^2-x^2(m-1),\quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\overline{y})^2\sim x^2(n-1)$$

可得$\frac{1}{\sigma^2}(\sum(x_i-\overline{x})^2+\sum(y_i-\overline{y})^2)\sim x^2(m+n-2)$。此时可以推导得

$$t=\frac{(\overline{x}-\overline{y})-(\mu_1-\mu_2)}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim t(m+n-2)$$

这就给出了t检验统计量的表达式

$$t=\frac{\overline{x}-\overline{y}}{s_w\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

对检验问题 I，检验的拒绝域和 p 值分别为

$$W_I=\{t\ge t_{1-\alpha}(m+n-2)\}\qquad p_I=1-F(t)$$

原假设成立时， t是服从自由度是$n+m-2$的t分布的统计量。对检验问题 II，检验的拒绝域和 p 值分别为

$$W_{II} = \{t \le t_{1-\alpha}(m+n-2)\} \quad p_{II} = F(t)$$

对检验问题 III，检验的拒绝域和 p 值分别为

$$W_{III} = \{|t| \ge t_{1-\alpha/2}(m+n-2)\} \quad p_{III} = 2*(1-F(|t|))$$

当我们不能确定两样本的未知方差相等时，常使用如下的近似检验。若 $\bar{x} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right), \bar{y} \sim N(\mu_2, \frac{\sigma_2^2}{m})$，且两者独立，则

$$\bar{x} - \bar{y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

构造 t 化统计量

$$t^\star = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

$t^\star$不服从 t 分布，但是其形式非常像 t 统计量。这里不加证明地给出，$t^\star$近似服从自由度为 l 的 t 分布，其中，

$$l = \left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2 / \left[\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}\right]$$

（2）实例分析

为了从词语角度出发，看红楼梦作者在虚词使用习惯上是否存在差异，统计了红楼梦原著每一回的文言虚词使用频数，这里以“之”和“亦”两个具有代表性的虚词为例。将 41-80 回以及 81-120 回看作两个总体[1]，探索虚词在使用频率上是否有显著差异。在显著性水平$\alpha = 0.05$下，尝试判断 41-80 回中“亦”使用的频数均值是否比 81-120 回中高？

解：用X表示 41-80 回词频，Y表示 81-120 回词频，则由假定，$X \sim N(\mu_1, \sigma^2), Y \sim N(\mu_2, \sigma^2)$，要检验的假设是：$H_0: \mu_1 = \mu_2 \ \ vs \ \ H_1: \mu_1 > \mu_2$，由于二者方差未知但相等，故采用两样本$t$检验，经计算：

$$\bar{x} = 3.90, \quad \bar{y} = 0.75$$

$$\sum_{i=1}^{12}(x_i - \overline{x})^2 = 770.8718, \quad \sum_{i=1}^{12}(y_i - \overline{y})^2 = 44.6154$$

[1] 实际数据中，很难验证总体服从正态分布。但一般两总体检验的方法论依然可用，这是因为在样本较大的前提下可以理论上验证，假设检验的结论基本成立。

从而$s_w = \sqrt{\frac{1}{40+40-2}(0.1109+0.0771)} = 3.2334$

$$t = \frac{3.90-0.75}{3.2334 \cdot \sqrt{\frac{1}{40}+\frac{1}{40}}} \approx 4.4$$

通过 R 计算可知$t_{0.95}(78) \approx 1.6646$，由于$t > t_{0.95}(78)$，因此拒绝原假设，认为 41-80 回“亦”使用的平均数量显著高于 81-120 回，后期语言向白话文靠拢。

下面用 p 值再做一次检验，因t是服从自由度为 78 的t分布的统计量，则$p = P(t \geq 4.4)$，使用 R 计算得到 p 值约为$1.6 * 10^{-5}$，由于p值小于事先给定的显著性水平 0.05，因此拒绝原假设，结论是相同的。

在 R 中进行该问题的两样本t检验如下所示：

```
wenyan <- read.csv("红楼梦虚词词频统计.csv", stringsAsFactors = F)
freqx <- wenyan$`亦`[41:80]
freqy <- wenyan$`亦`[81:120]
t.test(freqx, freqy, alternative = "greater", var.equal = T, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  freqx and freqy
## t = 4.4123, df = 78, p-value = 1.621e-05
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.961594      Inf
## sample estimates:
## mean of x mean of y
##      3.90      0.75
```

该检验为单边检验，因此函数的输出结果中，给出的两样本均值差$\bar{x} - \bar{y}$的 0.95 置信区间为$[1.96, \infty)$。

3. 成对数据检验

在对两个总体均值进行比较时，有时数据是成对出现的，那么此时若采用二样本t检验。得出的结论有可能是不合理的。下面以减肥药的受试者体重试验为例进行说明。

在减肥药案例数据中，假定受试者的体重数据服从正态分布，试问：试验前后的受试者体重在显著性水平$\alpha = 0.05$上有无差异？

解：假定$x \sim N(\mu_1, \sigma_1^2), y \sim N(\mu_2, \sigma_2^2)$，且$x$和$y$相互独立，这里假定两个总体的方差相等是合理的（因为保持了培养皿试验前后的其他条件不变）。我们先用两样本t检验讨论此问题。为此，记该试验前后含量的样本均值分别为 $\bar{x}, \bar{y}$，样本方差分别为s_x^2, s_y^2，如今要判断检验问题：

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

在假定$\sigma_1^2 = \sigma_2^2 = \sigma^2$下，采用两样本$t$检验，检验统计量$t_1$与拒绝域$W_1$分别是

$$t_1 = \frac{\overline{x} - \overline{y}}{s_w/\sqrt{n/2}}$$

$$W_1 = \left\{|t_1| > t_{1-\frac{\alpha}{2}}(2n-2)\right\}$$

其中$s_w^2 = \frac{(s_x^2 + s_y^2)}{2}$，$\alpha$是给定的显著性水平，由给定的数据可算得

$$\overline{x} = 55.7, \quad \overline{y} = 53.1, \quad s_x^2 = 14.23, \quad s_y^2 = 33.21, \quad s_w^2 = 23.722$$

从而可算得两样本的t检验统计量的值

$$t_1 = \frac{55.7 - 53.1}{4.8705/\sqrt{10/2}} = 1.1937$$

若给定$\alpha = 0.05$，查表得$t_{0.0975}(18) = 2.1009$，由于$|t_1| < 2.1009$，故不应拒绝原假设，即认为试验前后的受试者体重没有显著差别，此处检验的 p 值为0.2467。

下面我们换一个角度来讨论此问题。在这个问题中出现了成对数据，同一条件下的一个受试者在试验前后有两次体重测量值，其差$d_i = x_i - y_i \sim N(\mu, \sigma_d^2)$，其中$\mu = \mu_1 - \mu_2, \sigma_d^2 = \sigma_1^2 + \sigma_2^2$，原先要比较$\mu_1$和$\mu_2$的大小，如今则转化为考察$\mu$

是否为零，即考察如下的检验问题：

$$H_0: \mu = 0 \quad vs \quad H_1: \mu \neq 0$$

即把双样本的检验问题转化为单样本t检验问题。这时检验的t统计量为

$$t_2 = \frac{\bar{d}}{\left(\frac{s_d}{\sqrt{n}}\right)}$$

其中

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad s_d = \left(\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2\right)^{\frac{1}{2}}$$

在给定的显著性水平α下，该检验问题的拒绝域是

$$W_2 = \{|t_2| \geq t_{1-\frac{\alpha}{2}}(n-1)\}$$

这就是成对数据的t检验。

在本例中，可以算得

$$n = 10, \ \overline{d} = 2.6, \ s_d = 3.5024$$

于是

$$t_2 = \frac{2.6}{3.5024/\sqrt{10}} = \frac{2.6}{1.1076} = 2.3475$$

对于给定的显著性水平$\alpha = 0.05$，可以查表得到$t_{0.975}(9) = 2.2622$。由于$|t_2| > 2.2622$，故应拒绝原假设$H_0: \mu = 0$，即可认为试验前后的受试者体重有显著差异。此处检验的 p 值为 0.0435。进一步，平均含量差值的估计量为 $\hat{\mu} = \bar{x} - \bar{y} = 2.6$，可见服用药物后的体重要低于服用药物前的体重。

在 R 中进行成对样本的t检验如下所示：

```
x <- c(50,59,55,60,58,54,56,53,61,51)

y <- c(43,55,49,62,59,49,57,54,55,48)

# 成对数据 t 检验

t.test(x, y, alternative = "two.sided", paired = T, var.equal = F, conf.level
= 0.95)
```

```
##

##  Paired t-test

##
```

```
## data:  x and y
## t = 2.3475, df = 9, p-value = 0.04348
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.09454818 5.10545182
## sample estimates:
## mean of the differences
##                     2.6
```

本问题中，成对数据t检验方法更加合理。这是因为成对数据的差d_i消除了受试者之间的差别（如不同体质的差异），从而用于检验的标准差$s_d = 3.5024$已经排除了体质之间差异的影响，只保留服用药物前后体重之间的差异。而两样本t检验中用于检验的标准差$s_w = 4.8705$还包含了不同受试者体质之间的差异，从而使得标准差增大，导致因子不显著。所以成对数据场合化为单样本t检验所作的结论更可靠。

应注意，成对数据的获得事先要做缜密的试验设计，在获得成对数据时不能发生“错位”，从而准确获得“成对数据”的信息。

5.4 单因素方差分析

方差分析（Analysis of Variance，简称 ANOVA），是用于对两个或者两个组别以上样本均值差别的显著性检验，它关注的重点是对于组别差异的分析。例如，不同城区的房价是否存在差异？不同岗位的薪资是否会有区别？不同年份的电影评分是否不同？这就需要研究一个连续型变量（例如：房价、薪资、评分）在定性变量（城区、岗位、年份）各个组别之间的差异。方差分析是是进行这类分析的统计分析手段。

当方差分析只涉及一个分类变量时，就称为单因素方差分析（one way ANOVA）。这里的因素（factor）指的是多水平（level）变量（即分类变量，例如：数据集中的制片国家、上映年份等）。方差分析中的因素，就是方差分析中要检

验的对象；而水平则指的是一个因素中的不同组别。方差分析本质上是通过检验各组别的均值是否存在显著差异，来判断分类变量对因变量的影响程度。

例如在本案例中，对于是否为美国制片这一因素我们关心的是：美国制片与非美国制片的电影，它们的评分是否显著不同？哪一种电影的评分最高？在本节接下来的部分，将结合实际案例，从单因素方差分析开始，详细介绍方差分析的原理及基本步骤。

5.4.1 单因素方差分析的基本思路

1. 提出假设

与第五章中讲到的参数检验类似，方差分析的第一步是要提出假设。方差分析是为了检验因素的k个水平对应的因变量均值$\mu_i(i = 1,2,\dots,k)$是否相等。因此，需要提出的假设为：

原假设：$H_0{:}\,\mu_1 = \mu_2 = \cdots = \mu_k$，即该因素对因变量没有影响；

备择假设：$H_1{:}\,\mu_i, i = 1,2,\dots,k$不全相等，即该因素对因变量有影响。

在本案例中，我们首先选取只有两个水平的因子型变量（制片国家是否为美国）为例进行方差分析，这里对应的两水平为：美国制片/非美国制片，即$k = 2$。原假设为H_0:美国制片的电影与非美国制片的电影评分相等；备择假设为：H_1：美国制片的电影与非美国制片的电影评分不相等。接下来，将构造检验统计量来进行统计决策。

2. 构造检验统计量

在方差分析中构造的检验统计量是 F 统计量，主要通过计算多种组内及组间平方和求得。设自变量共有 k 个水平，第 i 个水平的第 j 个因变量观测值记为y_{ij}。记$\bar{y}_i$为该水平组内的因变量均值，n_i为该水平内观测值的样本数。$\bar{\bar{y}}$是全部观测数据因变量的均值，n为全部观测值的个数。以下介绍三种重要的平方和。

a. 总平方和（SST）

总平方和为全部观测值y_{ij}与总体平均值$\bar{\bar{y}}$之间的误差平方和，反映了这些所有观测值与平均值之间的离散程度，计算公式为：

$$\mathrm{SST}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2$$

总平方和的自由度为$\mathrm{n}-1$。

b. 组间平方和（SSA）

组间平方和是各组平均值$\bar{y}_i$与总平均值$\bar{\bar{y}}$之间的误差平方和，反映了组与组之间的差异程度，计算公式为：

$$\mathrm{SSA}=\sum_{i=1}^{k}n_i(\bar{y}_i-\bar{\bar{y}})^2$$

组间平方和的自由度为$\mathrm{k}-1$，组间均方误差$\mathrm{MSA}=\frac{\mathrm{SSA}}{\mathrm{k}-1}$。这里需要注意到的是，如果组别之间的差异越大，那么组间平方和SSA和组间均方误差MSA的值也就更大。

c. 组内平方和（SSE）

组内平方和是每个水平或组的各样本数据y_{ij}与组内平均值$\bar{y}_i$的误差平方和，反映了每个组别内观测值的离散程度，也称为误差平方和。计算公式为：

$$\mathrm{SSE}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

组内平方和的自由度为$\mathrm{n}-\mathrm{k}$，组内均方误差 $\mathrm{MSE}=\frac{\mathrm{SSE}}{\mathrm{n}-\mathrm{k}}$。与组间平方和不同的是，组内样本值的差异越大，那么组内平方和SSE以及组内均方误差 MSE也就越大。

三个平方和之间的关系为：$\mathrm{SST}=\mathrm{SSA}+\mathrm{SSE}$。

如果原假设成立，则表明各组别均值之间没有显著性差异。那么，组间平方和 SSA 就不会太大；如果组间均方显著地大于组内均方，说明各水平(总体)之间的差异不仅有随机误差，还有组间差异。SSA 多大才叫大呢？这里我们可以使用 F 检验统计量：

$$\mathrm{F}=\frac{\mathrm{MSA}}{\mathrm{MSE}}$$

可以证明，F 统计量服从分布$F(k-1,n-k)$。F 统计量可以看作某种标准化后的组间差异。

为了使计算过程和步骤更加清晰明确，可以将方差分析的结果以方差分析表的形式呈现，如表格 6-2 所示：

表格 6- 1 单因素方差分析表

误差来源	平方和	自由度	均方误差	F 值
组间（因素影响）	SSA	$k-1$	MSA	$F=\frac{MSA}{MSE}$
组内（误差）	SSE	$n-k$	MSE	
总和	SST	$n-1$		

在 R 语言当中，函数 aov()可计算出 F 统计量的值，对电影是否为美国电影的评分水平进行计算，结果如下所示：

```
aov_nation <- aov(score ~ is_US, data = dat)
summary(aov_nation)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## is_US         1    9.5   9.506   1.128  0.289
## Residuals   248 2089.7   8.426
```

结果中的 Df 为自由度；Sum Sq 下的结果分别对应误差平方和 SSA 和 SSE；Mean Sq 下为均方误差 MSA 和 MSE；F 统计量的值为 1.128，p 值为 0.289。如果在 0.05 的显著性水平下，可以认为是否美国制片对于电影的评分没有显著影响。根据以上方差分析结果，下面介绍如何做出统计决策。

3. 做出统计决策

根据所给定的显著性水平$\alpha=0.05$，可以做出统计决策。在上面的结果里，p 值大于 0.05，因此我们不能拒绝原假设：即可以认为在 95%的置信水平下，电影是否为美国制作对于电影的评分没有显著影响。接下来，将进一步利用方差分析，分析其他因素对电影评分的影响程度。

5.4.2 实例分析

多水平的单因素方差分析（例如上映年份）与两水平的单因素方差分析步骤完全一致，通过同样的方式构造 F 统计量进行检验。我们在 R 语言中使用函数

aov()来对三个解释变量：是否为美国制片、电影类别、上映年份分别依次进行单因素方差分析，结果如表格 6-3 所示：

表 5.4.1 方差分析结果表

变量名称	自由度	F 值	p 值	显著性
是否美国制片	1	1.128	0.289	
电影类别	5	0.544	0.743	
上映年份	2	4.723	0.01	***

根据表 5.4.1 的结果，我们发现，电影类别的方差分析 p 值大于 0.05，而上映年份的方差分析 p 值小于 0.05，可以得出结论：在 95%置信水平下，不同的电影类别对电影评分的影响没有显著性差异，而不同的上映年份对电影评分的影响具有显著性差异。

然而，仅仅通过方差分析的 F 统计量，是无法判断各个水平因变量均值的相对大小的。可以调用 R 语言里的函数 TukeyHSD()实现各个水平之间的比较。该函数可以计算各个组别之间的差异，并找到均值最大和最小的组别，对结果进行可视化。以公司类别为例，分析各个组别的均值差异。

```
# 均值比较
TukeyHSD(aov_year)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ year_break, data = dat)
##
## $year_break
##                                 diff       lwr          upr     p adj
## 2000年以前-2000-2010年     0.257479 -0.747529  1.262486950 0.8179909
## 2010年及以后-2000-2010年 -1.120423 -2.242597  0.001751048 0.0504590
## 2010年及以后-2000年以前   -1.377902 -2.467637 -0.288166687 0.0088322
```

由于上映年份共有 3 个水平，故两两的差值共有 3 对。输出的结果中，diff

代表两个组别之间差异的均值，以第一行为例，2000 年以前的电影比 2000-2010 年上映的电影评分平均高 0.26 分，但它对应的 p 值大于 0.05，因此该差异并不显著；lwr 和 upr 分别代表了该差异的 95%置信下限和上限。使用绘图函数 plot() 可以将该结果可视化：

```
# 设置字体大小和方向
par(family = "STXihei",las = 1, mai = c(1.02,3,0.82,0.42))
# 将方差分析结果可视化
plot(TukeyHSD(aov_year))
```

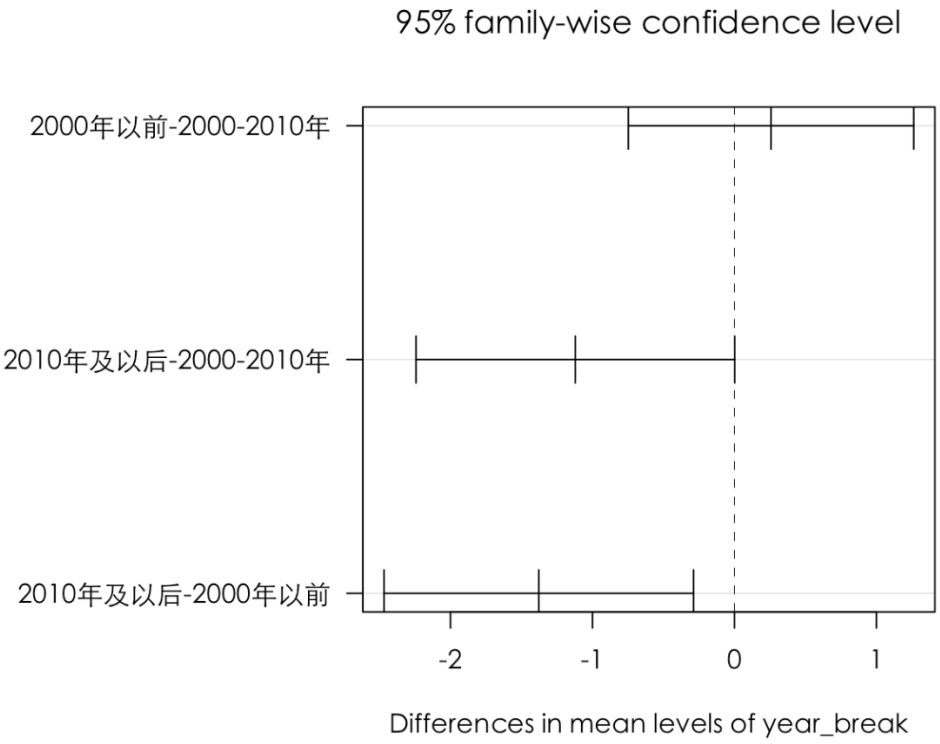

图 5.4：各组别均值差异 95%区间

在图 5.4 中，中间的竖线代表了两个不同水平，为因变量电影评分差值的均值。而以均值为中点，左右两根竖线内的范围，则代表了该差值的 95%置信区间。图中有一条垂直的虚线，代表 0。可以看到，若该 95%的置信区间若包含了 0，则这两水平内薪资水平的差值是不显著的（例如 2000 年以前的电影与 2000-2010 的电影）；反之若不包含 0，则该差异显著（例如 2010 年及以后的电影与 2000 年以前的电影）。

5.5 本章小结

本章主要介绍了统计学中三个重要问题：参数估计、假设检验与方差分析。参数估计是通过收集到的样本信息，估计总体的某些参数。本章主要介绍了三种重要的参数估计方式：矩估计、极大似然估计与区间估计。其中，矩估计与极大似然估计是重要的点估计方式，区间估计则可以帮助获得参数估计的置信区间。接下来，本章介绍了假设检验的概念和用法。假设检验的基本步骤分为提出假设、选择检验统计量、确定拒绝域的形式和给出显著性水平。假设检验按照需比对的样本数目可以分为单样本检验和两样本检验；按照问题类型可以分为单边检验和双边检验。最后，本章还介绍了单因素方差分析，用于对两个或者两个组别以上样本均值差别的显著性检验，它关注的重点是对于组别差异的分析。

5.6 本章习题

1. 设 $x_1,\cdots,x_n$ 是来自于均匀分布 $U(a,b)$ 的一组样本，请推导 a,b 的矩估计。

2. 假设某产品在制造时分为合格品与不合格品两类，我们用一个随机变量X来表示某个产品经检查后是否合格，$X=0$表示合格品，$X=1$表示不合格品，则X服从两点分布$b(1,p)$，其中p是未知的不合格品率。现抽取 20 个产品，看其是否合格，得到样本如下所示：

1 0 0 0 1 0 1 1 1 0 1 0 0 1 0 0 1 1 1 1

求p的最大似然估计，并在 R 中进行实现。

3. 假设灯泡的寿命服从正态分布。为了估计某种灯泡的平均寿命，现随机地抽取 15 只灯泡测试，得到它们的寿命（单位：小时）如下：

998　1021　988　986　1018　997　996

973　925　985　981　1007　1011　968　1002

试求灯泡平均寿命的 0.95 置信区间，并在 R 中进行实现。

5.7 参考答案

1. 根据总体均值和总体方差的表达式，使用样本的前二阶矩估计总体的均值和

方差。

2. 在这个问题中，似然函数为：

$$L(p) = P(X_1 = x_1, \dots, X_n = x_n; p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

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

通过令对数似然函数导函数为 0，求解最大似然估计的表达式。两点分布的参数p的最大似然估计和矩估计结果相同。

3. 可以通过公式法计算灯泡平均寿命的 0.95 置信区间，也可以使用 R 中的‘t.test()’函数进行计算。