R 统计分析

杨春生

献给

我的同事和学生

目录

表格

插图

前言

这是专门为我的同事和学生写的一本书。我的同事们统计学表现出浓厚的兴趣，但他们没有接触过 R 软件，不知道怎样用 R 来处理统计数据；我的学生则对数据处理心怀畏惧，他们觉得统计学理论比较难，而且也没有任何数据处理的软件操作经验。因此，本书的内容不仅要包括统计学的基本理论，还包括 R 软件的基本操作。但真正让我下决心写这本书的，却是和一个我不认识的学生的一次谈话。她告诉我学习 R 软件的艰辛，以及用 R 软件撰写统计报告的困难。我震惊于她使用 R 软件的方式，更觉得要改变利用 R 软件进行统计学教学中存在的理念上的误区。很多人利用 R 软件讲授统计学时，主要是基于 R 软件的免费、开源性质，却不知道 R 软件强大的文档处理能力。因此，利用 R 软件进行统计学教学，可以将统计学的基本原理、R 软件的统计分析和撰写实验报告有机的结合起来，我将其称为“一体化”教学模式。

在一本书中涵盖统计学的基本原理、R 软件的统计分析编程以及利用 R 撰写动态统计报告等内容，本身就是一个非常大胆的想法，但并不意味着它不能实现。本书就是这种尝试的结果。

本书是 R 软件系列书中的第一本，我将其命名为《R 统计分析》，同期写作的还有《R 计量分析》，只不过《R 统计分析》是利用谢益辉博士提供的 **bookdown** 模板写作；《R 计量分析》则是利用 **R**+LaTeX 的方式写作，它的方式可能更加自由。在进行这两本书写作的同时，还需要处理 **Wrod** 文档，因此，整个暑假，我需要不停的在三种写作方式之间切换，真是一个让人“神经错乱”的暑假。

本书的主要内容包括三个部分：第一部分是基础部分，

和现有的统计学教材相比，本书有三个方面鲜明的特点：

1. 利用 R 软件的动态报告撰写功能，实现统计学教学中的理论教学、实践操作和结果展示的 “一体化”。传统的统计学教学中理论讲授、实践操作和结果展示是相对独立的三个过程，导致统计学的教学效率相对低下。即使是现有的利用 R 软件的统计学教材，仍然没有摆脱传统教学模式的束缚。本教材和其他教材不同之处在于，充分利用 **rmarkdown+project** 的动态统计报告和项目管理功能，实现三者的有机结合。使得统计报告的撰写不再是让人厌烦的复制、黏贴和格式调整等文档操作，而是确定并组织其核心内容的过程。

2. 利用了 R 软件中自带的大量数据，使得本书呈现的数据处理过程更加贴近现实。统计学的核心技能是对数据进行分析。但数据分析的基础技能则是数据的搜集、整理和处理等过程。一篇完整的统计报告中，撰写统计理论，统计分析过程和展示统计结果的工作大致占整个统计过程的 5% 的工作量；其余 95% 的时间都用于数据的搜集、数据的整理和处理。即使是利用二手数据进行统计分析，其数据的搜集、整理和处理过程所占的时间也不少于 80%。但这一过程在现有的统计学教材中却没有得到充分的重视。它们所提供的数据处理案例往往和现实无关。R 软件自带的大数据，往往来源于现实的调查，因此，利用这些数据讲授数据的整理和处理过程，无疑具有很强的现实针对性。

3. 本书中所有的统计分析过程的编程都是可重复的。和其他菜单式操作软件、或采用命令行方式运行的统计软件不同，R 软件是一款编程式的统计软件。因此，本书中所有的统计分析过程，都采用编程的方式进行。因此，本书中所有的源代码都利用 R 软件的 **chunk** 进行编码，在其后则展示编码的结果。我们可以直接将源代码复制到 R 软件的脚本文件中去，运行脚本文件即可得到相应的结果。不仅如此，这些源代码还可以使用与其他数据的数据处理过程。当然，本书作者推荐的方式并非如此。本书作者认为，若想对 R 软件的编程有一个大致的了解，最好的办法就是将本书中所有的程序逐字逐句地敲入电脑中。

本书的服务对象是统计学的初学者，因此，本书假定它们并不知道统计学的基本理论，以及统计软件的使用经验，更不要求它们掌握 **rmarkdown** 的基

本语法。当然，即使是已经掌握统计学理论，并具有一定的 R 软件操作经验的读者，也会对本书的动态统计报告感兴趣。

在本书的写作和编译过程中，作者使用了 R 软件的 **rmarkdown**、**knitr** 等包，本书写作的 R 软件的版本信息可以利用

```
sessionInfo()
```

命令查看。

致谢

本书作者首先要感谢安徽大学的荣兆梓教授。本书作者在遇到恩师的时候，唯一可以养活自己的技艺就是开拖拉机，是荣教授将我带进了丰富多彩的经济学的世界。当然，这本书是我“跑偏”的一个证据。所以我不知道恩师在看到这本书的时候是高兴还是生气。不过，按照恩师的理论高度，我已经无法做到见于师齐，只能根据自己的爱好，在 R 软件里面下功夫，做一片不一样的树叶，或许可以寻找到自己存在的价值；另外，我要感谢武夷学院的梁丽萍教授和黄冈学院的李新光副教授，他们在本书的写作过程中都给出过很多中肯的意见；感谢我的学生袁泽华对本书的编写投入极大的热情，不仅参与书稿结构的讨论，还撰写了 R 软件作图部分的内容；感谢谢益辉老师提供的 **bookdown** 模板，使得撰写书稿的过程可以只专注与内容，不需要对本书的格式进行调整。虽然该模板仍然存在一些细微的问题，但最起码，它可以让书籍撰写成为一件让人感到愉快的事；最后，还要感谢经管学院的领导和同事，占院长对本书的写作相当的关心，在我写书过程中给予我最大的自由，科研院长郝西文在本书写作过程中不仅从专业方面给出了中肯的意见，还帮我处理了很多琐事，让我在本书写作过程中始终保持愉快的心情。也要感谢经管学院的学生，本书中的大部分内容在经管学院的统计学教学过程中进行了将近三年的试讲，感谢我的同事和学生对程序中很多错误的包容，也感谢他们对这本书的校对。当然，本书的所有错误和他（她）们无关。最后，我还要感

谢我的挚友雷英德，虽然他没有给本书提出任何具体的建议，但和他的数次交谈让我明白了一个实际从事统计工作的人最需要哪些基本的统计技能。当然，本书和下面的一首诗也献给他，以此纪念我们那随着岁月流逝仍然不会褪色的友谊。

无题

风雨不弃血泪篇，蒯缑壁上悲华年；

豪气仗剑分三晋，白发吹箫过五关。

千金不为相如赋，一意独守尾生言；

最是庐阳伤别梦，兰台走马莫寻欢。

杨春生

二 0 二一年八月

第一章　统计概述

在日常生活中，统计是一个自带流量的词汇，报纸、期刊或杂志有之，网络、电视则更多。本章主要介绍统计学的定义和应用、统计学中的基本概念、统计数据和统计软件等。

1.1　什么是统计学

1.1.1　统计学的定义

统计学（**statistics**）是搜集、处理、分析、解释数据并从数据中得出结论的科学。

从统计学的定义可以发现，统计学提供的是一套关于搜集数据、处理数据、分析数据的方法。在我们的日常生活、学习和工作中，我们会面对各种数据。大数据时代，尤为明显。而统计学则提供了对这些数据进行搜集、处理和分析的工具。一言以蔽之，统计学是一门让数据“说话”的科学。例如，下面的现象，就需要通过数据来“讲故事”。

- 虽然我们都觉得吸烟有害健康，但我们印象中吸烟的人和不吸烟的人的寿命好像差距并不大；
- 有人经过调查得出的结论是左撇子的平均寿命低于使用右手的人的平均寿命；

1.1.2 统计学分类

按照统计学在进行数据分析时所用的方法，可以将其大体上分为**描述性统计学**和**推断性统计学**。它们也是我们通常所说的统计学的两大分枝。

描述性统计学（**descriptive statistics**）是研究数据的收集、处理和描述的统计学方法。其内容包括：

- 如何取得研究所需的数据；
- 如何对所取得的数据进行总体上的描述；
- 如何使用统计图或统计表更直观的展示数据。

推断性统计学（**inferential statistics**）则研究怎样利用样本数据来推断总体特征的统计学方法。其基本的内容主要包括两类：

- 区间估计
- 假设检验

区间估计是利用样本的数量特征来推断总体的数量特征。例如，从一批灯泡中随机的抽取一定数量的灯泡，测出这批灯泡的平均使用寿命，然后，利用它去推断这批灯泡的平均使用寿命。假设检验则是利用所得到的样本信息，来判断对总体的某个假定是否成立。例如，若随机的从一批灯泡中抽取一定数量的灯泡，测出它们的平均使用寿命，然后根据这一数据来判断关于这批灯泡的平均使用寿命的假定是否成立。

描述性统计学和推断性统计学是统计学的两大分枝。

1.1.3 统计学的应用

一门学科之所以被称为一门学科，是因为它有自己固定的研究对象。例如，物理学研究的是自然现象的运动规律；化学研究的是物质的组成、结构、性质及其变化规律；生物学则研究生命现象及生物活动规律。

统计学也是一门独立的学科，这一点似乎没有人会怀疑，但如果要问，统计学的研究对象是什么？答案则五花八门。例如，有人认为，统计学是一门独特的学问，没有固定的研究对象。即统计学是一门工具性的学科，需要和其他学科相结合，才能进行研究。当然，从经济学的视角出发，统计学所研究的是经济现象的数量特征及其变化的规律。所以，在这门课程的学习中，我们经常会碰到 **GDP**、**人均 GDP**、**通货膨胀率**以及**失业率**等一系列的概念。而统计学对这些经济现象进行研究的最终目的，就是提供一套在经济科学领域通用的搜集数据、处理数据并从数据中得出结论的原则和方法。

由于统计学研究的是经济现象的数量特征及其变化规律，因此，在研究过程中，必然涉及到对数据的分析。统计学对数据进行分析的原则与方法，不仅适用于对经济现象的研究，也适用于对社会现象，甚至是自然现象的研究。所以，统计学具有工具性学科的特征，即它可以和其他学科相结合，形成新的交叉性学科。仅仅在经济领域，统计学就可以应用有经济、金融、保险、审计、市场营销以及企业管理等领域。同样，统计学也可以应用于社会科学和自然科学中的各个领域，由此可见，将统计学看成是一门工具性学科是有一定道理的。

统计学的工具性，决定了其应用的广泛性，但这并不意味着统计学是万能的。例如，在经济领域，引用统计学虽然可以分析经济变量之间的相关性，但对变量之间的因果关系，统计学却无能为力，需要应用经济理论进行定性分析。

1.2 统计中的基本概念

统计学中的概念众多，其中，最常用的概念包括：总体和个体、变量和数据、抽样和样本以及参数和统计量等。在本节中，我们将对这些概念作一个系统的介绍。

1.2.1 总体和个体

总体（**population**）指的是所有考察对象的全体。

个体（**unit**）指的是组成总体的每一个考察对象。

在这里面需要注意的是，总体中的个体虽然各有特色，可是，既然它们能够组成一个总体，说明它们必然在某一方面具有共同的特征，即我们常说的同质性。例如，作为个体的每一个中国人，我们虽然在年龄、性别、民族、宗教信仰和受教育水平方面各不相同，但作为总体的中国人，我们有一个共同的特征，那就是我们一般认为自己是 “炎黄子孙” 或 “龙的传人”。

在实际调查研究中，有时候总体的范围很容易确定。例如，要检验一批灯泡的使用寿命，则这批灯泡所构成的集合就是研究的总体，而每一个灯泡就是个体。但有些时候，总体的确定却相当复杂，例如，农夫山泉想知道它所生产的一款新产品——Scream 系列饮料，在某一地区是否受到消费者的喜爱。在这个案例中，首先要确定该饮料的消费者总体。但事实上，我们很难确定哪些消费者购买了该饮料。此时，我们可以根据研究的目的确定总体。假定我们认为，只有青年人才是饮料的购买主体，那么，我们就可以将该研究的总体固定为该地区所有的青年人。

根据总体中所包含的单位数量是否可数，可以将总体分为有限总体和无限总体。有限总体是指总体的范围能够明确确定，而且总体中元素的数量是有限可数的。例如，在检验灯泡使用寿命的实验中，这批灯泡的数量就是有限可数的；若总体中包含的元素的数量是无限的、不可数的，那么这个总体就是无限总体。例如，我们想研究抛一枚硬币，其正面朝上的可能性，那么，每一次实验的结果可以看成是一个个体，由于该实验可以无休止的进行下去，因此，该实验结果构成的总体就是一个无限总体。

最后，需要进一步说明的是，由于统计学研究的是经济现象的数量特征，因此，在实际调查中，统计学关注的是总体在某一个方面的数量特征，而不是总体本身。在灯泡使用寿命检验中，我们关注的是灯泡的使用寿命而不是灯泡本身，所以，在该检验中，我们也可以把这批灯泡的使用寿命的集合当作总体，把每一只灯泡的使用寿命看成是个体。由此可以看出，统计学意义上的总体，通常不是一群人或物品的集合，而是一组观测数据的集合。

1.2.2 变量和数据

变量（**variable**）指的是描述总体中单个个体的某种特征的量。变量的特点在于变化，即从一个个体到另一个个体，这种特征会发生改变。例如，观测一只股票的价格，会发现今天的股价和昨天、前天、$\cdots$ 的股价都不尽相同；观测一个企业的销售额，会发现这个月的销售量和上个月、上上个月、$\cdots$ 也都不一样；如此等等。

数据（**data**）指的是对变量的观测结果。

按照不同的标准，可以对变量和数据进行分类。

根据观测结果的特征，可以将变量分为类别变量和数值变量两种。

类别变量（**catigorical variable**）也称分类变量或定性变量，是取值为个体属性、类别或区间值的变量。例如，对个体的性别进行观测，得到的结果为“男” 或 “女”；用户对某种商品的评价，一般可以划分为 “很好”、“好”、“一般”、“差” 和 “很差” 等几类；在上面的观测中，无论是人的性别还是对商品的评价，其结果都不是数字，而是个体的属性或类别。另外，按照人的年龄可以对人群进行分类。例如，一般将 18 周岁以下的人称为 “未成年人”，将 18 周岁到 45 周岁的人称为 “年轻人”，将 45 到 60 岁的人称为 “中年人”，将 60 岁以上的人称为 “老年人”，等等。在这个观测中，变量的观测结果也是类别变量，但其年龄取值却是一个区间。根据类别变量的取值是否有序，可以将其分为名义值类别变量和顺序值类别变量两类。名义值类别变量也称无序类别变量，其观测结果是不可以排序的。最典型带动无序类别变量有：性别、宗教信仰、籍贯、民族以及企业的行业属性等，这些变量的取值之间不存在顺序关系。顺序值类别变量也称有序类别变量，即其观测结果可以排序。例如，对商品的评价中，其观测结果按照 “很好”、“好”、“一般”、“差” 和 “很差” 的结果排列，说明用户对商品的评价是逐步降低的。当类别变量的观测结果只有两种可能时，我们将其称为二值类别变量。人的性别就是一个典型的二值类别变量。

数值变量（**metric variable**）指的是取值为数字的变量。数值变量又被称为定量变量。例如 “产品的产量”、“股票的价格”、“生活费支出” 等，都是常见

的数值变量。

数值变量按照其取值的不同，可以分为离散变量和连续变量。其中，离散变量指的是只能取有限值的变量，而且其取值可以一一列举。例如，“产品的产量”、“年龄” 和 “年末人口数量” 等都是离散变量；连续变量可以在一个或多个区间中取任意值的变量，它的取值是连续的，因此无法一一枚举。常见的连续变量包括人的 “体重”、“身高” 以及 “温度” 等。

由于数据是变量的观测结果，因此，数据的分类基本和变量的分类一致，可以分为类别数据和数值数据两大类。

类别数据（**catigorical data**）指的是对类别变量的观测结果。类别数据也称为分类数据或定性数据。与类别变量相对应，类别数据相应的也分为名义值类别数据和顺序值类别数据两类，其中，只有两个观测结果的类别数据又被称为二值类别数据。

数值数据（**metric data**）指的是对数值变量的观测结果。它有时又被称为定量数据。

这里需要特别说明的是，为了数据处理的方便，我们一般都会将类别变量赋予一定的数值，这一过程就是数字化编码。例如，我们一般对二值类别变量的数字化编码为 “1” 和 “0”，在人的 “性别” 的数字化编码中，我们可以将 “男性” 数字化编码为 1，则 “女性” 的数字化编码就是 “0”。若 “民族” 进行编码，可以将 “汉族” 数字化编码为 1，“蒙古族” 数字化编码为 2，···。在对顺序类别变量进行编码时，则要遵循一定的规范，例如，对商品的评价结果的数字化编码，在确定编码的基准和间隔后，其编码也就确定了。若我们将编码的基准是将 “一般” 数字化为 0，编码的间隔为 1，则 “很好”、“好”、“一般”、“差” 和 “很差” 的数字化编码就分别为 “2”、“1”、“0”、“-1”、“-2”。

对类别变量进行数字化编码，只是其表现形式发生了变化，即取得了数字的 “外衣”，并非是变量的属性发生变化。

数据除了上面的分类方式外，还有其他的分类标准。例如，按照数据的收集方法不同，数据可以分为观测数据和实验数据。

观测数据（**observational data**）指的是通过调查或观测收集到的数据。这

类数据的获得过程中，没有对被观测的个体实施人为的控制，在自然状态下进行的。在绝大多数情况下，有关社会经济现象的数据都是观测数据。

实验数据（**emperimental data**）指的是在试验中通过控制实验对象而收集得到的数据。例如，比较两组种子不同的发芽率的实验，就要求这两组种子所面临的其他条件，如温度、湿度等都相同。一般而言，自然科学领域的大部分数据都是实验数据。

按照数据和时间之间的关系，可以将数据分为横截面数据、时间序列和面板数据三大类。

横截面数据（**cross-sectional data**）指的是在相同或相近的时间点上收集的不同个体的数据。这类数据是在不同空间上获得的，用于描述数据在某一时间上的空间分布状况。例如，2020 年世界各国的经济增长率就是横截面数据；

时间序列（**time series data**）指的是相同个体在不同时间点上收集到的数据。一般情况下，这类数据是按照时间的顺序收集的。例如，新中国成立 70 年，我国的 GDP 总量数据等，就是典型的时间序列数据。

面板数据（**panel data**）又被称为 “平行数据”，是指在时间序列上取多个截面，在这些截面上同时选取相同的个体而得到的数据。面板数据有两个维度，即个体维度和时间维度；纵向看，面板数据描述的是同一个体在不同时间上的数据表现，因此是一个时间序列；横向看，面板数据描述的是同一时间点上不同个体的数据表现，因此是一个横截面数据集。例如，中部六省从 2000 年到 2014 年的人均 GDP，就是一个典型的面板数据。

对数据进行不同的分类，在统计学的研究中是必要的，因为，不同类型的数据，在分析时采用的方法和手段会发生变化，例如，对数值变量，我们一般会求它的均值。但对于类别变量，我们一般会求它的比例；对多个数值变量，我们会进行相关性分析或回归分析，但对多个类别变量，我们更多的会采用列联分析和方差分析。一般而言，对数值型变量的分析方法更加多样，如计算各种统计量、进行参数估计和假设检验等。

1.2.3 样本和抽样

统计学所研究的是经济现象总体的数量特征，从理论上说，需要采用全面调查的方法，逐个统计每一个个体的数量特征，进行汇总。但在很多时候，我们无法做到这一点。其原因大致如下：

- 对于无限总体，采用全面调查的方法根本行不通；
- 即使是有限总体，采用全面调查的方法在很多时候在时间和财力方面支出巨大，显得没有必要。最典型的例子就是人口普查，一般 10 左右才进行一次。
- 有些时候，全面调查可能具有破坏性，使得调查结果没有任何意义。在灯泡检验中，对每一个灯泡的使用寿命进行统计，可以得到最精确的结果，但种种结果毫无意义。原因在于，这批灯泡在检验中全都损毁了。

因此，在实际的统计研究中，我们往往从总体中抽取一部分个体，通过对这些个体的数量特征的研究，来估计总体的数量特征。这里面就涉及到两个基本的概念，样本和抽样。

样本（**sample**）指的是从总体中抽取的一部分个体的集合。

抽样（**sampling**）指的是从总体中选取一部分代表性样本的方法。

在抽样过程中，若每个个体是否被抽中是完全随机的，而且其被抽中的概率相等，我们将这种抽样称为概率抽样。常见的概率抽样有简单随机抽样、分层抽样、系统抽样和整群抽样等。

简单随机抽样（**simple random sampling**）是从含有 N 个元素的总体中，随机地抽取 n 个元素组成一个样本。按照抽样过程中，被抽取的样本是否放回总体，我们可以将简单随机抽样分为重复抽样和不重复抽样。很显然，重复抽样是等概率抽样。但在无限总体中，我们也可以将不重复抽样看成为等概率抽样。

分层抽样（**stratified sampling**）也被称为分类抽样，即在抽样之前，先将总体划分为若干层（类），然后从各个层（类）中抽取一定数量的个体组成一

个样本。例如，想研究某大学学生的生活费支出，可以先将学生按照年级分层。然后，从各个年级中抽取一定数量的学生组成一个样本。分层抽样的优点是可以使样本分布在各个层内，从而增加样本的代表性。

系统抽样（**systematic sampling**）也称等距抽样，即在抽样之前，首先将总体中的各个个体按照某种顺序排列，然后随机确定一个起点，每隔一定的间隔抽取一个元素组成一个样本。例如，在上面的研究中，可以先按照学生的学号进行顺序排列，然后用随机数的方法找到一个起点，按照一定的间隔依次抽取个体，就得到一个样本。

整群抽样（**cluster sampling**）首先将总体划分为若干群，以群作为抽样的单元，然后随机抽取群，并对群内个体进行全面调查。例如，在上面的研究中，可以将学生按照其所住的宿舍划分为若干群，然后随机抽取一个或若干个宿舍，对被抽中宿舍的每一个学生的消费情况进行调查。一般而言，整群抽样要求个体在群间分布的差异小，而在群内分布的差异大。

另一个和样本相关的概念是**样本容量**（**sample size**）。它指的是通过抽样方法选取的样本中包括的个体数量。

1.2.4 参数和统计量

统计学的根本任务是研究经济现象的数量特征与变化规律。但在统计学的研究中，总体的数量特征往往是未知的。在这种情况下，就需要从总体中抽取一部分样本，利用这些样本的数量特征来推断总体的数量特征。这就涉及到另外两个相关的概念：参数和统计量。

参数（**parameter**）是描述总体特征的概括性数字度量，它是研究者想要了解的总体的某种特征值。一般情况下，我们将总体中未知的数量特征都称为参数。

统计量（**statistic**）指的是由样本确定的函数。在统计学中，常见的统计量包括样本均值和样本方差等。

在统计学中，一般情况下，总体的参数是未知的，但总体参数是不变的；而

样本的统计量则是已知的，但样本统计量却是可变的，会随着抽取的样本不同而发生变化。因此，在统计学中，一般情况下，都将统计量看成是一个随机变量。

下面，我们仍然以灯泡的使用寿命为例，来说明参数和统计量的含义，并进一步说明统计量的构造。

在灯泡寿命的检验中，它的目的是估计这批灯泡的使用寿命。由于每个灯泡的使用寿命都不尽相同，因此，用一个指标来概括这批灯泡的使用寿命，就需要有高度的综合性，这个指标就是这批灯泡的平均使用寿命。它的计算方法也比较简单，就是将每一个灯泡的使用寿命加起来，然后除以总体中的观测值的个数。但在现实中，由于该实验有极强的破坏性，我们采用的另外一种方法：首先，从总体中抽取一个样本，然后，确定样本中每个灯泡的使用寿命，并确定其平均寿命，样本中的灯泡平均使用寿命就是一个统计量。我们可以利用样本的平均使用寿命，来推断总体的平均使用寿命。

由此可以看出，统计量的构都来源于所抽取的样本，若抽取的样本不同，则得到的统计量也会发生变化。

1.3 统计软件

随着科学技术和生产力的发展，经济活动中的数据可得性越来越容易，因此，现实的统计分析中，数据的量非常大，统计分析中所用的方法也越来越复杂，对数据的可视化要求也越来越高。使用计算机和统计软件进行数据的处理和分析，已经成为一种潮流。

在计算机开始普及之前，统计分析中的计算问题使得统计学的应用受到了极大的限制。但随着计算机技术的发展，尤其是笔记本电脑的普及，统计学的教学发生了颠覆性的变化。对于统计学的学习者而言，复杂的计算问题可以由电脑完成，人们可以将更多的时间用于对统计分析的思想和原理的理解。

在当今，统计学的教学离不开电脑，更离不开统计软件。由于绝大部分的统计方法都由统计软件来完成，因此，学习统计学必须要精通一款统计软件。在

本节，我们将对常见的统计软件进行简单的介绍。

一般而言，常见的统计软件按照其进行统计分析的操作方式，可以分为菜单式统计软件，如 SPSS 和 Eviews 软件等；命令行式统计软件，如 SAS 和 Stata 等；编程式统计软件，如 Matlab 和 R 等。

1.3.1 菜单式统计软件

所谓菜单式统计软件，指的是该统计软件在进行统计分析时，主要通过下拉菜单的方式来进行。这类统计软件的代表是 SPSS 和 Eviews 等。

1.**SPSS**

SPSS 是统计产品解决方案 (Statistical Product and Service Solutions) 的缩写，也是世界上著名的标准统计软件之一。SPSS 软件的特点是简单、易懂且操作方便。另外，该软件具有较强的绘制图形和表格的能力，输出的结果也比较规范和直观，适合进行社会科学研究中的数据分析。

SPSS 软件很早就被引入到我国的统计学、尤其是医学统计学的教学中去。该软件的 Windows 版本目前已经有比较成熟的汉化版。使用该软件的用户，只要掌握一定的 Windows 系统的操作技能，了解统计分析的基本原理，就可以利用 SPSS 软件进行数据分析。利用 SPSS 软件，可以进行描述性统计学分析、区间估计、假设检验、相关分析、回归分析以及方差分析等。但 SPSS 软件的时间序列分析的能力相对较弱。另外，作为菜单式操作软件，SPSS 统计分析的效率相对较低。

2.**Eviews**

Eviews 软件是计量经济学观点（Econometrics Veiws）的缩写，是最常见的计量经济学软件包。在上世纪 80 年代中期，Eviews 软件由高铁梅老师从美国引进到我国的计量经济学课程的教学。和 SPSS 一样，Eviews 软件也是菜单操作软件。虽然 Eviews 软件主要应用于计量经济学中的模型估计、模型检验和模型的预测等，但该软件也可以用于统计分析，同样，Eviews 软件的

绘制图形功能也十分强大。另外，由于该软件上手比较快，操作也比较简单，因此，Eveiws 在统计分析中也相当的流行。

1.3.2 命令行式统计软件

命令行式统计软件，指的是在进行统计分析时，主要通过在其命令输入窗口输入相应的命令的方式来实现对数据的处理和分析。和菜单式统计软件相比，命令行式统计软件的下拉菜单的功能比较简单，无法通过菜单实现绝大部分的统计分析。因此，只能在命令输入窗口书仍然相应的命令，实现对数据的分析，并利用相应的命令，展示统计分析的结果。具有代表性的命令行式统计软件是 SAS 和 Stata。

1.**SAS**

SAS 是统计分析系统（Statistical Analysis System）的缩写，是在业界享有盛誉的统计软件。在数据处理和统计分析领域，SAS 是国际上的标准软件和最权威的应用软件。SAS 提供了能够从事统计分析、经济计量分析、时间序列分析、决策分析、财务分析和全面质量管理的工具。和菜单式软件相比，SAS 是一个由 30 多个模块构成的组合式软件系统。另外，SAS 是用汇编语言编写而成的统计软件，因此，使用 SAS 通常需要编写程序。但由于 SAS 的编程语言是非矩阵语言，因此，在进行统计分析时，SAS 编程比较繁琐，最好的方式仍然是逐条输入命令的方式来运行它。

2.**STATA**

STATA 与 SAS、SPSS 被共同成为三大权威统计软件，它被广泛地应用于统计学、经济学、生物学、医药学、社会学、人口学等等一系列学科的研究。

STATA 软件具有数据处理、绘图、统计分析、回归分析和编程处理这五大主要功能，其相互配合，可以完成系统完整的数据分析和处理任务。

与其他统计软件一样，STATA 具有正常的标题栏、菜单栏、工具栏和状态栏，由于 Stata 主要是通过命令进行操作，因此，这些工具栏的统计分析功能相对较弱。

和 Eviews 软件一样，STATA 也可以很方便的应用到计量经济学课堂教学中去。

1.3.3 编程式统计软件

编程式统计软件指的是采用编程的方法进行数据处理和统计分析的统计软件。和菜单式统计软件、命令行式统计软件相比，编程式统计软件的菜单栏的统计分析功能更弱。编程式统计软件的优点是数据处理和统计分析过程的效率更高，缺点是需要一定的时间才能初步掌握这类软件的使用。典型的编程式统计软件有 MATLAB 和 R。由于 MATLAB 主要用于数值计算领域，在统计分析中应用并不是很广泛，这里就不再对该软件进行介绍。

R 软件

R 软件是基于 S 语言的一款免费、开源的软件。R 软件提供了相当丰富的数据分析技术，功能十分强大。和其他软件相比，R 软件的更新速度更快、使用更加灵活，而且包含了很多最新的统计实现方案。另外，R 软件的绘图功能是其他任何软件都无法比拟的。由于 R 软件具有强大的功能，且使用是分灵活，因此，在实际的统计分析和研究中，R 软件正在被越来越多的人使用。可以说，R 软件最终必然会成为统计分析软件的主流。[1]

课后练习

一、思考题

1. 举例说明无序类别变量和有序类别变量；
2. 概率抽样的方法有哪些？

二、简答题

[1]这里写得比较乱，希望后面有时间修订它。

1. 数据的类型有哪些？

2. 统计学有哪些应用领域？

第二章　R 软件基础

本章主要内容：

- R 软件的下载、安装和设置；
- R 软件中数据的种类；
- R 软件的基本运算。

2.1　R 软件的下载、安装和设置

为了使用 R 软件进行统计分析，首先，必须下载该软件；然后，需要在自己的电脑中安装 R 软件；安装完成后，在使用 R 软件前，还需要对 R 软件进行正确的设置。

2.1.1　R 和 Rstudio 软件的下载

1. **下载 R 软件**

在 **CRAN** 网站http://www.r-project.org/（该网站就是我们所说的 R 软件官网。）可以下载最新的 R 软件版本。进入 R 软件官网，如图2.1所示：

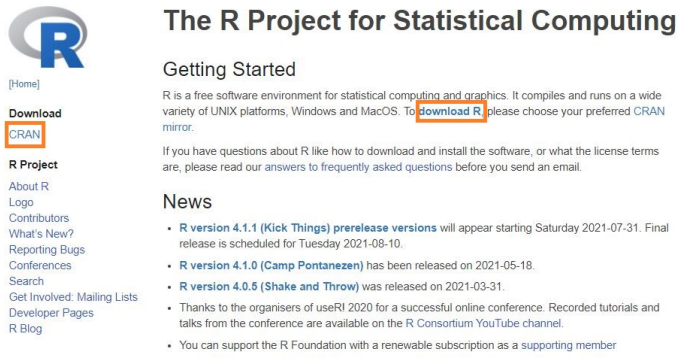


图 2.1: R 软件官网

在 R 软件官网中，点击网页上的 **downloadR** 或左侧栏上的 **CRAN**，即可进入下载 R 软件的镜像文件网站。在镜像文件网站中，找到 **China**，如图2.2所示：

图 2.2: 镜像网站设定

由图可以发现，在 **China** 下面，总共有 10 个镜像网站可供选择。可以根据自己所在的地区，选择离自己最近的镜像网站下载 R 软件。本书作者选择的是位于合肥的中国科学技术大学的镜像网站。也就是说，我们将从中科大的镜像网站下载 R 软件，其速度远远快于从官方网站下载的速度。

设定好下载 R 软件下载的镜像网站后，页面会自动进入下载 R 软件的版本

选择页面，如图2.3界面，此时，我们需要根据电脑的操作系统来选择所下载的 R 软件的版本。当然，对绝大多数人来说，使用的是 Windows 操作系统，因此选择 **Download R for Windows**。

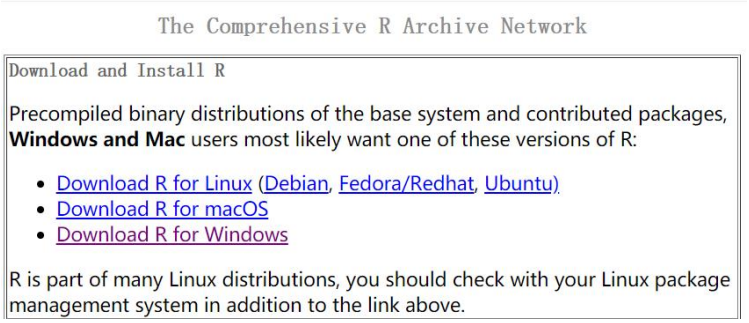


图 2.3: R 软件的版本选择页面

此时，页面会自动进入 Windows 版本的 R 软件下载页面，在此页面中点击 **install R for the first time**，即可进入最新版本的 R 软件下载页面，如图2.4所示，点击即可自动下载最新版的 R 软件。

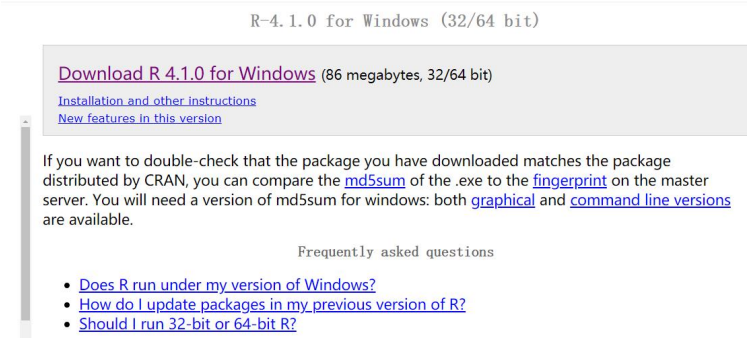


图 2.4: 最新版本的 R 软件下载页面

2. 下载 RStudio 软件

虽然 R 软件的功能相当强大，但 R 软件的使用界面相当简单，而且很不友好。例如，在 R 的命令输入窗口，就没有命令的自动补全功能，其实，R 软

件的使用界面很不友好。

为了更高效的使用 R 软件，需要下载一个 R 软件的 **IDE**(Integrated development environment 的缩写，意思是集成开发环境，其实是 R 软件的一个“外壳”。)。R 软件最好的 **IDE** 就是 **RStudio** 软件。同样，RStudio 软件也是一款免费开源软件，可供人们下载使用。

下载 Rstudio 软件的网址是：https://www.rstudio.com/。点击该网址进入 RStudio 软件下载的官方网站，如图2.5所示：

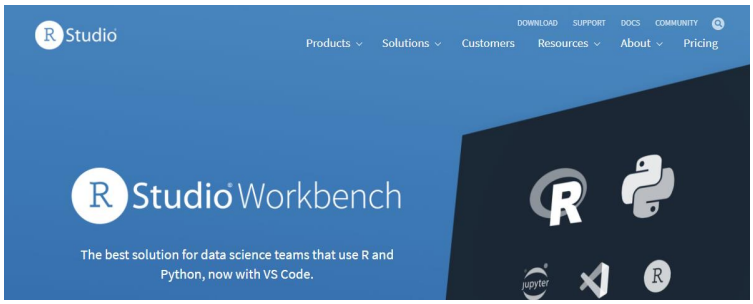


图 2.5: RStudio 官网

在 RStudio 官网首页，选择网页工具栏上的 **Products**，在其下拉菜单中点击 **RStudio**，浏览新弹出的网页，找到 **RStudio Desktop** 版本的 Rstudio，点击它下面的 **DOWNLOAD RSTUDIO DESKTOP**，即可进入 RStudio 软件的下载页面，如图2.6所示：

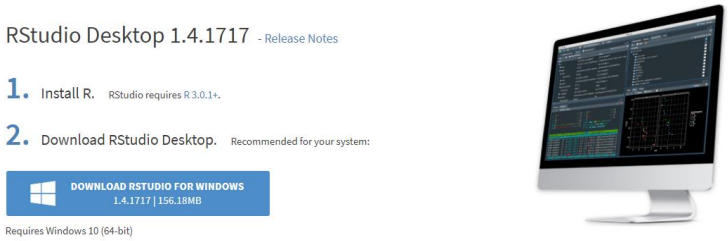


图 2.6: RStudio 下载页面

点击图中的下载链接，即可下载 RStudio。

2.1.2 R 和 Rstudio 软件的安装

在 R 软件和 RStudio 软件下载完成后，就可以安装它们。一般情况下，先安装 R 软件，再安装 Rstudio 软件。R 软件的安装比较简单，点击下载的.exe 文件，启动该文件的安装程序。选择中文为软件的安装语言，将 R 软件的安装路径选择在 D: 盘根目下一个名为 R 的文件夹中。剩下的全部选确定，即可顺利安装 R 软件。

可以用同样的方法安装 RStudio 软件，在安装时，将该软件的安装路径选择为 D: 盘根目录下一个名为 RStudio 的文件夹。其余的全部选择确定，即可顺利安装 RStudio 软件。

这里需要特殊说明的是，第一、在安装 R 软件和 RStudio 软件时，安装的路径最好的全英文路径；第二、安装 R 软件和 RStudio 软件的电脑的用户名最好是全英文的。若以中文作为用户名，即使能够安装成功，在使用过程中仍然会发生各种错误。

一般情况下，RStudio 软件安装完成后，会在桌面上出现一个快捷方式，若桌面上没有显示该快捷方式，可以在 RStudio 软件的安装路径 D:/RStudio 中的一个名为 bin 的文件夹中找到一个名为 rstudio.exe 的文件，利用该文件可以在桌面创建一个快捷方式。点击该快捷方式，即可启动 RStudio 软，其结果如图2.7所示：

由图2.7可以看出，R 软件的工作界面大致可以分为三个区域，最左边的为命令的输入和结果输出窗口。该窗口的主要功能是输入命令，并将结果展示出来；右侧靠上的窗口，主要存储输入到 R 软件中的各种数据；右侧靠下的窗口主要功能如下：

1. 展示 R 软件当前工作路径下的所有文件；
2. 展示 R 软件绘制的所有图形；

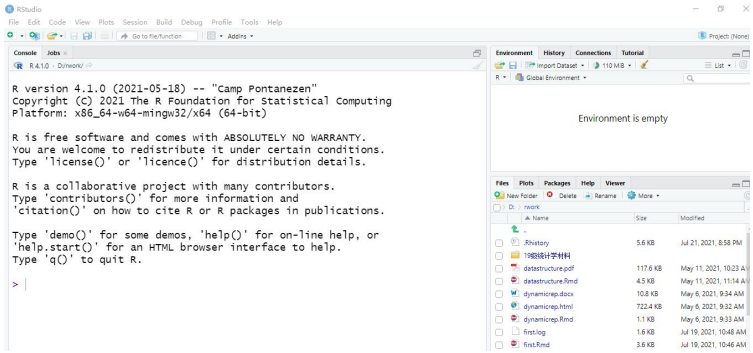


图 2.7: R 软件的启动界面

3. 展示 R 软件当前安装的所有的包；

4. 展示 R 软件相关命令的帮助内容。

或许，你已经迫不及待，好吧，然我们先输入一个简单命令。在左侧的命令输入窗口，有一个 >，将光标移到该符号的右侧，即可输入相关的命令。

让我们输入一个最简单的命令：

```
2+3
```

看看 R 软件有什么变化吧。现在，我们终于可以用 R 软件来进行统计分析了。当然，前面的路途仍然漫长。

2.1.3 R 和 Rstudio 软件的设置

虽然，R 软件和 RStudio 软件安装之后即可使用，但是，为了更加方便、高效地使用 R 软件，在正式使用之前，我们还需要对 R 软件进行必要的设置。这些设置包括：

- R 软件的工作路径设置；

- R 软件的镜像网站设置；
- R 软件的编码保存设置；

1. R 的工作路径设置

R 软件的工作路径（working directory）是 R 用来读取文件和保存结果的默认目录（文件夹）。我们可以用

```
getwd( )
```

命令来查看我们当前的工作路径。一般情况下，首次安装 R 软件，其工作路径都位于 C: 盘 users 下面一个名为 administrator 的文件夹中。为了更方便地使用 R 软件，我们可以自己设定 R 软件的工作路径。本书作者就将 R 软件的工作路径设定在 D: 盘根目录下一个名为 rwork 的文件夹中。其方法如下：

1. 在 D：盘新建一个名为 rwork 的文件夹；
2. 点击 R 软件菜单栏中的 **Tools**，在其下拉菜单中选择 **Global Options** 进入 R 软件的设置界面，如图2.8所示：

可以发现，在图2.8中，R 软件的工作路径为 C: 盘 uers 文件夹中一个名为 administrator 的文件夹。

3. 点击右侧的 **Browse**，就可以将该路径设定为 D: 盘的一个名 rwork 的文件夹。点击下面的 **OK** 设定完成。

完成设定后，可以先关闭 R 软件，然后再重新启动 R 软件，在其命令输入窗口输入：

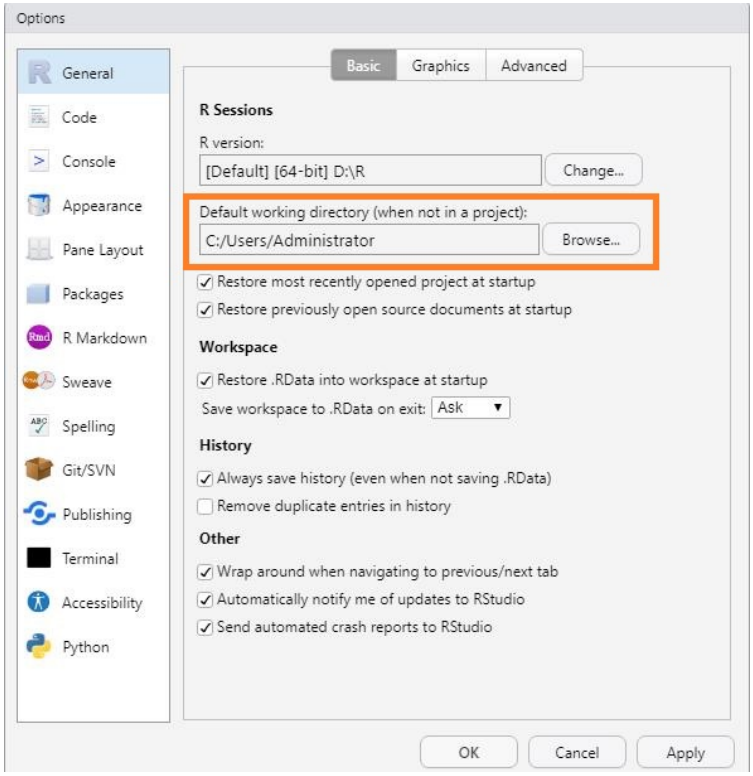


图 2.8: R 软件的设置界面

```
getwd()
```

即可以看到，我们的工作路径发生了改变。在后面的章节，我们会发现，还可以利用项目来重新设定 R 软件的当前工作路径，这里不再赘述。

2. R 的编码保存设定

由于 R 软件是一款编程软件，因此，利用 R 软件进行统计分析的时候往往需要编程。R 软件对这些程序文件的编码保存格式有一定的要求，即所有的代码必须以 UTF-8 的形式保存。为此，我们需要对 R 软件的编码保存进行

重新设定。重新设定 R 软件程序文件的编码保存方式的操作如下：

1. 点击 R 软件菜单栏中的 **Tools**，在下拉菜单中选择 **Global Options**，进入 R 软件的设定界面，如图2.8所示，
2. 在图2.8中，点击最左侧的 **Code**，进入 R 软件的编码保存设置页面。
3. 在该页面，点击右侧上方的 **Saving** 选项，在其弹出的界面中，在 **Default text encoding** 下面的选项右侧的 **Change**，在弹出的选项中选择 UTF-8 选项。
4. 分别点击下面的 **Apply** 和 **OK**，就可以完成 R 软件的编码保存设置。

3. R 的镜像网站设定

镜像，原意是光学里指的物体在镜面中所成之像。引用到电脑网络上，一个网站的镜像是指对一个网站内容的拷贝。镜像通常用于为相同信息内容提供不同的源，特别是在下载量大的时候提供了一种可靠的网络连接。制作镜像是一种文件同步的过程。“镜像网站”(英文:Mirror sites)，又译作 “镜像站点” ，亦即把一个互联网上的网站数据” 拷贝” 到本地服务器，并保持本地服务器数据的同步更新。R 软件在全球都设有镜像网站，为全世界 R 软件的使用者下载 R 软件，安装 R 软件的包提供了巨大的方便。R 软件的使用者可以根据自己所在的地域来设定 R 软件的镜像站点。设定 R 软件的镜像站点操作如下：

1. 点击 R 软件菜单栏中的 **Tools**，在其下拉菜单中点击 **Golobal Optionns**，进入 R 软件的设定界面；
2. 点击图2.8中左侧的 **Packages**，进入 R 软件的镜像网站设定界面；
3. 在新弹出的镜像网站设定界面，点击 **Primary CRAN repository** 下方右侧的 **Change**，选择离自己最近的镜像网站。
4. 点击设定界面下方的 **OK**，完成 R 软件的镜像网站设定。

当然，R 软件还有很多设置，如外观设置、字体设置等，这里不再一一说明。但完成上面设置后，会让 R 软件使用起来更方便。

2.2 R 软件中的数据

作为统计软件，其首要功能就是进行数据处理。在统计实践中，不同的统计软件的数据格式与种类不尽相同。因此，学习一款统计软件，首先就必须了解在该软件中，数据是如何被组织起来的。

在 R 软件中，所有的结果都被以对象的形式保存起来，同样，R 拥有许多保存数据的对象类型，它们包括：标量、向量、矩阵、数组、数据框和列表。其中我们着重要掌握的有向量、矩阵和数据框。本节将对它们进行详细的介绍。

2.2.1 标量和向量

向量是用于存储数据的一维数组。可以用 `c()` 命令来创建一个向量。例如，在 R 软件的命令输入窗口的命令输入符号后面，分别输入下列三条命令，就可以得到三个不同的向量。

```
a <- c(1,2,3)
b <- c("a","b","c","d")
c <- c(T,F,F,F,T)
```

这里需要注意的是，在 R 软件中，等号的类型有三种，第一种是最常见的赋值运算：< −，它表示的是一种赋值运算，例如上面的第一个命令的含义是将 1、2、3 赋值给向量 a。第二种则经常出现在 R 软件命令的选项设定中，例如：

```
d <- matrix(1:20,nrow=4,byrow=T)
```

第三种则出现在逻辑判断中，例如，下面的运算：

```
e <- 3==5
f <- 3!=5
```

上面命令的含义是，判断 3 和 5 是否相等，并将结果赋值给对象 e；判断 3 和 5 是否不相等，并将结果赋值给 f。因此，上面两个命令的结果分别是 FALSE 和 TRUE。

在输入上面命令后，我们可以用 `ls()` 命令来查看 R 内存中的对象由哪些：

```
ls()
```

```
## [1] "a" "b" "c" "d" "e" "f"
```

在 R 软件的内存中，其对象包括 a、b、c、d、e 和 f。除了 d 我们暂时不知是什么外，其他的都是一个向量，其中，e 和 f 中仅有一个元素，就是我们常说的标量。

在 R 软件中，按照向量的取值形式，可以将其分为三类：数值型向量、字符型向量和逻辑型向量。在上面的向量中，a 是一个数值型向量，b 是一个字符型向量，c 是一个逻辑型向量。

这里需要注意的是，一个向量中的所有元素的类型或模式必须相同，同一向量中无法混同不同类型的数据。可以用 `class()` 命令或 `mode()` 命令来查看对象的类型或模式。可以在 R 软件中输入下列命令，并查看其结果：

```
clss(a)
class(b)
class(c)
```

和向量密切相关的另一个命令是 `length()`。该命令可以查看向量的长度，即向量中观测值的个数。例如下面的命令，生成了一个长度为 100 的随机数向量，并利用 `length()` 命令查看该向量的长度。

```
g <- rnorm(100)
length(g)
```

```
## [1] 100
```

2.2.2 矩阵

矩阵是一个二维的数组。在 R 软件中，可以用 `matrix()` 命令来生成一个矩阵。例如：

```
d <- matrix(1:20,nrow=4)
```

该命令生成了一个 4×5 的矩阵 d。若想看看 d 的具体结构，可以在命令输入窗口输入

```
d
```

可以得到如下结果：

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    5    9   13   17
## [2,]    2    6   10   14   18
## [3,]    3    7   11   15   19
## [4,]    4    8   12   16   20
```

利用 matrix() 命令生成矩阵时，其中的元素默认是按照列进行填充，若希望其中的元素按照行进行填充，则可以将其中的选项 byrow 设定为 TRUE。[1]例如：

```
e <- matrix(1:20,nrow=4,byrow=TRUE)
```

该矩阵的具体结构如下：

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    2    3    4    5
## [2,]    6    7    8    9   10
## [3,]   11   12   13   14   15
## [4,]   16   17   18   19   20
```

通过比较，就可以发现矩阵 d 和矩阵 e 中的元素的填充方式差异。

我们可以用 mode() 命令来查看一个矩阵的类型，用 dim() 命令来查看矩阵的维度。请自行输入下列命令，并查看其结果：

```
mode(d)
dim(d)
```

在这里需要注意的是，在矩阵中，每一个元素的取值类型也必须一样，同一矩阵中无法混同不同类型的数值。

[1]或者更简单地将其设定为 T。

2.2.3 数据框

由于数据框的不同列可以容纳不同类型的数据，因此，它是 R 软件中最常见的数据结构。在 R 软件中，可以用 data.frame() 命令生成一个数据框。

假定对四个学生的数学、语文和英语成绩的观测结果如下：

```
math <- c(80,85,75,95)
Chinese <- c(80,85,70,98)
English <- c(75,85,65,100)
grade <- c("中等","良好","及格","优秀")
```

由此，我们得到了三个向量。利用这三个向量，可以得到一个数据框。其编程和结果如下：

```
score <- data.frame(math,Chinese,English,grade)
score
```

```
##   math Chinese English grade
## 1   80      80      75  中等
## 2   85      85      85  良好
## 3   75      70      65  及格
## 4   95      98     100  优秀
```

从 R 软件展示的结果可以发现，在数据框的前面，有一列有序的数组，表示的观测对象。当然，在实际数据收集中，这四个行的名称应当是四位同学的姓名。若这四位同学的姓名是已知的，我们可以用它来表示行的观测对象的名称。

```
rownames(score)<- c("张三","李四","王五","马六")
score
```

```
##      math Chinese English grade
## 张三   80      80      75  中等
## 李四   85      85      85  良好
## 王五   75      70      65  及格
## 马六   95      98     100  优秀
```

因此，数据框的结构和我们的分数册比较相似。其第一列表示的是被观测对象的名称，后面是变量的观测结果。从该例子可以看出，数据框和矩阵的不同，在矩阵中，各列的数值类型应当是相同的，但在数据框中，各列的数值类型可以不相同。例如，在数据框 score 中，前三列都是数值型数据，第四列是字符串型数据。

我们可以分别用 class() 命令、dim() 命令、rownames() 命令和 colnames() 命令来查看一个数据框的数据结构、维度、行的名称，列的名称。请自行输入下列命令，并查看其结果：

```
class(score)
dim(score)
rownames(score)
colnames(score)
```

关于向量、矩阵和数据框中的数据管理，我们将放到后面的章节进行，这里不再做更多的说明。在下一节中，我们将进一步学习利用 R 软件进行简单的数值运算。

2.3 R 软件的基本运算

作为统计软件，必须能够胜任各种基本运算，R 也不例外。R 软件的基本运算大致可以分为赋值运算、代数运算和编程运算。我们将利用最简单的例子对此进行说明。

2.3.1 R 软件的赋值运算

所谓赋值运算，就是将具体的数值赋予某个变量。例如

```
a <- 3
b <- 4
```

在上面的命令中，R 软件给变量 a 和 b 分别赋值，使得变量 a 的取值是 3，变量 b 的取值是 4。这一点可以通过下面的命令来验证。

```
a
## [1] 3
b
## [1] 4
```

赋值是 R 软件最基本的运算，几乎所有的运算，都是从赋值运算开始的。下面的程序则是利用了赋值运算的结果，分别进行不同的运算。

2.3.2 R 软件的代数运算

R 软件的代数运算包括加法、减法、乘法、除法、乘方、开方、求对数等。下面的程序则分别展示了各种代数运算的输入格式，读者可以自己输入到 R 软件的命令界面，看看结果，就可以知道下面的运算类型。需要注意的是，一

般情况下，R 软的命令可以分行输入，若要在同一行输入不同的 R 软件命令，则它们之间需要用 ; 隔开。另外，需要注意的是，R 软件输入命令时，需要将输入方式切换到英文半角状态，才能输入相应的命令。

```
a+b;a-b;a*b;a/b;
a^b;a^(0.5)
```

对于求算术平方根的运算，也可以写成下面的形式

```
sqrt(a)
```

在 R 软件进行代数运算时，有两个常数值需要注意其输入方式，一个是圆周率 π，另一个是求自然对数时的底数 e。R 软件调用这两个常数的命令分别为

```
a <- pi
b <- exp(1)
```

在 R 软件中，一般情况下，求对数运算指的是求一个数的自然对数，例如

```
a <- log(10)
a
```

```
## [1] 2.3
```

若要求以 10 为底的对数，则可以用下面的命令来完成：

```
a <- log(100,10)
b <- log(100,base=10)
```

R 软件还可以进行混合运算，请自行计算下面的结果：

```
a <- 2; b <- 3; c <- 4
res1 <- (a+b)/c-b
res2 <- (a-b)^(c/a)
```

2.3.3 R 软件的编程运算

在初等的统计学教材中，一般不要求掌握 R 软件编程运算的方法，这里仅仅给出两个实例，展示 R 软件的编程运算的大致框架，希望大家可以猜出下面程序中的计算内容和结果。

```
i = 1
sum1 = 0
while(i <= 100){
  sum1 = sum1 + i
  i = i + 1
}
sum1
sum2 <- 0
for(i in 1:100){
  sum2 <- sum2+i
}
sum2
```

2.4 R 软件中命令的运行方式

利用 R 软件进行统计分析，必须运行相应的命令或程序，一般 R 软件的运行方式大致有三种：

- 命令行方式
- 脚本文件方式
- 动态统计报告方式

2.4.1 命令行

R 软件最简单的运行方式是命令行式。即在 R 软件的命令输入窗口输入相应的命令，经过 R 软件处理后，将其结果直接显示在结果输出窗口的一种运行方式。我们将通过一个例子来说明命令行的运行方式的运行过程

例 2.1. 假定圆的半径分别为 1,2 和 3，求圆的周长和面积。

要解决上面的问题，可以在 R 软件的命令输入窗口依次输入如下命令：

```
rm(list=ls())
r <- c(1,2,3)
circle <- 2*pi*r
square <- pi*r^2
circle
square
```

采用命令行的方式进行计算，是 R 软件最简单的一种运行方式。这种方式目前被绝大部分的教科书所采用。

采用命令行式的运行方式，优点是比较简单，并且能够及时看到运算的结果。当然，采用这种方式运行 R 软件，其缺点也比较明显。首先，运算的结果无法保存；其次，运算中的命令无法重复使用，或者是重复使用比较困难。为了提高 R 软件数据处理的效率，在实际的统计分析中，一般都不采用这种运行方式。

2.4.2 脚本文件

利用脚本文件的方式运行 R 软件，最大的优点是效率高。一般情况下，该运行方式需要新建脚本文件、编写脚本文件、运行和保存脚本文件。

- 新建脚本文件

利用脚本文件计算时，首先要新建一个脚本文件。新建脚本文件的方式是：点击 RStudio 工具栏中的 `File`，在其下拉菜单中选择 `New File`，再在下拉菜单中选择 `R Script`，则在一个新的窗口中出现名称为 `Untitled` 的脚本文件。

- 编写脚本文件

在这个名为 `Untitled` 文件中，可以根据所分析的问题来进行 R 软件的编程。仍然以求圆的周长和面积为例，在脚本文件中，其大致的内容如图2.9所示：

通过对比可以发现，R 软件脚本文件中的命令和前面的没有任何不同，但在脚本文件中，多了一些以 # 号开头的文字。这些文字主要用于解释或说明程序中各命令的计算内容。在 R 的脚本文件中，所有位于 # 号后面的文字都被 R 软件所忽略。

- 运行脚本文件

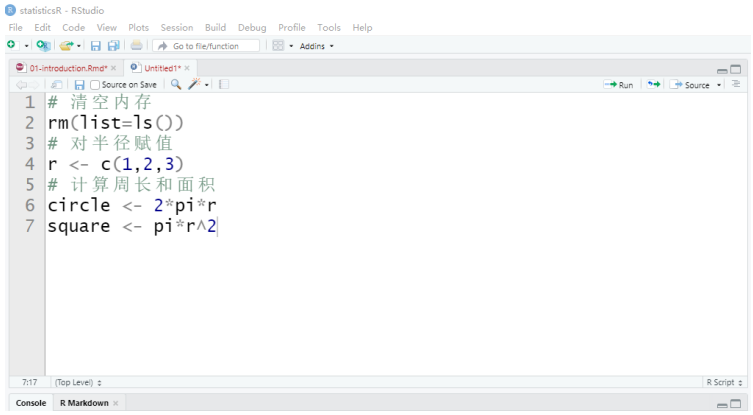


图 2.9: R 软件的脚本文件

编写脚本文件后，可以立即运行脚本文件。其方法是用鼠标选中所要运行的命令，然后点击脚本文件右上角的 Run，即可运行脚本文件，并得到相应的结果。

- 保存脚本文件

和在 R 软件的命令输入和输出界面输入命令方式不同，R 软件的脚本文件可以被保存在电脑中且被重复使用。因此，在脚本文件被运行后，可以对其进行保存。在保存脚本文件最简单的方法是点击脚本文件的右上角关闭该文件时，R 软件会提示是否保存文件，点击 Save 并给文件命名，则该文件可以保存的当前工作路径的文件夹中。

从上面的过程可以看出，和命令行式运行模式相比，脚本文件的编程效率更高。另外，由于有对程序解释说明的文字，让程序的阅读变得更容易。但脚本文件得到的结果并不能直接输出到相关文档中去，还是需要我们采用手工的办法复制、黏贴到 word 文档中去。也就是说，脚本文件虽然提高了程序的运行效率，但对结果的处理效率仍然有待提高。

2.4.3 动态统计报告

由于 R 软件的脚本文件也无法将结果直接输出到 word 等电子文档中去，使得统计分析的结果整理变得十分的繁琐，降低了统计分析的效率。因此，脚本文件也不是 R 软件运行的最佳模式。在 `rmarkdown` 文件中运行 R 软件的命令或编程，是 R 软件的最佳运行模式。因为利用 R 软件的 `rmarkdown` 包，R 软件可以直接撰写统计报告，实现统计报告的“一体化”处理。

利用 R 软件撰写统计报告，需要下载 R 软件的 `rmarkdown` 包和 `knitr` 包，其安装命令如下：

```
install.packages(c('rmarkdown','knitr'))
```

安装完成后，可以用下列命令看看是否成功安装：

```
library('rmarkdown')
library('knitr')
```

若没有出现错误信息，则可以确认上述宏包已经成功的安装。

成功加载上面两个包后，就可以用它们来生成动态统计报告，其具体的做法如下：

1. 点击 R 软件菜单栏上的 **File**，在下拉菜单栏中依次选择 **New File**、**R Markdown**；结果如图2.10所示：

在图中，有三种可供选择的文档类型，它们分别为网页格式、PDF 格式和 Word 格式。一般我们可以选择网页格式和 Word 格式，至于 PDF 格式，由于它的编译涉及到 LaTeX 软件，更为复杂，这里不作要求。

2. 点击选择 Word 格式的文档，得到一个 rmarkdown 文件如图2.11所示：

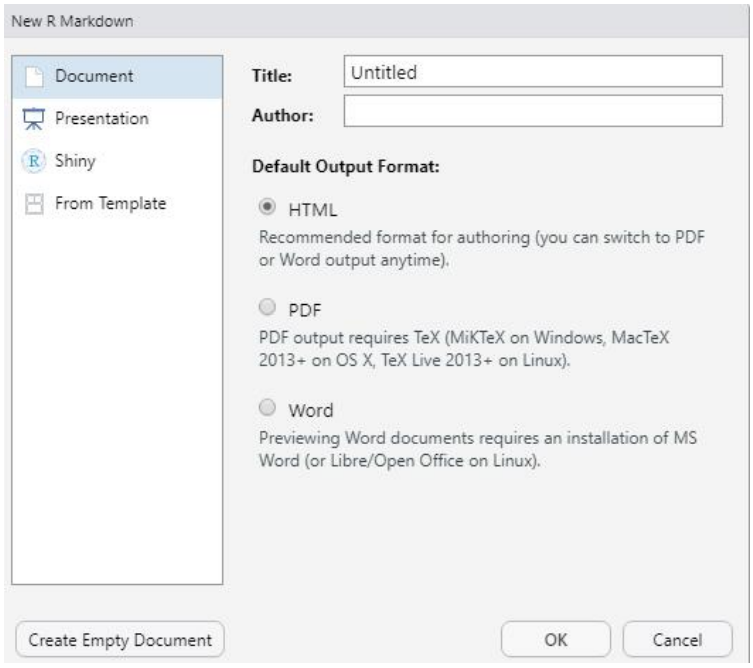


图 2.10: 新建 rmardown 文件

```
---
title: "Untitled"
author: "张三"
date: "2021/7/24"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting
syntax for authoring HTML, PDF, and MS Word documents. For more
details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that
includes both content as well as the output of any embedded R code
chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
```

图 2.11: rmarkdown 文件

这是一个已经设定好格式的 rmarkdown 文件，它的前四行分别是：题目、作者、日期和输出格式。其中日期和输出格式已经给出，我们只有填上动态统计报告的题目和作者就可以了。

下面的例子给出了利用圆的半径计算周长和面积的 rmarkdown 文件的示例，虽然简单，却能够说明所有问题。

将上图中的标题和作者进行修改，并写出该统计报告的文档结构，如图2.12所示：

图 2.12: 撰写 rmarkdown 文件

3. 点击菜单栏中的 **File**，在弹出的对话框中填写文件保存的名称，R 软件会自动将该文件已 rmd 的格式保存到当前工作路径中去。

4. 点击题目上方的 **Knit** 按钮，即可生成 Word 格式的统计报告如图2.13所示：

这里需要作特别说明的是，该报告之所以被称为动态统计报告，根本原因在于，若其他部分保持不变，将程序中的圆的半径改成 3，则该报告中展示出来

图 2.13: 编译得到的动态统计报告

的圆的周长和面积也会发生相应的变化，而无需人手工去干预。

撰写动态统计报告的语法是 rmarkdown 语法，对该语法的说明见附录。

课后练习

一、软件操作题

1. 通过对本章内容的学习，自行下载、安装并设置 R 软件。
2. 当圆的半径 $R = 3$ 时，分别在下列两种情形下计算圆的周长和面积：

- 利用 R 的命令输入输出窗口；
- 利用 R 的脚本文件。

二、编程题

1. 分别利用 R 的命令输入窗口和脚本文件，计算当圆的半径分别是从 1 到 10 的整数时，该圆的周长和面积。
2. 假定圆的半径是从 10 到 20 的整数，圆的周长和面积又会发生什么样的变化。

三、根据要求编写动态统计报告

1. 利用 rmarkdown 包，新建一个动态统计报告，介绍并举例说明 R 软件中的数据的结构。
2. 利用 rmarkdown 包，撰写统计报告，计算当圆的半径为从 1-10 的整数时，圆的周长和面积各是多少？并将该结果在动态统计报告中展示出来。

第三章　数据管理

本章主要内容：

1. 基本的数据处理；
2. 复杂的数据处理；
3. 数据的输出和读入。

3.1　简介

在整个统计研究过程中，数据处理大约占据 90% 的时间，是统计分析过程中最枯燥也是最容易出错的步骤之一。从本章开始，我们将介绍利用 R 软件进行数据管理，它的主要内容有：数据的输出和读入；对数据的基本管理和对数据的高级管理三个部分。另外，从本章开始，所有的统计分析过程都利用 R 软件的 **wooldridge** 包中的 **wage2** 数据集来进行。因此，在学习统计学基本原理和 R 软件的操作之前，我们首先来了解该数据集的安装和使用。

首先，我们需要安装 **wooldridge** 包，其安装命令为

```
install.packages('wooldrige')
```

在该包完成安装后，可以利用

```
library('wooldridge')
```

命令调用该包以及包中的所有数据。

为了进一步了解该包中我们将要使用的数据集 **wage2**，可以用

```
help('wage2')
```

在 R 软件的右下角查看该数据集的帮助文件。由帮助文件可以发现，数据集 **wage2** 是一个有 935 个观测对象和 17 个观测变量的 **data.frame**(数据框)。在其中的 17 个变量中，我们要重点关注的变量主要有：**wage**、**educ** 等数值型变量；以及 **married**、**black** 等虚拟变量下面，我们利用该数据集，讲授 R 软件的数据管理。

3.2 R 软件基本数据管理

R 软件基本数据管理的内容主要包括查看数据的结构、变量名称、以及数据集取子集等。

3.2.1 查看数据结构

在 R 软件中，利用 *View( )* 命令可以查看整个数据集的结构，例如，想查看 R 软件的 **wooldridge** 包中的 **wage2** 的具体的内容和结构：

```
View(wage2)
```

在 R 软件中，还利用 *class( )* 命令可以查看该数据集的数据类型：

```
class(wage2)
```

```
## [1] "data.frame"
```

由此可知，该数据集的数据结构是 R 软件中最常见的数据框。利用 *dim()* 命令可以查看该数据框的具体结构

```
dim(wage2)
```

```
## [1] 935  17
```

和前面利用帮助文件查看的结果一样，该数据框的观测对象由 935 个，观测的变量有 17 个，它们的具体名称如下：

```
colnames(wage2)
```

```
##  [1] "wage"    "hours"   "IQ"      "KWW"     "educ"
##  [6] "exper"   "tenure"  "age"     "married" "black"
## [11] "south"   "urban"   "sibs"    "brthord" "meduc"
## [16] "feduc"   "lwage"
```

该结果和在帮助文件中看到的结果是一样的。

3.2.2 数据集的子集

数据集的子集操作的方法很多，其本质就是将原数据集的数据拆分。就选取子集的方式而言，一般包括按列选取子集、按行选取子集和按逻辑选取子集三种方式。

按列选子集就是按照变量的名称来选择数据集的子集的一种方法。例如在数据集 **wage2** 中，若想选择前面四列，组成一个新的数据集，并对它们进行统计分析，最简单的办法是：

```
newdata <- wage2[,c(1,2,3,4)]
```

若想选择它的第 1 列、第 5 列、第 6 列和第 8 列组成数据集，则有：

```
newdata2 <- wage2[,c(1,5,6,8)]
```

当然，按列选子集还可以按照列的名称来选择子集。例如，**newdata2** 也可以按照下面的方式生成：

```
newdata22 <- wage2[,c("wage","educ","exper","age")]
newdata23 <- wage2[,c("meduc","feduc")]
```

按行选子集就是按照被观测的个体来选取数据集的子集的一种方法。例如，在上面的数据集 **newdata2** 中，我们可以将它拆分为三个数据集，第一数据集包括前面 500 个观测记录，第二个数据集则包括第 501 到第 700 个观测值，第三个数据集则包括剩余的数据集。其拆分方法如下：

```
subnew1 <- newdata2[1:500,]
subnew2 <- newdata2[501:700,]
subnew3 <- newdata2[701:935,]
```

按逻辑选子集指的是利用 *subsec( )* 命令选择子集的一种方法，例如在数据集 **newdata2** 中，我们想选择所有受教育年限不高于 11 年的观测对象，组成一

个子集，而将受教育年限高于 11 年的所有观测对象，组成另外一个子集，其命令为：

```
subeduc1 <- subset(newdata2,educ <= 11)
subeduc2 <- subset(newdata2,educ > 12)
```

当然，利用该命令，还可以进行多种条件的组合，例如：

```
subeduc3 <- subset(newdata2,educ<=11|exper >12)
subeduc4 <- subset(newdata2,educ <=11 & exper >12)
```

利用 R 还可以对数据集进行随机抽样，所用的命令为 *smaple( )*。一般在抽样前需要用 *set.seed( )* 命令设置种子数，以保证抽样结果的可重复性。下面的命令则是对数据集 **newdata2** 进行不可重复抽样得到的新的数据集。

```
set.seed(123)
subsmp <- newdata2[sample(1:nrow(newdata2),100,replace = F),]
```

3.2.3 生成新的变量

在 R 软件中，可以利用已有的变量生成一个新的变量，这个变量可以是数值型变量、也可以是字符串变量或因子变量。

1. 数值型变量。利用 R 软件中的已有变量生成一个新的数值型变量，是 R 软件数据管理中最常见的情形。例如，在上面的名为 **mydata** 的数据集中，利用它的一个名为 **wage** 的变量，生成一个新的变量 **ln.wage**，是变量 **wage** 的自然对数形式。其命令如下：

```
rm(list=ls())
library(wooldridge)
data(wage2)
mydata <- wage2[,c("wage","educ","age","married")]
```

```
mydata$lnwage <- log(mydata$wage)
```

上面的程序在数据集 **mydata** 中生成了一个新的变量，我们可以用 **colnames( )** 来查看其中列变量的变化。

```
colnames(mydata)
```

```
## [1] "wage"    "educ"    "age"     "married" "lnwage"
```

由结果可知，在数据集 **mydata** 中，总共有五个变量，最后哦一个就是刚刚生成的 *lnwage*。利用 *class()* 命令可以查看该变量的类型。

```
class(mydata$lnwage)
```

由结果可知，这是一个数值型的变量。

我们也可以利用已有的变量的取值不同，生成一个具有不同取值的新的变量。在下面的编程中，生成了一个新的名为 **marstat** 的变量，它的赋值规律为，当变量 **married** 的取值为 1 的时候，赋值为 **Married**，反之，当 **married** 的取值为 0 时，赋值为 **Unmarried**。

```
mydata$marstat[mydata$married==1] <- "Married"
mydata$marstat[mydata$married==0] <- "Unmarried"
```

在上面的编程中，首先是判断变量 *married* 的取值，其中的 == 表示的是一种逻辑判断关系。

这里的代码输入比较麻烦，原因在于 R 软件变量的调用规则，一种更简单的方法是利用 **within( )** 函数，直接创建一个新的变量：

```
mydata <- within(mydata,{
  marst <- NA
  marst[married==1]<-"Married"
  marst[married==0]<- "Unmarried"})
```

在上面的编程中，首先创建一个新的变量 **marst**，它的取值为缺失值。然后，根据数据集 **mydata** 中的 **married** 的取值，来确定该变量的取值。这里的变量 **marst** 是一个字符串变量，和前面的结果是一样的。

在 R 软件中，有一类字符串变量可以将其看成为因子变量，这样做是为了统计上的方便。将字符串变量转换为因子变量的是 *factor( )* 命令。

```
mydata$marst2 <- factor(mydata$marst,
                        levels=c("Married","Unmarried"),
                        ordered = T)
```

在 R 软件中，可以利用 *as.numeric()* 命令将因子变量转换为数值变量，但该转换有一个缺点，即比较的基础是从 1 开始，因此，在对 *marst2* 的转换过程中，得到了一个元素取值分别为 1 和 2 的变量，分别对应的是 **Married** 和 **Unmarried**，而不是一个全部元素为 1 和 0 的向量，因此，利用编程的

方法，给出了一个全部元素都为 1 和 0 的向量的生成过程，其基本原理和前面的创建新变量并没有本质的区别。

```
mydata$marst3 <- as.numeric(mydata$marst2)
mydata<- within(mydata,{
  marst4 <- NA
  marst4[marst2=="Married"]<- 1
  marst4[marst2 =="Unmarried"]<-0
})
```

第四章　R 软件作图

本章主要内容：

1. 基本的数据处理；
2. 复杂的数据处理；
3. 数据的输出和读入。

4.1　简介

在整个统计研究过程中，数据处理大约占据 90% 的时间，是统计分析过程中最枯燥也是最容易出错的步骤之一。从本章开始，我们将介绍利用 R 软件进行数据管理，它的主要内容有：数据的输出和读入；对数据的基本管理和对数据的高级管理三个部分。另外，从本章开始，所有的统计分析过程都利用 R 软件的 **wooldridge** 包中的 **wage2** 数据集来进行。因此，在学习统计学基本原理和 R 软件的操作之前，我们首先来了解该数据集的安装和使用。

首先，我们需要安装 **wooldridge** 包，其安装命令为

```
install.packages('wooldrige')
```

在该包完成安装后，可以利用

```
library('wooldridge')
```

命令调用该包以及包中的所有数据。

为了进一步了解该包中我们将要使用的数据集 **wage2**，可以用

```
help('wage2')
```

在 R 软件的右下角查看该数据集的帮助文件。由帮助文件可以发现，数据集 **wage2** 是一个有 935 个观测对象和 17 个观测变量的 **data.frame**(数据框)。在其中的 17 个变量中，我们要重点关注的变量主要有：**wage**、**educ** 等数值型变量；以及 **married**、**black** 等虚拟变量下面，我们利用该数据集，讲授 R 软件的数据管理。

4.2 R 软件基本数据管理

R 软件基本数据管理的内容主要包括查看数据的结构、变量名称、以及数据集取子集等。

4.2.1 查看数据结构

在 R 软件中，利用 *View( )* 命令可以查看整个数据集的结构，例如，想查看 R 软件的 **wooldridge** 包中的 **wage2** 的具体的内容和结构：

```
View(wage2)
```

在 R 软件中，还利用 *class( )* 命令可以查看该数据集的数据类型：

```
class(wage2)
```

```
## [1] "data.frame"
```

由此可知，该数据集的数据结构是 R 软件中最常见的数据框。利用 *dim()* 命令可以查看该数据框的具体结构

```
dim(wage2)
```

```
## [1] 935  17
```

和前面利用帮助文件查看的结果一样，该数据框的观测对象由 935 个，观测的变量有 17 个，它们的具体名称如下：

```
colnames(wage2)
```

```
##  [1] "wage"    "hours"   "IQ"      "KWW"     "educ"
##  [6] "exper"   "tenure"  "age"     "married" "black"
## [11] "south"   "urban"   "sibs"    "brthord" "meduc"
## [16] "feduc"   "lwage"
```

该结果和在帮助文件中看到的结果是一样的。

4.2.2 数据集的子集

数据集的子集操作的方法很多，其本质就是将原数据集的数据拆分。就选取子集的方式而言，一般包括按列选取子集、按行选取子集和按逻辑选取子集三种方式。

按列选子集就是按照变量的名称来选择数据集的子集的一种方法。例如在数据集 **wage2** 中，若想选择前面四列，组成一个新的数据集，并对它们进行统计分析，最简单的办法是：

```
newdata <- wage2[,c(1,2,3,4)]
```

若想选择它的第 1 列、第 5 列、第 6 列和第 8 列组成数据集，则有：

```
newdata2 <- wage2[,c(1,5,6,8)]
```

当然，按列选子集还可以按照列的名称来选择子集。例如，**newdata2** 也可以按照下面的方式生成：

```
newdata22 <- wage2[,c("wage","educ","exper","age")]
newdata23 <- wage2[,c("meduc","feduc")]
```

按行选子集就是按照被观测的个体来选取数据集的子集的一种方法。例如，在上面的数据集 **newdata2** 中，我们可以将它拆分为三个数据集，第一数据集包括前面 500 个观测记录，第二个数据集则包括第 501 到第 700 个观测值，第三个数据集则包括剩余的数据集。其拆分方法如下：

```
subnew1 <- newdata2[1:500,]
subnew2 <- newdata2[501:700,]
subnew3 <- newdata2[701:935,]
```

按逻辑选子集指的是利用 *subsec( )* 命令选择子集的一种方法，例如在数据集 **newdata2** 中，我们想选择所有受教育年限不高于 11 年的观测对象，组成一

个子集，而将受教育年限高于 11 年的所有观测对象，组成另外一个子集，其命令为：

```
subeduc1 <- subset(newdata2,educ <= 11)
subeduc2 <- subset(newdata2,educ > 12)
```

当然，利用该命令，还可以进行多种条件的组合，例如：

```
subeduc3 <- subset(newdata2,educ<=11|exper >12)
subeduc4 <- subset(newdata2,educ <=11 & exper >12)
```

利用 R 还可以对数据集进行随机抽样，所用的命令为 *smaple( )*。一般在抽样前需要用 *set.seed( )* 命令设置种子数，以保证抽样结果的可重复性。下面的命令则是对数据集 **newdata2** 进行不可重复抽样得到的新的数据集。

```
set.seed(123)
subsmp <- newdata2[sample(1:nrow(newdata2),100,replace = F),]
```

4.2.3 生成新的变量

在 R 软件中，可以利用已有的变量生成一个新的变量，这个变量可以是数值型变量、也可以是字符串变量或因子变量。

1. 数值型变量。利用 R 软件中的已有变量生成一个新的数值型变量，是 R 软件数据管理中最常见的情形。例如，在上面的名为 **mydata** 的数据集中，利用它的一个名为 **wage** 的变量，生成一个新的变量 **ln.wage**，是变量 **wage** 的自然对数形式。其命令如下：

```
rm(list=ls())
library(wooldridge)
data(wage2)
mydata <- wage2[,c("wage","educ","age","married")]
```

```
mydata$lnwage <- log(mydata$wage)
```

上面的程序在数据集 **mydata** 中生成了一个新的变量，我们可以用 **colnames( )** 来查看其中列变量的变化。

```
colnames(mydata)
```

```
## [1] "wage"    "educ"    "age"     "married" "lnwage"
```

由结果可知，在数据集 **mydata** 中，总共有五个变量，最后哦一个就是刚刚生成的 *lnwage*。利用 *class()* 命令可以查看该变量的类型。

```
class(mydata$lnwage)
```

由结果可知，这是一个数值型的变量。

我们也可以利用已有的变量的取值不同，生成一个具有不同取值的新的变量。在下面的编程中，生成了一个新的名为 **marstat** 的变量，它的赋值规律为，当变量 **married** 的取值为 1 的时候，赋值为 **Married**，反之，当 **married** 的取值为 0 时，赋值为 **Unmarried**。

```
mydata$marstat[mydata$married==1] <- "Married"
mydata$marstat[mydata$married==0] <- "Unmarried"
```

在上面的编程中，首先是判断变量 *married* 的取值，其中的 == 表示的是一种逻辑判断关系。

这里的代码输入比较麻烦，原因在于 R 软件变量的调用规则，一种更简单的方法是利用 **within( )** 函数，直接创建一个新的变量：

```
mydata <- within(mydata,{
  marst <- NA
  marst[married==1]<-"Married"
  marst[married==0]<- "Unmarried"})
```

在上面的编程中，首先创建一个新的变量 **marst**，它的取值为缺失值。然后，根据数据集 **mydata** 中的 **married** 的取值，来确定该变量的取值。这里的变量 **marst** 是一个字符串变量，和前面的结果是一样的。

在 R 软件中，有一类字符串变量可以将其看成为因子变量，这样做是为了统计上的方便。将字符串变量转换为因子变量的是 *factor( )* 命令。

```
mydata$marst2 <- factor(mydata$marst,
                        levels=c("Married","Unmarried"),
                        ordered = T)
```

在 R 软件中，可以利用 *as.numeric()* 命令将因子变量转换为数值变量，但该转换有一个缺点，即比较的基础是从 1 开始，因此，在对 *marst2* 的转换过程中，得到了一个元素取值分别为 1 和 2 的变量，分别对应的是 **Married** 和 **Unmarried**，而不是一个全部元素为 1 和 0 的向量，因此，利用编程的

方法，给出了一个全部元素都为 1 和 0 的向量的生成过程，其基本原理和前面的创建新变量并没有本质的区别。

```
mydata$marst3 <- as.numeric(mydata$marst2)
mydata<- within(mydata,{
  marst4 <- NA
  marst4[marst2=="Married"]<- 1
  marst4[marst2 =="Unmarried"]<-0
})
```

第五章　描述性统计学

通过统计图形或统计表格展示数据，其最大的优点在于可以得到关于数据的最直观的印象。为了进一步掌握数据分布的特征和规律性，还需要对数据分布的特征进行定量分析，它通常由描述性统计学来完成。在本章中，我们将从以下桑方面来测度或描述数据的分布特征：

- 分布的集中趋势
- 分布的离散趋势
- 分布的形状描述

本章将重点讨论以上三个方面的分布特征值的计算方法、特点及应用场合。

5.1　分布的集中趋势

集中趋势 (**Central tendency**) 指的是一组数据向某一中心值靠拢的倾向。集中趋势的测度就是找出这一中心值。在本节中，我们将按照变量的数据类型，分别介绍分类数据、顺序数据和数值型数据的集中趋势的测度。

5.1.1　分类数据的集中趋势

测度分类数据的集中趋势，最常见的方法是用众数。

众数 (**Mode**) 指的是一组数据中出现次数最多的变量值。例如，一个班级有 58 人，按照性别划分，其中男生 32 人，女生 26 人。这里的变量 性别就是一个分类变量，它的众数是 男性，即 $M_0 =$ 男性。

例 5.1. 利用 **wooldridge** 包中的数据集 **wage1**，计算该数据集中的分类变量 *female* 和 *married* 的众数。

解：计算变量 *female* 的众数的 R 语言程序如下：

```
library(wooldridge)
mydata <- wage1[, c("female","married")]
sum.f <- sum(mydata$female)
n <- nrow(mydata)
sum.m <- n-sum.f
cbind(n,sum.f,sum.m)
```

```
##         n sum.f sum.m
## [1,] 526   252   274
```

由最终结果可以看出，在数据集 **wage1** 中，总共有 526 人，其中，已经女性的人数为 252 人，男性的人数为 274 人。因此，变量 *female* 的众数为男性。

运用同样的方法，可以计算变量 *married* 的 R 语言编程如下：

```
sum.ma <- sum(mydata$married)
sum.ua <- n-sum.ma
cbind(n,sum.ma,sum.ua)
```

```
##         n sum.ma sum.ua
## [1,] 526    320    206
```

因此，在数据集 **wage1** 中，变量 *married* 的众数为已婚。在本例中需要注意的是，由于数据集中的分类变量 *female* 和 *married* 的取值只有两种可能，因此，可以用 1 和 0 来分别表示。即当 *female* 的取值为 1 时，表示的是女性，其取值为 0 则表示男性；同样的道理，当 *married* 的取值为 1 时，表示已婚，其取值为 0 则表示未婚。

用于测度分类变量的集中趋势的众数，同样适用于顺序数据和数值型数据的集中趋势的测度。当然，一般情况下，只有在观测值比较大的场合，利用众数来测度变量的集中趋势才有意义。

5.1.2 顺序数据的集中趋势

测度顺序数据集中趋势的统计量或指标是分位数。常见的分位数包括中位数、四分位数、十分位数和百分位数等。这里我们只介绍中位数和四分位数的计算方法，以及它们和百分位数在 R 软件中的调用。

1 中位数

中位数 (**Median**) 指的是一组数据排序后处于中间位置上的变量值。一般情况下用 M_e 表示中位数。

中位数将数据分为数量相同的两部分，每一部分都包含 50% 的数据，其中一部分数据比中位数小，另一部分则比中位数大。中位数可以用来测度顺序数据和数值型数据的集中程度。中位数是一个位置平均数，其特点是不受极端值的影响。

计算中位数，一般分两步：首先确定中位数的位置，然后确定中位数的大小。

对于未分组的原始数据，需要首先需要对数据进行排序，然后确定中位数的位置，即

$$\text{中位数位置} = \frac{n+1}{2} \tag{5.1.1}$$

对于顺序数据或分组数据，中位数的位置为：

$$\textbf{中位数位置} = \frac{n}{2} \tag{5.1.2}$$

其中，n 为数据的个数。

在确定中位数的位置后，可以确定中位数的具体数值。

假定有一组数据为 $x_1, x_2, \cdots, x_n$，按照从小到大排序后，该组数据变为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$，则该组数据的中位数为

$$\boldsymbol{Me} = \begin{cases} x_{(\frac{n+1}{2})} & \text{n 为奇数} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{\frac{n}{2}+1}) & \text{n 为偶数} \end{cases} \tag{5.1.3}$$

2 四分位数

和中位数相似的还有四分位数、十分位数和百分位数等。它们分别将数据分为相等的 4 份、10 份和 100 份后，各个分位点上的值。由于它们的性质基本相似，因此这里只对四分位数进行简单的介绍。

四分位数四分位点指的是一组数据排序后，处于 25% 和 75% 位置上的值。

四分位数通过三个点将全部数据分为相等的四个部分，其中每一部分包含 25% 的数据。很明显，中间的四分位数就是中位数，因此，通常所说的四分位数指的就是处在 25% 和 75% 位置上的数值，前者又被称为下四分位数，后者则被称为上四分位数。而上四分位数和下四分位数的差，恰好包含了一半的数据。

与中位数的计算方法相似，根据未分组数据计算四分位数时，首先对数据进行排序，确定四分位数位置。

例如，假定下四分位数为 Q_L，上四分位数为 Q_U，对于未分组的原始数据，各四分位数的位置为

$$Q_L\text{的位置} = \frac{n+1}{4}; \qquad Q_U\text{的位置} = \frac{3(n+1)}{4}$$

这里需要说明的是，若四分位数的位置不是某一个整数，则需要根据它的位置，按比例分摊四分位数位置两侧数值的差值。

5.1.3 数值型数据的集中趋势

可以用平均数来度量数值型数据的集中趋势。平均数主要适用于数值型数据，而不适用于分类数据和顺序数据。一般情况下，平均数的定义如下：

平均数指的是一组数据相加后除以数据的个数所得到的结果。

根据所掌握的数据不同，平均数可以有不同的形式。

1 算术平均数

算术平均数包括简单算术平均数和加权算术平均数两种。一般多在未分组数据中使用简单算术平均数，而在分组数据中使用加权算术平均数。假定有一组数据 $x_1, x_2, \cdots, x_n$，其样本容量为 n，则该组数据的平均数 $\bar{x}$ 的算术平均数为：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5.1.4}$$

在分组数据中，常用加权算术平均数来计算平均数。假定数据被分为 k 组，各组的组中值分别为 $M_1, M_2, \cdots, M_k$，各组变量值出现的频数分别为 $f_1, f_2, \cdots, f_k$，则其算术平均数为

$$\bar{x} = \frac{M_1 f_1 + M_2 f_2 + \cdots + M_k f_k}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{k} M_i f_i}{\sum f_i} \tag{5.1.5}$$

其中，$\sum f_i$ 就是观测值的个数。

根据式 (5.1.5) 计算平均数时，采用各组的组中值代表这组数据的一般水平，其实它暗含了数据在这一组是均匀分布的假定，若实际数据与假定相符，则计算的结果比较准确，否则，则会产生较大的误差。

表 5.1.1: 三种水果批发价格和成交数据

蔬菜名称	批发价格（元）	成交量（公斤）	成交额（元）
甲	6.0	15000	90000
乙	8.0	25000	200000
丙	15	10000	150000
合计	-	50000	440000

利用式 (5.1.4) 计算简单算术平均数时，计算结果仅仅和观测值的大小有关。但在利用式 (5.1.5) 计算平均数时，其结果不仅受到各组的组中值的大小影响，还受到各组变量值出现的频数 f_i 的影响。若某一组的权数越大，说明该组的数据越多，则该组数据的大小对平均数的影响就越大。反之，则影响越小。可将式 (5.1.5) 改写成如下形式：

$$\bar{x} = \frac{\sum\limits_{i=1}^{n} M_i f_i}{\sum f_i} = \sum_{i=1}^{n} M_i \frac{f_i}{\sum f_i} \tag{5.1.6}$$

由上式可以看出，对分组数据而言，有两类因素影响其均值大小：一类是各组的水平，即各组的组中值 M_i 大小；另一类是分组数据的结构，即各组在总体中所占的比重 $f_i/\sum f_i$。平均数在描述数据的集中程度时具有极其重要的地位，它是进行统计分析和统计推断的基础。从统计思想上看，平均数是一组数据的重心所在，是统计误差相互抵消后的必然结果，因此，可以反映被研究的经济现象的必然性的数量特征。关于算术平均数的重要性质，我们在数据分布的离散程度中会进行详细的说明。

例 5.2. 某批发市场中，有三种水果的日成交数据如下表所示，计算三种水果的平均批发价格。

解：这是一道典型的加权算术平均数的计算题，可以直接用公式计算

$$\bar{x} = \frac{\sum M_i f_i}{\sum f_i} = \frac{6 \times 15000 + 8 \times 25000 + 15 \times 10000}{15000 + 25000 + 10000} = 8.8(\text{元})$$

当然，这里还可以这样算

$$\bar{x} = \sum M_i \frac{f_i}{\sum f_i} = 6 \times \frac{3}{10} + 8 \times \frac{1}{2} + 15 \times \frac{1}{5} = 8.8(\text{元})$$

从第二种算法中，我们可以很明显看出，均值不仅受到各种水果的价格水平的影响，也受到批发的商品结构的影响。

本题的 R 软件计算编程如下：

```
P <- c(6,8,15)
Q <- c(15000,25000,10000)
TR <-t(P)%*%Q
mean.P <- TR/sum(Q)
```

则可以得到三种水果价格的均值为 8.8 (元)。

2 调和平均数

在实际的均值计算中，由于所获得的数据的结构不同，有时，我们无法直接应用式 (5.1.5) 和式 (5.1.6) 计算平均数，而采用调和平均数的形式计算均值。调和平均数的定义如下：

调和平均数是各变量倒数的平均倒数，它是平均数的另一种表现形式。

一般情况下，调和平均数用 H_m 来表示。我们可以用一个例子来说明调和平均数的计算方法。

例 5.3. 假定在上例中，我们只知道各种水果的价格和各种水果批发的金额如下表所示，求三种水果的平均价格。

解：本题中没有各种水果的成交量信息，因此，需要首先求出各种水果的成交量，然后用成交金额除以成交量来计算水果的平均价格，即

表 5.1.2: 三种水果批发价格和成交数据

蔬菜名称	批发价格（元）	成交额（元）
甲	6.0	90000
乙	8.0	200000
丙	15	150000
合计	-	440000

$$H_m = \frac{\sum 成交金额}{\sum \frac{各组成交额}{各组的批发价格}}$$

所以，本题的计算方法如下：

$$H_m = \frac{90000 + 150000 + 250000}{\frac{90000}{6} + \frac{200000}{8} + \frac{15000}{15}} = \frac{440000}{15000 + 25000 + 10000} = 8.8(元)$$

由此可见，调和平均数和算术平均数计算得到的结果完全相同，根本原因在于两者计算时所依据的信息不同，本质上却是完全一样的。

同样，本题也可利用 R 软件编程计算，如下

```
P <- c(6,8,15)
R <- c(90000,200000,150000)
Hm.P <- sum(R)/sum(R/P)
```

同样，可以得到水平的平均价格为 8.8 (元)。

3 几何平均数

定义 5.1. 几何平均数指的是 n 个变量乘积的 n 次方根。

一般情况下，几何平均数可以用 G_m 来表示，其计算公式为

$$G_m = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{5.1.7}$$

几何平均数是一种特殊形式的平均数，一般用来对时间序列进行分析。在实际应用中，几何平均数主要用于计算社会经济现象的平均增长率。

假定某经济社会的年经济增长率分别为 $G_1, G_2, \cdots, G_n$，我们可以求该社会的平均年经济增长率，其基本的思路如下：

假定该社会基期的 GDP 为 y_0，根据已知的信息，可得该经济社会的第 n 年的 GDP 为

$$y_n = y_0(1+G_1)(1+G_2)\cdots(1+G_n) = y_0\prod_{i=1}^{n}(1+G_i)$$

再假定从基期到第 n 年，每年的经济增长率相同，则有

$$y_n = y_0(1+G)^n$$

由此可得

$$(1+G)^n = \prod_{i=1}^{n}(1+G_i)$$

由此可得

$$\bar{G} = \sqrt[n]{\prod_{i=1}^{n}(1+G_i)} - 1$$

当然，在上面的计算中，我们也可以用求自然对数的方法，将其转变为算术平均数，即

$$ln(1+\bar{G}) = \frac{\sum_{i=1}^{n} ln(1+G_i)}{n}$$

在这里需要注意的是，无论是几何平均数的哪一种算法，都要求 $1+G_i>0$。

例 5.4. 假定一位投资人持有的股票在最近五年的收益率分别为 5%、6%、12%、4% 和 6%，求该投资人在这五年的平均收益率。

解：利用 R 软件计算编程如下：

```
n <- 5
g<- c(1.05,1.06,1.12,1.04,1.06)
fg <- prod(g)
mean.g <- 100*(fg^(1/n)-1)
```

由此可得，该投资人持有股票的年平均增长率为 6.56%。

5.1.4 众数、中位数和平均数的比较

众数、中位数和平均数是衡量集中趋势的三个主要的测度值，它们有不同的特点和应用场合。简单而言，众数的特点是不受极端值的影响，具有较强的稳定性，但众数一般用来衡量分类变量的集中程度，对顺序变量和数值型变量，有时并不适用；中位数和众数一样，也不受极端值的影响，因此比较稳定，但中位数则主要用于顺序变量和数值型变量，对分类变量，其意义不大；平均数经常被用来衡量数据的集中程度，但它容易受到极端值的影响，同样，平均数只能用于衡量数值型变量的集中程度，而不适用于分类变量和顺序变量。

1. 众数、中位数和平均数之间的关系

从分布的角度看，众数是一组数据分布的最高峰值，中位数则处于一组经过排序后的数据的中间位置，而平均数则是全部数据的算术平均。因此，对于具有单峰分布的数据而言，众数、中位数和平均数具有以下关系：

第一、若数据的分布是对称的，则众数 M_O、中位数 M_e 和平均数 $\bar{x}$ 必定相等，即 $M_o = M_e = \bar{x}$；第二、若数据的分布是左偏的，说明数据存在异常的极小值，必然拉动平均数向极小值方向靠近，而众数和中位数则不受极值的影响，因此，必然有 $M_o = M_e > \bar{x}$；反之，若数据的分布是有偏的，则说明数据存在异常的极大值，必然拉动平均数向极大值方向靠近，因此，必然有 $M_o = M_e < \bar{x}$。

众数、中位数和平均数之间的关系可以用下图来表示：（这儿需要插一幅图）。

2. 平均数在应用时应注意的事项

平均数，无论是算术平均数还是调和平均数，主要适用于衡量数值型变量的集中程度。虽然，对数值型变量而言，也可以采用众数和分位数来衡量集中程度，但总体上看，平均数无疑是使用最为广泛的指标。但需要注意的是，在数据是对称分布或接近对称分布情形下，利用平均数衡量数据集中趋势无疑是合适的，但在数据分布发生较大偏斜的情形下，用平均数来衡量总体的集中程度，无疑是不适合的，此时，应当在给出平均数的同时，也需要给出相应的众数或中位数来进行辅助的说明。下面，我们将使用案例对这一点进行详细的说明。

例 5.5. 利用 wooldridge 包中的 bwght 数据集，计算变量 cigs 的均值，并说明该值能否代表孕妇每天抽烟数量的一般水平或集中程度。

解： 首先加载数据集并计算变量 cigs 的均值

```
library(wooldridge)
data(bwght)
mean.cigs <- mean(bwght$cigs)
```

由此，得到孕妇每天吸烟数量的平均数为 2.09 根。但浏览该数据集可以看出，绝大部分的孕妇在怀孕期间其实都是不抽烟的，为此，我们可以计算变量 `cigs` 的众数和中位数。

```
n <- length(bwght$cigs)
n.ncigs <- sum(bwght$cigs==0)
prop <- 100*round(n.ncigs/n,2)
```

由上面的计算结果可以看出，在所有的 1388 名孕妇中，有 1176 名孕妇其实是不抽烟的，即有 85% 的孕妇在怀孕期间是不抽烟的，因此，用 2.09 根来表示孕妇怀孕期间每天抽烟的根数，本身就是由问题的。因此，在本题中，我们最好用 0 来表示孕妇每天吸烟的一般水平。

5.2 分布的离散趋势

集中趋势和离散趋势是数据分布的两个不同的特征。前者反映的各个变量值向其中心值聚集的程度，后者反映的是各变量值和一般水平之间的差异程度。根据前面的知识可知，集中程度的各测量值是对数据一般水平的概括性度量，但这一概括性度量是否具有代表性，则取决于数据的离散程度。数据的离散程度越小，集中程度的测量值对一般水平概括的代表性就越高，反之，集中程度的测量值对一般水平概括的代表性就越低。而离中趋势的各个测度值就是对数据离散程度的衡量。

常见的描述数据离散程度的测度值主要包括异众比率、四分位差、方差和标准差。此外，还包括极差、平均差和测度相对离散程度的离散系数等测度值。另外，在本节，我们还将学习如何将一组数据标准化。

5.2.1 异众比率

异众比率主要用来衡量众数对一组数据的代表程度。其定义为：

定义 5.2. 异众比率指的是非众数组的频数占总频数的比率。

异众比率一般用 V_r 表示，其计算公式为

$$V_r = \frac{\sum f_i - f_m}{\sum f_i} = 1 - \frac{f_m}{\sum f_i} \tag{5.2.1}$$

其中，$\sum f_i$ 表示变量值的总频数，f_m 表示众数的频数。

由其定义可以看出，异众比率越大，说明非众数组的频数占总频数的比重越大，众数的代表性就越差；异众比率越小，说明非众数组的频数占总频数的比重越小，则众数的代表处就越好。虽然异众比率主要用于测度分类数据的离散程度，但它也可以用于测度顺序数据和数值型数据的离散程度。

例 5.6. 利用 `wooldridge` 包中的 `bwght` 数据集，说明变量 `cigs` 的众数及其异众比率。

解：根据前面的计算结果，变量 `bwght` 的异众比率计算如下：

```
v.r <- 100*(1-n.ncigs/n)
v.r <- round(v.r,2)
```

由此可得，该变量值的异众比率为 15.27%，说明在所有的 1388 名孕妇中，只有 15.27% 的孕妇在怀孕期间吸烟，而高达 `100-v.r`% 的孕妇在怀孕期间不吸烟。很明显，和平均数相比，众数无疑具有更好的代表性。

5.2.2 四分位差

例 5.7. 四分位差指的是上四分位数与下四分位数之间的差，又被称为内距或四分间距。

一般情况下，四分位差用 Q_d 表示。其计算公式为

$$Q_d = Q_U - Q_L \tag{5.2.2}$$

其中，Q_U 和 Q_L 分别表示上四分位数和下四分位数。

四分位差反映了中间 50% 的数据的离散程度，其数值越小，说明中间的数据越集中，反之，则表明中间的数据越分散。四分位差不受极值的影响。此外，由于中位数位于上下四分位数的中间位置，因此，四分位差在一定程度上也反映了中位数对一组数据的代表程度。

虽然四分位差主要用来测度顺序变量的离散程度，但也可以用于测度数值型数据的离散程度，但它并不适用于测度分类数据的离散程度。

例 5.8. 利用 wooldridge 包中的 bwght 数据集，计算变量 bwght 的四分位差。

解：计算四分位差的第一步是确定其位置：

```
sort.bwght <- sort(bwght$bwght)
n <- length(sort.bwght)
p.l <- (n+1)/4
p.u <- 3*(n+1)/4
```

即变量 bwght 的下四分位数和上四分位数的位置分别为 347.25 和 1041.75，因此，变量 bwght 的上、下四分位数

```
f.l <- floor(p.l)
f.u <- floor(p.u)
ratio1 <- p.l-f.l
ratio2 <- p.u-f.u
q.l <- sort.bwght[[f.l]]*(1-ratio1)+sort.bwght[[f.l+1]]*ratio1
q.u <- sort.bwght[[f.u]]*(1-ratio2)+sort.bwght[[f.u+1]]*ratio2
d.ul <- q.u-q.l
```

经计算的出生婴儿的体重的下四分位数为 107 盎司，上四分位数为 132 盎司，因此，出生婴儿的体重的四分位差为 25 盎司。

5.2.3 方差和标准差

测度数值型变量离散程度的方法主要有方差和标准差，除此外，还有极差和平均差。

1 极差

定义 5.3. 极差又称全距，是一组数据的最大值和最小值之间的差。

一般情况下，极差用 R 表示，其计算公式为

$$R = max(x_i) - min(x_i) \tag{5.2.3}$$

其中，$max(x_i)$ 和 $min(x_i)$ 分别表示一组数据的最大值和最小值。

极差是描述数据离散程度最简单的测度值，其特点是计算简单，容易理解。但它容易受到异常值的影响。另外，由于极差只利用了一组数据中两个端点的信息，不能反映中间数据的离散状况，因此，利用极差无法准确测度数据的分散程度。

2. 平均差

定义 5.4. 平均差又称平均离差，是各变量值和其平均数离差的绝对值的平均数。

平均差一般用 M_d 表示，其计算公式如下：

$$AD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{5.2.4}$$

若是依据分组数据计算平均差，则其公式是：

$$AD = \frac{\sum_{i=1}^{k} |M_i - \bar{x}| f_i}{\sum f_i} \tag{5.2.5}$$

其中，M_i 表示各组的水平，f_i 表示各组的频数。

例 5.9. 利用 wooldridge 包中的 bwght 数据集，计算变量 bwght 的极差和标准差。

解：利用 R 软件计算变量 bwght 的极差如下：

```
max.w <- max(bwght$bwght)
min.w <- min(bwght$bwght)
range <- max.w-min.w
```

由此可知，出生婴儿体重的极差为 248 盎司。

利用 R 软件计算出生婴儿体重平均差的程序如下：

```
mean.w <- mean(bwght$bwght)
de.w <- abs(bwght$bwght-mean.w)
ad.w <- round(sum(de.w)/length(de.w),2)
```

得到出生婴儿体重的平均差为 15.49 盎司。

平均差以平均数为中心，反映了每个数据和平均数之间的平均差异程度，因此可以比较全面的反映变量的离散情况。平均差的优点是实际意义清楚，容易理解。但其缺点在于，由于变量值和均值的离差和为 0，因此，计算平均差时需要对离差进行绝对值计算，造成了平均差的计算比较复杂，在实际中应用较少。

3. 方差和标准差

和平均差通过求离差的绝对值避免离差和等于 0 的情形相似，方差对离差采取平方的方法来避免离差和等于 0 的情况发生。这样做的好处是在数学上处理更方便。因此，方差和标准差是实际应用最为广泛的两个离散程度的测度值。

定义 5.5. 方差是各变量值与其均值离差平方和的平均数。

一般用 σ^2 表示总体方差，用 s^2 表示样本方差。它们的计算公式如下：

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{5.2.6}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{5.2.7}$$

可以发现，两个公式的区别在式 (5.2.6) 在计算方差时，其分母的数值为 n，式 (5.2.7) 在计算方差时，其分母的数值为 $(n-1)$。[1]

定义 5.6. 标准差指的是变量值和均值离差的平方和的均值的算术平方根。

同样，标准差也可分为总体标准差和样本标准差，

其中，总体标准差的计算公式为

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} \tag{5.2.8}$$

样本标准差的计算公式为

[1]这里我们不再讨论自由度问题。希望了解该概念的同学参阅相关书籍。

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{5.2.9}$$

例 5.10. 利用数据集 bwght，计算其变量 bwght 的方差和标准差。

解：计算变量 bwght 的方差和标准差的程序如下

```
var.w <-round(var(bwght$bwght),2)
s.w <- round(sqrt(var.w),2)
```

由此可知，出生婴儿的体重的方差为 414.28，体重的标准差为 20.35 盎司。[2]

5.2.4 数据的标准化

有了均值和标准差后，可以利用它们对数据进行标准化，从而判断数据中是否存在异常值。

1. 标准化值

定义 5.7. 标准化值又称标准分数或 z 分数，指的是变量值与其均值的离差除以标准差后的值。

标准化值一般用 z 表示，根据定义得

$$z_i=\frac{x_i-\bar{x}}{s} \tag{5.2.10}$$

标准化值正好给出了一组数据中各变量值的相对位置。例如，若某个变量值的标准化值为 1，则表明该数值高于均值 1 倍标准差。当需要对多个具有不同量纲的变量进行处理时，常常需要用式 (5.2.10) 对各个变量进行标准化处理。

[2]R 软件在计算方差和标准差时，默认所计算的是样本标准差而非总体标准差。

例 5.11. 根据数据集 bwght 中的变量 bwght 的信息，求一个出生时体重为 130 盎司的婴儿的体重的标准化值，并说明该值的含义。

解：计算标准化值的 R 软件程序如下：

```
m.w <- mean(bwght$bwght)
s.w <- sd(bwght$bwght)
bw <- 132
z <- (bw-m.w)/s.w
zval <- round(z,3)
```

经计算的一个出生体重为 132 盎司的婴儿的标准化值为 0.653，它的含义是该婴儿体重位于婴儿体重均值右侧的 0.65 个标准差上。

标准化值实际上对原始数据进行了线性变换，使其均值为 0，标准差为 1，但它没有改变数据原始数据中的位置，也没有改变原始数据的分布形状。我们可以对变量 bwght 进行标准化并验证上面的结论。

```
s.bw <- (bwght$bwght-m.w)/s.w
m.sw <- mean(s.bw)
s.sw <- sd(s.bw)
```

由此可得标准化后的婴儿体重的均值为 0，标准差为 1。

2. 数据分布的经验法则

数据分析表明，当数据呈对称分布时，有如下经验法则：

1）大约有 68% 的数据位于其均值加减 1 倍标准差范围之内；

2）大约有 95% 的数据位于其均值加减 2 倍标准差范围之内；

3）大约有 99% 的数据位于其均值加减 3 倍标准差范围之内。

利用数据 bwght 中的 bwght 变量，可以对该经验法则进行验证。

```
n1<- sum(s.bw >=-1 & s.bw <=1)
ratio1 <- 100* n1/length(s.bw)
n2 <- sum(s.bw>=-2 & s.bw <=2)
ratio2 <- 100* n2/length(s.bw)
n3 <- sum(s.bw >=-3 & s.bw <= 3)
ratio3 <- 100* n3/length(s.bw)
n4 <- sum(s.bw<-3)
n5 <- sum(s.bw>3)
```

上面的计算结果表明，大约有 72% 的数据位于 1 倍标准差范围之内，大约有 95% 的数据位于 2 倍标准差范围之内，大约有 99% 的数据位于 3 倍标准差范围之内，和我们的经验法则基本差不多。

以上的计算结果同样表明，一组数据中位于其均值 3 倍标准差之内的数据几乎包括了所有的数据，位于数据均值 3 倍标准差之外的数据是相当罕见的。我们将位于均值 3 倍标准差之外的数据称为异常值或离群点。对数据集 bwght 中的变量 bwght 的分析表明，有 3 个数据位于均值左侧 3 倍标准差以外，它们就是异常值或离群点。

3. 切比雪夫不等式

上面的经验法则成立与否，主要取决于所分析的数据是否呈对称分布。若数据分布不是对称的，经验法则就有可能不在适用。此时可以使用切比雪夫不等式来确定其下界。即确定所占比例的最小值。切比雪夫不等式的主要内容为：

设随机变量 X 具有数学期望 $E(X)=\mu$，方差 $D(X)=\sigma^2$，则对于任意正数 ε，有不等式

$$P(|X-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

成立。

按照切比雪夫不等式，存在任意大于 1 的 k，使得

$$P(|X-\mu| \leq \varepsilon) \geq 1-\frac{\sigma^2}{\varepsilon^2}$$

将数据标准化后，一个很明显的结论是：至少有 $(1-1/k^2)$ 的数据位于 k 倍标准差范围之内。若 k 的取值分别为 2、3 和 4，则有下面的结论：

- 至少有 75% 的数据落入均值的 2 倍标准差范围之内；
- 至少有 89% 的数据落入均值的 3 倍标准差范围之内；
- 至少有 94% 的数据落入均值的 4 倍标准差范围之内。

利用数据集 bwght 中的变量 cigs 来验证切比雪夫不等式，可得

```
m.cigs <- mean(bwght$cigs)
sd.cigs <- sd(bwght$cigs)
s.cigs <- (bwght$cigs-m.cigs)/sd.cigs
n1 <- sum(s.cigs>=-2 & s.cigs<=2)
ratio1 <-100*n1/length(s.cigs)
n2 <- sum(s.cigs >= -3 & s.cigs <= 3)
ratio2 <- 100*n2/length(s.cigs)
n3 <- sum(s.cigs >= -4 & s.cigs <=4)
ratio3 <- 100*n3/length(s.cigs)
```

由前面的结果可知，数据集 bwght 中的变量 cigs 是非对称的，根据程序的计算结果可知，

- 有 93% 的数据位于 2 倍标准差范围之内；
- 有 99% 的数据位于 3 倍标准差范围之内；
- 有 99% 的数据位于 4 倍标准差范围之内。

5.2.5 离散系数

由前面的学习可知，在反映离散程度的测量值包括极差、平均差、方差和标准差等，它们反映的是数据离散程度的绝对值。一般情况下，它们的大小和变量的平均水平有关，平均水平越高，离散程度也越大。另外，它们的大小还和数据的计量单位有关，采用不同单位计量的变量值，其方差和标准差的差别却相当大。因此，为了正确衡量不同均值水平下数据的波动程度，可以计算离散系数。

定义 5.8. 离散系数又称变异系数，指的是一组数据的标准差和其均值的比。

可以用 CV 来表示离散系数，其计算公式为

$$CV = \frac{s}{\bar{x}} \tag{5.2.11}$$

离散系数是测度数据离散程度的相对统计量。因此，它主要被用来比较具有不同均值水平的数据的波动性。一般而言，离散系数小的说明数据的波动性小，反之，离散系数大的则表明数据的波动性大。

例 5.12. 利用 bwght 中的数据，分组计算不抽烟的孕妇和抽烟的孕妇所生孩子的体重的标准差和标准差系数。并解释它们的含义。

解：首先对家庭收入分组，将孕妇在怀孕是不抽烟的家庭收入分为一组，将孕妇在怀孕期间抽烟家庭收入分为另一组。

```
inc0 <- bwght$faminc[bwght$cigs==0]
inc1 <- bwght$faminc[bwght$cigs>0]
```

现在，可以分组计算两组的均值和标准差。

```
m.inc0 <- mean(inc0)
m.inc1<- mean(inc1)
s.in0 <- sd(inc0)
s.in1 <- sd(inc1)
```

很明显，怀孕是不抽烟的孕妇的家庭收入的均值为 30.49，而怀孕是抽烟的孕妇的家庭收入为 20.92，两者的家庭收入存在较大差距。

最后，计算这两组家庭的收入的离散系数

```
cv0 <- round(100*s.in0/m.inc0,2)
cv1 <- round(100*s.in1/m.inc1,2)
```

经计算得到怀孕时不抽烟的孕妇的家庭收入的离散系数是 62.18%，怀孕时抽样的孕妇的家庭收入的离散系数是 72.39%。它们的含义是：若将出生的因而的体重假定为 100 盎司，则前者的波动程度为平均每 100 盎司波动 `r,cv0` 元，后者的波动程度为平均每 100 盎司波动 72.39 元，前者的波动程度要明显小于后者。

5.3 分布的形状

集中程度和离散程度是数据分布的两个重要特征，但要全面了解数据分布的特征，还需要知道数据分布的形状，偏度和峰度就是对分布形状进行测度的两个统计量。

5.3.1 偏度

偏度是对数据对称性的测度，由统计学家 Pearson 于 1895 年提出。

定义 5.9. 偏度又称偏态，是对数据分布不对称性的测度。

其实，判断数据分布是否对称并不困难，在前面，我们利用众数、中位数和平均数之间的关系，就可以判断数据分布的形状是否发生偏斜以及偏斜的方向。但若具体测度偏斜的程度，则需要计算偏度系数。

定义 5.10. 偏度系数是对数据分布不对称性的度量。

偏度系数一般用 S 表示。在一个样本容量为 n 的样本中，偏度经典的计算公式为

$$S = \frac{\frac{1}{n}\sum(x_i - \bar{x})^3}{(\frac{1}{n-1}\sum(x_i - \bar{x})^2)^{\frac{3}{2}}} \tag{5.3.1}$$

由偏度的计算公式可以发现，它是变量值和其均值的离差的三次方的平均数。若 $S = 0$，则表示数据的分布是对称的；若 $S > 0$，此时，S 值越大，数据的分布越往右偏；反之，则越往左偏。

5.3.2 峰度

峰度是对数据分布的尖峰或平峰程度的测度，它由统计学家 Pearson 于 1905 提出。

定义 5.11. 峰度指的是数据分布的平峰或尖峰程度。

对数据分布的尖峰程度的衡量，同样需要计算峰度系数。

定义 5.12. 峰度系数指的是对数据尖峰程度的测度。

峰度一般用 K 表示，其计算公式为

$$K = \frac{\frac{1}{n}\sum(x_i - \bar{x})^4}{(\frac{1}{n}\sum(x_i - \bar{x})^2)^2} - 3 \tag{5.3.2}$$

需要注意的是，在式 (5.3.2) 中，当 $K > 0$ 时表示的是数据的尖峰分布，反之，则表示数据的平峰分布。

例 5.13. 利用数据集 bwght 中的数据，计算家庭收入 faminc 的偏度和峰度。

解：计算变量 faminc 的偏度和峰度的 R 软件编程如下

```
rm(list=ls())
data(bwght)
n <- length(bwght$faminc)
m.finc <- mean(bwght$faminc)
s.finc <- sd(bwght$faminc)
sf1 <- sum((bwght$faminc-m.finc)^3)/n
sf2 <- s.finc^3
s <- round(sf1/sf2,3)
kf1 <- sum((bwght$faminc-m.finc)^4)/n
kf2 <- (sum((bwght$faminc-m.finc)^2)/n)^2
k <- round(kf1/kf2-3,3)
```

由此可得，家庭收入的偏度系数是 0.617，峰度系数是-0.527。

在 R 软件的 **fBasics** 包中，可以直接利用 *skewness()* 命令和 *kurtosis()* 命令直接计算变量的偏度和峰度。由于 **fBasics** 包并非 R 软件自带的宏包，因此，在使用该命令前需要安装它。

```
install.packages('fBasics')
```

安装完成后，可以利用

```
library(fBasics)
```

命令调用该包。利用该包中的命令直接计算偏度和峰度的命令如下

```
skew <- skewness(bwght$faminc)
kurt <- kurtosis(bwght$faminc)
```

同样可以得到家庭收入的偏度系数为 0.617，峰度系数为-0.53。

当然，计算偏度系数和峰度系数的另一种方法是采用自编公式的办法，即先编写计算偏度系数和峰度系数的计算公式，将其存入当前工作路径，使用的时候调用它即可。其具体步骤如下：

- 在 R 软件的菜单栏点击 **File**、**New File**、**R Script** 新建一个脚本文件；
- 在该脚本文件中编写自己的公式，其内容如下：

```
mysk <- function(x){
n <- length(x)
mean.x <- mean(x)
sd.x <- sd(x)
sf1 <- sum((x-mean.x)^3)/n
sf2 <- sd.x^3
sk <- round(sf1/sf2,3)
```

```
kf1 <- sum((x-mean.x)^4)/n
kf2 <- (sum((x-mean.x)^2)/n)^2
kurt <- round(kf1/kf2-3,3)
return(cbind(sk,kurt))
}
```

- 点击菜单栏上的 **File** 下拉菜单中的 **Save**，将其保存为后缀为 *.R* 的程序文件，其名称和公式的名称一致，这里应当是一个名为 `mysk.R`。
- 用 source('mysk.R') 命令调用该自编函数，即可计算变量的偏度系数和峰度系数。

这里应用自编函数计算变量 faminc 的偏度系数和峰度系数的编程如下：

```
source('mysk.R')
mysk <- mysk(bwght$faminc)
```

同样可以得到变量 faminc 的偏度系数为 0.617，峰度系数为-0.527。

第六章　参数估计

参数估计是推断统计学的重要组成部分。它在抽样基础上，根据样本统计量对总体参数进行推断。本章的主要内容包括：参数估计概述，主要介绍统计推断中的基本概念和方法；一个总体参数的区间估计主要有总体均值和总体比例的区间估计、总体方差的区间估计；两个总体参数的区间估计则包括总体均值或成数差的区间估计、总体方差之比的区间估计等；本章最后则对参数估计中的精度和样本容量的确定问题进行探讨。

6.1　参数估计概述

在统计调查中，在得到总体的全部数据情形下，只需要对总体进行简单的统计描述，就可以得到总体的数量特征，它们主要有总体均值、总体成数或总体标准差等。但在实际的统计调查中，在很多情形下，得到总体的全部数据既不可能，也没有必要。更多的时候，我们得到的仅仅是关于总体的一个抽样的样本数据，需要根据样本数据来推断总体参数。参数估计问题由此产生。

6.1.1　估计量、估计值和参数估计

估计量指的是用来估计总体参数的统计量。一般情况下，估计量可以用 $\hat{\theta}$ 来表示。常见的参数估计量有样本均值、样本比例、样本方差或标准差等。有的时候，估计量又被称为样本估计量或参数估计量。

估计值指的是在估计总体参数时，和某一特定样本相联系的估计量的具体数值。例如，为了估计某一班级学生的平均月消费支出，我们从中抽取一个随机样本，假定我们事先确定平均月消费支出的计算公式为

$$\bar{x} = \frac{\sum x_i}{n}$$

根据这一计算公式，我们可以计算出抽取的一个特定的随机样本的平均月消费支出为 1800 元。则在上述过程中，$\bar{x}$ 的计算公式就是一个估计量，而根据样本计算得到的 1800 元，则是一个估计值。由此可见，参数估计量指的是估计总体参数的方法，和具体的样本无关，而参数估计值则是根据参数估计量，利用特定的样本数据，计算出来的具体数值。因此，参数估计量是一个随机变量，会随着样本的变化而变化，而估计值则是指一个具体的数值。

参数估计指的是用样本估计量去估计总体参数。常见的参数估计包括：利用样本均值 $\bar{x}$ 去估计总体均值 μ；利用样本比例 p 去估计总体成数 π；利用样本标准差 s 去估计总体标准差 σ 等。一般情况下，总体参数可以用 θ 表示。参数估计就是用样本估计量 $\hat{\theta}$ 去估计总体参数 θ。

6.1.2 参数估计量的评判标准

从前面的论述可以看出，参数估计的本质是利用样本统计量 $\hat{\theta}$ 去估计总体参数 θ。实际上用于估计总体参数 (θ) 的样本统计量 $(\hat{\theta})$ 往往不止一个，例如，我们可以用样本均值作为总体均值的估计量，也可以用样本数据的众数作为总体均值的估计量，甚至可以用该样本中的任意一个数作为总体均值的估计量，等等。因此，在参数估计时，我们首先要确定，到底哪种估计量才是最好的估计量，这就需要给出对估计量的评判标准。统计学家给出了评价估计量的标准主要有三点，即无偏性、一致性和有效性。

1. 无偏性

无偏性指的是参数估计量的数学期望等于总体参数。假定总体参数为 θ，参数估计量为 $\hat{\theta}$，参数估计量的无偏性可以表示为

$$E(\hat{\theta}) = \theta \tag{6.1.1}$$

常见的无偏估计量很多，例如，样本均值 $\bar{x}$、样本比例 p 和样本标准差 σ 等，它们分别为总体均值 μ、总体比例 π 和总体标准差 σ 的无偏估计量。因此，有

$$E(\bar{x}) = \mu \qquad E(p) = \mu \qquad E(s) = \sigma$$

2. 有效性

参数估计量的**有效性**指的是，在所有的无偏估计量中，具有最小方差的参数估计量。一个无偏估计量并不意味着它就一定接近于总体参数。一个估计量是否靠近总体参数，不仅需要它具有无偏性，还需要他的离散程度较小，即具有较小的方差或标准差。假定一个参数的所有无偏估计量为 $\hat{\theta}_1$，$\hat{\theta}_2$，……，$\hat{\theta}_n$，其中若 $\hat{\theta}_i$ 具有最小的方差，则称该估计量是参数 θ 的无偏估计量。在实际的参数估计中，我们往往无法找出所有的无偏估计量。例如，假定用于估计总体参数 θ 的无偏估计量分别为 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，它们的抽样分布的方差分别为 $D(\hat{\theta}_1)$ 和 $D(\hat{\theta}_2)$，若 $\hat{\theta}_1$ 具有较小的方差，即 $D(\hat{\theta}_1) < D(\hat{\theta}_2)$，则称 $\hat{\theta}_1$ 是一个比 $\hat{\theta}_2$ 更有效的估计量。在无偏估计量的假定前提下，若估计量的方差越小，那么该估计量就越有效。

3. 一致性

有效性指的是随着样本容量的增大，点估计量的值越来越接近被估计的总体参数。

也就是说，对于同一参数进行估计，和从较小的样本得到的参数估计量相比，从一个较大样本得到的参数估计量，要更加接近总体的参数。根据中心极限定理可得，样本均值的抽样分布的标准差为

$$\sigma_x = \sigma/\sqrt{n}$$

由此可见，样本均值的标准差和样本容量的大小相关，样本容量越大，样本均值的标准差就越小，因此，大样本容量给出的估计量就更接近于总体均值 μ。从这个意义上说，样本均值是总体均值的一个一致估计量。

6.1.3 参数估计的方法

参数估计的方法有两种，一种是点估计，另一种是区间估计。

1. 点估计

点估计指的是直接将样本估计量 $\hat{\theta}$ 的值当作总体参数 θ 的估计值。

点估计的优点是方法简单。例如，可以直接将样本均值 $\bar{x}$ 作为总体均值 μ 的估计值；将样本比例 p 作为总体比例 π 的估计值；将样本方差 s^2 作为总体方差 σ^2 的估计值，等等。假定要估计一个班级学生的平均月生活费支出，根据抽取的一个随机样本计算得到的平均月生活费支出为 1600 元，就可以用 1600 元作为全班学生的平均月生活费支出的估计值。这就是点估计。同样的道理，假定要估计一批产品的合格率，根据抽样的结果得到的合格率为 90%，那么，就可以将 90% 直接作为这批产品合格率的估计值，这也是点估计。

但点估计的缺点也十分明显。在用点估计值代表总体参数值的时候，点估计无法给出一个用于衡量点估计值可靠程度的度量。也就是说，点估计无法给出点估计值和总体参数的真实值之间接近程度的信息。要做到这一点，需要对总体参数进行区间估计。

2. 区间估计

区间估计通过从总体中抽取的样本，根据一定的精确程度与置信程度的要求，构造出适当的区间，以作为总体参数在范围的估计。

在区间估计中，有两个重要的概念，一是精确程度，二是置信程度。所谓精确程度，指的是但估计值和参数真实值之间的差异，即点估计值对参数真实值的接近程度；所谓置信程度，则是表明该估计的把握程度，即在多次这样的实验中，通过该方法所构造的区间中，包含参数真实值的实验所占的比例。其他条件不变时，若想提高精确程度，则必然会降低把握程度。

下面，我们以总体均值为例来说明区间估计的基本原理。

由大数定律和中心极限定量可知，在可重复抽样或无限总体的大样本抽样条件下，样本均值的数学期望值等于总体均值，即 $E(x) = \mu$，样本均值的标准差为 $\sigma_x = \sigma/\sqrt{n}$。因此，样本均值 $\bar{x}$ 落在总体均值 μ 两侧各一个抽样标准差范围内的概率为 67.83%，落在总体均值两侧各两个抽样标准差范围内的概率为 95.45%，落在总体均值两侧各三个抽样标准差范围内的概率为 99.73%，等等。我们可以求出样本均值落入总体均值 μ 的两侧任意一个抽样标准差房屋内的概率。也就是说，在总体均值已知时，给定了概率，我们可以用总体均值构造出样本均值落入的区间。

但在实际抽样估计中，情形恰好相反，总体均值是一个未知的待估计参数，已知的是样本均值 $\bar{x}$。由于样本均值 $\bar{x}$ 和总体均值 μ 之间的距离是对称的，即若样本均值落入以总体均值 μ 为中心的左右两侧各两个标准差范围之内，那么，总体均值 μ 也应当落入以样本均值 $\bar{x}$ 为中心的左右各两个标准差范围之内。因此，若以样本均值为中心构造区间，则总体参数 μ 落入样本均值左右两侧各 1 个标准差范围内的概率约为 67.83%，落入样本均值左右两侧各两个标准差范围捏的概率约为 95.45%，等等。因此，可以给出置信区间的定义如下：

置信区间指的是由样本统计量所构造的总体参数的估计区间。其中，区间的最小值被称为置信下限，区间的最大值被称为置信上限。

之所以将以样本均值为中心构造的区间称为置信区间，其根本原因在于统计学家在某种程度上确信这个区间会包含总体参数。例如，统计学家有 95.45% 的把握程度认为，总体参数会落入以样本均值为中心，左右各两倍标准差的范围之内，其中，95.45% 就是统计学家的把握程度。因此，通常又将该把握程度称为置信程度或置信水平。置信水平为 95.45% 的含义为：若我们抽取

表 6.1.1: 常用置信水平及 $z_{1-\alpha/2}$

置信水平 (%)	α	$\alpha/2$	$z_{1-\alpha/2}$
90	0.10	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.58

100 个这样的样本，利用这 100 个样本的样本均值，以及抽样误差的 2 构建 100 个以样本均值为中心的的区间，那么，大约有 95 个这样的区间中包含了总体参数的真实值。因此，也可以用概率来定义置信水平。

置信水平指的是若我们将构造置信区间的步骤重复多次，则置信区间中包含总体参数真实值的次数所占的比率，就是置信水平或置信系数。

在构造以样本均值为中心的置信区间时，我们可以预先设定所希望的置信水平的值。比较常用的置信水平值、对应的 α 值、单侧的概率 $\alpha/2$ 和对应的右侧临界值 $z_{1-\alpha/2}$ 如表 (6.1.1) 所示：

对于置信区间的理解，需要注意以下几点：

首先，置信区间是一个随机区间，他会因抽取样本的不同而有所差异，而总体参数的真实值是一个未知的固定值，因此，在利用样本信息构造置信区间时，并不是所有的区间都包含真实值。

其次，如果用某种方法构造的所有区间中，有 95% 的区间包含总体参数的真实值，5% 的区间不包含总体参数的真实值，那么用该方法构造的区间称为置信水平为 95% 的置信区间。同样的道理，也可以用这样的方法构造任意一个置信水平的区间。

最后，置信水平是统计规律，只有在大量实验的基础上才能得出，但在实际抽样估计时，往往只抽取一个样本，所构造的是和该样本相联系的一个特定的置信区间。因此，我们无法知道由这个样本所构造的置信区间是否包含总体参数的真实值。所以，我们只能希望我们所构造的这个特定的区间，是大量包含总体参数真实值的区间中的一个，但该区间也可能是少数几个不包含

参数真实值的区间中的一个。

下面，我们以一个案例来说明置信区间的真实含义。假定我们要对某一学校学生每月平均的生活费支出。利用样本构建的每月平均的生活费用的 95% 的置信区间为 1600–1800 元之间。它的真实含义是，若这样的抽样实验进行 100 次，根据样本构造的区间中包含参数真实值的次数大概是 95 次，根据样本构造的区间中不包含参数真实值的次数大概有 5 次。但在某一次构造的区间中，是否包含总体参数的真实值，只有两种结果，要么包含总体参数的真实值，要么不包含总体参数的真实值，而不是以 95% 的可能性包含总体参数的真实值。也就是说，若该校学生每月平均的生活费支出是 1650 元，则该置信区间就包含了总体的真实值，反之，若该校学生每月平均生活费支出为 1550 元，则该置信区间就不包含总体参数的真实值。

显然，区间估计比点估计的应用要更广泛，在本章后面的章节中，我们所进行的参数估计，基本上都是建立在区间估计基础上的。

，

6.2 一个总体参数的区间估计

当我们研究一个总体时，所关心的参数主要有：总体均值 (μ)、总体成数 (π) 以及总体方差 (σ^2)。这一节的主要内容是利用样本统计量来构造总体参数的置信区间。

6.2.1 总体均值的区间估计

在对总体均值进行区间估计时，我们要考虑的因素主要有：总体是否服从正态分布、总体方差是否已知、用于构造总体参数置信区间的样本是否是大样本。由大数定律和中心极限定理，我们可以将上述因素归纳为正态总体，方差已知、正态总体，方差未知和非正态总体，大样本等三种情形。

1. 正态总体，方差已知

当总体服从正态分布且总体方差 σ^2 已知时，无论是大样本抽样还是小样本抽样，样本均值 $\bar{x}$ 都服从正态分布，即

$$\bar{x} \sim N(\mu, \sigma^2/\sqrt{n})$$

将样本均值标准化后可得：

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6.2.1}$$

根据标准正态分布的性质，可以构造出总体均值 μ 的置信水平为 $1-\alpha$ 的置信区间为

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.2.2}$$

其中，$\bar{x}$、σ 和 n 本别表示样本均值、总体方差和样本容量；α 是我们事先确定的一个概率值，也被称为风险值，它是置信区间中不包括总体均值 μ 的概率；$1-\alpha$ 是置信水平；$z_{1-\alpha/2}$ 是标准正态分布曲线下右侧面积为 $\alpha/2$ 时的 z 值；$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ 是估计总体均值时的估计误差或误差范围。

由式 (6.2.2) 可以看出，总体均值的置信区间由两部分组成：点估计值和误差范围。其两个边界分别是置信下限 $\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$，置信上相 $\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$。

若总体服从正态分布，且总体方差已知，即使是在小样本抽样条件下，仍然可以得到样本均值 $\bar{x} \sim N(\mu, \sigma^2/\sqrt{n})$，在给定的置信水平 $1-\alpha$ 下，同样可以像式 (6.2.2) 那样构造总体均值的置信区间。

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

例 6.1. 利用 R 软件自带的数据集 mtcars，估计变量 mpg 的均值的 95% 置信区间。（假定总体方差为 1600）

解：利用 R 进行区间估计的程序如下

```
n <- length(mtcars$mpg)
alpha <- 0.05
sigmasq <- 35
data(mtcars)
m.mpg <- mean(mtcars$mpg)
value.z <- qnorm(1-alpha/2)
se <- sqrt(sigmasq/n)
dev <- value.z*se
l.value <- m.mpg-dev
u.value <- m.mpg+dev

cbind(l.value,m.mpg,u.value)
```

```
##      l.value m.mpg u.value
## [1,]      18  20.1    22.1
```

2. 正态总体，方差未知

当总体服从正态分布时，无论是大样本抽样还是小样本抽样，在方差已知的条件下，样本均值都服从正态分布，就可以用公式 (6.2.2) 去构造总体均值的 $1-\alpha$ 的置信区间。但当总体方差未知时，则需要用样本方差去替代总体方差，因此，可得统计量

$$t = \frac{\bar{x}-\mu}{\hat{\sigma}^2/\sqrt{n}} \sim t(n-1) \tag{6.2.3}$$

这里需要注意的是，样本统计量 t 不再服从正态分布，而是服从自由度为 $(n-1)$ 的 t 分布。给定置信水平 $1-\alpha$，可以构造总体均值的 $1-\alpha$ 置信区间为

$$\bar{x} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \tag{6.2.4}$$

其中，s 表示样本的标准差。

这里需要进行说明的是：首先，当给定置信水平 $1-\alpha$ 后，可以利用 R 软件计算出 $t_{1-\alpha/2}(n-1)$ 的值，这一点我们在例题中会有说明；另外，当样本容量 n 增加到一定程度后，$t_{1-\alpha/2}(n-1)$ 的值逐渐向标准正态分布靠拢，也就是说，当样本容量足够大，我们可以用 z 统计量去替代 t 统计量。但这种替代并不适用于所有大样本条件，这里可以进行一个简单的说明，例如，当样本容量 n 为 40 时，给定置信水平 $1-\alpha=095$，相应的 z 统计量的临界值为 1.96，而此时相应的 t 统计量的临界值为 2.023。由此可见，两者的差异还是比较明显的。但比较遗憾的是，很多流行的《统计学》教材，在大样本条件下就直接采用 z 统计量去替代 t 统计量，这种做法明显是不足取的。

例 6.2. 利用 mtcars 数据集，估计 mpg 的均值的 95% 的置信区间。

解：利用 R 软件计算 mpg 均值的 95% 的置信区间的程序如下

```
n <- length(mtcars$mpg)
alpha <- 0.05
m.mpg <- mean(mtcars$mpg)
sd.mpg <- sd(mtcars$mpg)
tc <- qt(1-alpha/2,n-1)
dev <- tc*sd.mpg/sqrt(n)
l.mpg <- m.mpg-dev
u.mpg <- m.mpg+dev
result.t <-cbind(l.mpg,m.mpg,u.mpg)
```

这里需要注意的是，由于在本例中，样本容量已经达到 32，属于大样本抽样，若用 z 统计量来计算，其结果如下：

表 6.2.1: 两种方法估计的结果比较

	mpg 的下限	mpg 的均值	mpg 的上限
t 统计量	17.9	20.1	22.3
z 统计量	18.0	20.1	22.2

```
zc <- qnorm(1-alpha/2)
dev.z <- zc*sd.mpg/sqrt(n)
l.mpg<- m.mpg-dev.z
u.mpg <- m.mpg+dev.z
result.z <- cbind(l.mpg,m.mpg,u.mpg)
result <- rbind(result.t,result.z)
```

为了更直观的比较两种不同计算方法的结果，将它们用表格的形式输出如下：

例 6.3. 利用 wooldridge 包中的 bwght 数据集，计算变量 bwght 的均值的 95% 的置信区间。

解：在本例中，虽然总体方差未知，但由于样本容量为 1388，因此，即使总体方差未知，也可以用 z 统计量去估计变量 bwght 的置信区间。利用 R 语言估计如下

```
library(wooldridge)
data(bwght)
n <- length(bwght$bwght)
alpha <- 0.05

m.w <- mean(bwght$bwght)
```

```
sd.w <- sd(bwght$bwght)
z.val <- qnorm(1-alpha/2)
dev <- z.val*sd.w
l.w <- m.w-dev
u.w <- m.w+dev
cbind(m.w,l.w,u.w)
```

```
##      m.w  l.w u.w
## [1,] 119 78.8 159
```

3. 非正态总体，大样本

当总体不服从正态分布时，在现有的知识体系内，我们没有办法解决小样本抽样下的总体均值估计问题。因此，这里我们仅仅讨论大样本抽样的情形。假定总体的均值存在但是一个未知的参数 μ，总体的方差为 σ^2，根据中心极限定理，样本均值 $\bar{x} \sim N(\mu, \sigma^2/\sqrt{n})$，将其标准化可得

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{6.2.5}$$

因此，当总体不服从正态分布时，若方差已知，则在大样本抽样条件下，可以利用 z 统计量来构造总体均值 μ 的置信区间。假定置信水平为 $1-\alpha$，则其置信区间可以表示为

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.2.6}$$

其置信下限为 $\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$，期置信上限为 $\bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$。

若总体方差 σ^2 是一个未知数，则需要用样本方差 s^2 去替代，此时，样本统计量

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1) \tag{6.2.7}$$

也就是说，当总体方差未知时，构造的样本统计量不再服从标准正态分布，而是服从自由度为 $n-1$ 的 t 分布。在给定的置信水平 $1-\alpha$ 下，可以构造总体均值的置信区间为

$$\bar{x} \pm t_{1-\alpha/2}(n-1)\frac{s}{\sqrt{n}} \tag{6.2.8}$$

同样，若样本容量足够大，可以直接用 z 统计量替代 t 统计量构造总体均值的 $1-\alpha$ 置信区间。

将总体均值区间估计的各种情形总结如表 (6.2.2) 所示：

表 6.2.2: 总体均值区间估计的不同情形

总体分布	样本容量	σ 已知	σ 未知
正态总体	大样本（$n \geq 30$）	$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
	小样本（$n < 30$）	$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
非正态总体	大样本（$n \geq 30$）	$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

6.2.2 总体成数的区间估计

在总体成数估计中，我们只讨论大样本情形下的总体成数的区间估计问题。根据大数定律和中心极限定理，当样本容量足够大时，比例 p 的抽样分布可以用正态分布近似。[1]假定总体成数为 π，因此有

[1]在总体成数的区间估计值，确定样本容量是否足够大的一般经验法则为：$np \geq 5$ 且 $n(1-p) \geq 5$；或区间 $p \pm 2\sqrt{p(1-p)/2}$ 中不包括 0 和 1。因此，在本届中，无论总体方差是否已知，我们都可以用一个服从标准正态分布的随机变量 z 来构造总体成数置信区间

$$E(p)=\pi \qquad \sigma_p^2=\frac{\pi(1-\pi)}{n}$$

对其标准化后可得服从标准正态分布的随机变量

$$z=\frac{p-\pi}{\sqrt{\pi(1-\pi)/n}}\sim N(0,1) \tag{6.2.9}$$

和总体均值的区间估计相似，在给定的置信水平 $1-\alpha$ 下，由样本成数 p 可得总体成数 π 的 $1-\alpha$ 置信区间为

$$p\pm z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \tag{6.2.10}$$

在利用式 (6.2.10) 计算总体成数的置信区间时，总体方差应当是已知的。但在实际估计中，由于 π 恰好是我们要估计的总体参数，因此，是未知的，在这种情形下，可以用样本成数 p 来替代总体成数，从而得到总体比例的 $1-\alpha$ 置信区间

$$p\pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \tag{6.2.11}$$

由该表达式可以看出，和总体均值的区间估计一样，总体成数的区间估计也由两部分构成：一部分是点估计值；另一部分是和置信水平 $1-\alpha$ 相联系的估计误差。

需要特别说明的是，在总体成数的区间估计中，要求样本容量足够大，根据前面提到的经验法则可知，该样本容量的大小和 p 的大小有关，当样本成数 p 接近 0.5 时，用较小的样本就可以使得样本成数 p 的分布趋于正态分布，反之，当样本成数 p 的值越接近 0 和 1，则需要更大的样本容量，才能保证 p 的分布趋于正态分布。在总体成数的估计实践中，一个可供参考的标准如表 (6.2.3) 所示：

表 6.2.3: 比例近似服从正态分布的样本容量要求

样本成数 (p)	近似服从正态分布所要求的样本容量
0.5	30
0.4∼0.6	50
0.3∼0.7	80
0.2∼0.8	200
0.1∼0.9	600

例 6.4. 利用 wooldridge 包中的 bwght 数据集，推断当年出生的小孩是家中第二个孩子的比例的 95% 的置信区间。

解： 在数据集 **bwght** 中，变量 *parity* 表示的是出生的婴儿是家中第几个孩子。这里，我们要推断的是出生的婴儿是家中第二个孩子的置信区间，因此，首先需要进行点估计：

```
library(wooldridge)
data(bwght)
n <- length(bwght$parity)
n.s <- sum(bwght$parity==2)
portion <- n.s/n
```

由此，我们得到样本的点估计值为 28.03。在此基础上，我们可以对总体中出生婴儿是家中第二个孩子所占的比例进行推断：

```
alpha <- 0.05
s <- sqrt(portion*(1-portion)/n)
zc <- qnorm(1-alpha/2)
dev <- zc*s
```

```
l.p <- portion-dev
u.p <- portion+dev
result <- cbind(l.p,portion,u.p)
```

由估计的结果可以发现，出生的婴儿中是家中第二个孩子所占的比重的 95% 的置信区间为 (25.66 30.39)。由此我们也可以看出，在美国，超过四分之一的婴儿是家中的第二个孩子。

6.2.3 总体方差的区间估计

当总体服从正态分布时，样本方差服从自由度为 $n-1$ 的 χ^2 分布，即

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \tag{6.2.12}$$

在给定的置信水平 $1-\alpha$ 下，由卡方分布的特征可知，样本统计量满足

$$\chi^2_{\alpha/2}(n-1) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{1-\alpha/2}(n-1) \tag{6.2.13}$$

因此，总体方差 σ^2 的 $1-\alpha$ 的置信区间为

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)} \tag{6.2.14}$$

例 6.5. 利用 wooldridge 包中的数据集 bwght，根据变量 bwght 的信息，估计出生婴儿体重的方差的 95% 的置信区间。

解： 首先计算 bwght 的方差估计量

```
sd.bw <- sd(bwght$bwght)
n <- length(bwght$bwght)
hat.sg <- (n-1)*sd.bw^2
alpha <- 0.05
l.sqsig <- hat.sg/qchisq(1-alpha/2,n-1)
u.sqsig <- hat.sg/qchisq(alpha/2,n-1)
```

由此，可以得到出生婴儿体重方差的 95% 的置信区间为 (385.1　446.93)。

6.3 两个总体参数的区间估计

在两个总体的区间估计中，常见的情形包括：

- 两个总体的均值之差 $\mu_1 - \mu_2$；
- 两个总体的成数之差 $\pi_1 - \pi_2$；
- 两个总体的方差之比 σ_1^2/σ_2^2。

6.3.1 两个总体均值之差的区间估计

两个总体均值之差的区间估计的基本原理是：若两个总体的均值是两个未知参数 μ_1 和 μ_2，为了估计两个总体均值的差，从两个总体中分别抽取样本容量为 n_1 和 n_2 的随机样本，其样本均值为 $\bar{x}_1$ 和 $\bar{x}_2$。而样本均值之差 $\bar{x}_1 - \bar{x}_2$ 是总体均值之差 $\mu_1 - \mu_2$ 的估计量。

在实际应用中，根据所抽取的样本的方法不同，可分为独立样本和配对样本两种情况。

1. 独立样本

若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和另一个样本中的元素无关，则称这两个样本为独立样本。独立样本的总体均值之差的区间估计中，主要考虑以下四种情形。

1）独立大样本

若两个总体都服从正态分布，或抽取的两个样本都是大样本时，则有：

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

将两个样本均值之差标准化后得到

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{6.3.1}$$

当总体方差 σ_1^2 与 σ_2^2 已知时，在给定的显著性水平 $1-\alpha$ 条件下，两个总体均值之差 $\mu_1 - \mu_2$ 的置信区间为

$$(\bar{x}_1 - \bar{x}_2) \mp z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.3.2}$$

当总体方差未知时，可以用它们的样本方差替代总体方差，可以得到

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(n_1 + n_2 - 2) \tag{6.3.3}$$

在给定的置信水平 $1-\alpha$ 条件下，同样可得总体均值的置信区间为：

$$(\bar{x}_1 - \bar{x}_2) \mp t_{1-\alpha/2}(n_1 + n_2 - 2)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.3.4}$$

当样本容量足够大时，也可以直接用6.3.2计算样本总体均值之差从 $1-\alpha$ 的置信区间。

例 6.6. 在 wooldridge 包中，有数据集 wage1，利用该数据集计算男性和女性工资差异的 95% 的置信区间。

解：在数据集 wage1 中，变量 female 表示的是性别，其中，female 取值为 1 时，表示该对象是女性，取值为 0 时，表示该对象是男性。因此，首先应当将数据集分成两组。一种一组为女性，另一组是男性。

```
fdata <- subset(wage1[,c("wage","female")],wage1$female==1)
mdata <- subset(wage1[,c("wage","female")],wage1$female==0)
nf <- length(fdata$wage)
nm <- length(mdata$wage)
```

利用 R 软件，可以求出数据集中有女性 252 人，男性 274 人，假定抽样时个观测对象之间是相互独立的，可以认为这是两个独立的样本。在本例中，由于总体方差是未知的，需要用样本的标准差对其进行估计，又因为样本容量已经足够大，可以直接利用 z 统计量来计算。首先计算两个独立样本的平均工资和标准差。

```
wage.f <- mean(fdata$wage)
wage.m <- mean(mdata$wage)
sd.fw <- sd(fdata$wage)
sd.mw <- sd(mdata$wage)
```

现在，我们可以计算两个总体均值之差的区间估计。

```
alpha <- 0.05
zvalue <- qnorm(1-alpha/2)
dev.w <- wage.m-wage.f
```

```
hat.sgm <- sqrt(sd.fw^2/nf+sd.mw^2/nm)
error <- zvalue*hat.sgm
l.devw <- dev.w-error
u.devw <- dev.w +error
results <- cbind(l.devw,dev.w,u.devw)
```

由上面程序的结果可知，在这两个独立样本中，男性的小时工资和女性的小时工资的差的点估计是 2.51 美元。男性和女性之间的工资差的 95% 的置信区间为 (1.93 3.1)。

2) 独立小样本，方差相等

若两个独立的样本是小样本，为了估计两个总体均值的差，需要对两个总体进行一定的假设：

- 两个总体都服从正态分布；
- 两个随机样本分别独立地抽自两个总体。

在上述假定条件下，无论样本容量是多少，两个样本均值的差都服从正态分布，当两个总体方差已知且相等时，可以用式 (6.3.2) 来构建两个总体均值之差的置信区间。

当两个总体的方差 σ_1^2 和 σ_2^2 相等但未知时，需要用两个样本的方差 s_1^2 和 s_2^2 来估计总体方差。具体的做法是将两组数据合在一起，计算总体方差的合并估计量 s_p^2，其计算公式为：

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \tag{6.3.5}$$

由此可得一个自由度为 n_1+n_2-2 的 t 统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2) \tag{6.3.6}$$

在给定的置信水平 $1-\alpha$ 条件下，可以构建出两个总体均值之差的置信区间：

$$(\bar{x}_1 - \bar{x}_2) \mp t_{1-\alpha/2}(n_1 + n_2 - 2)s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{6.3.7}$$

例 6.7. 在 R 软件自带的数据集 mtcars 中，给出了随机抽样条件下 4 缸发动机和 8 缸发动机的油耗 (英里/加仑)。假定 4 缸发动机和 8 缸发动机的油耗都服从正态分布，且总体方差相同。试以 95% 的置信水平构建两个总体平均油耗之差的置信区间。

解：首先计算两个独立样本的均值和方差：

```
data(mtcars)
mpgf <- subset(mtcars[,c('mpg','cyl')],mtcars$cyl==4)
mpge <- subset(mtcars[,c('mpg','cyl')],mtcars$cyl==8)
m.f <- mean(mpgf$mpg)
m.e <- mean(mpge$mpg)
sd.f <- sd(mpgf$mpg)
sd.e <- sd(mpge$mpg)
```

由于已经假定 4 缸发动机和 8 缸发动机的油耗的方差相等，则可以利用数据计算总体方差的合并估计量。

```
n.f <- length(mpgf$mpg)
n.e <- length(mpge$mpg)
sq.sp <- ((n.f-1)*sd.f^2+(n.e-1)*sd.e^2)/(n.f+n.e-2)
```

```
alpha <- 0.05
tval <- qt(1-alpha/2,n.f+n.e-2)
```

在给定了显著性水平条件下，可以得到两个总体均值差的置信区间。

```
alpha <- 0.05
tval <- qt(1-alpha/2,n.f+n.e-2)
dev.s <- (m.f-m.e)
error <- tval*sqrt(sq.sp*(1/n.f+1/n.e))
l.dev <- round(dev.s-error,2)
u.dev <- round(dev.s+error,2)
```

由此可得，两个总体的平均油耗之差的 95% 的置信区间为 (8.61 14.52)。其含义为平均而言，和 8 缸发动机汽车相比，4 缸发动机每加仑的油要多跑 8 到 14 英里。[2]

3) 独立小样本，样本容量相同

独立样本的第三种情形是总体服从正态分布，但两个总体的方差不等且未知，但两个样本的样本容量相同，即 $n_1 = n_2 = n$。在给定的 $1-\alpha$ 的置信水平下，可以得到两个总体均值之差的置信区间：

$$(\bar{x}_1 - \bar{x}_2) \mp t_{1-\alpha/2}(n_1+n_2-2)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{6.3.8}$$

例 6.8. 利用数据集 wage1，从中抽取个男性和女性 20 名，假定两个群体的工资都服从正态分布，但方差不等，估计两类群体工资均值之差的 95% 的置信区间。

[2]需要注意的是这里的计量单位，mpg 的含义是每加仑行驶的英里数，它是一个逆指标。在具体分析前可以利用 help(mtcars) 命令来查看该数据集的帮助文件。

解：首先将数据集 wage1 分成两组，其中一组是女性，另一组是男性。

```
data(wage1)
data.f <- subset(wage1[,c('wage','female')],wage1$female==1)
data.m <- subset(wage1[,c('wage','female')],wage1$female==0)
```

为了使得样本容量相等，利用 R 软件对这两组数据抽样，抽取的样本容量为 $n_1 = n_2 = 20$。

```
n <- 20
set.seed(123456)
data.sf <- data.f[sample(1:nrow(data.f),n,replace = T),]
set.seed(12345)
data.sm <- data.m[sample(1:nrow(data.m),n,replace = T),]
```

计算两个样本的均值和标准差

```
m.fw <- mean(data.sf$wage)
sd.fw <- sd(data.sf$wage)
m.mw <- mean(data.sm$wage)
sd.mw <- sd(data.sm$wage)
```

现在，可以估计两个总体均值差的置信区间

```
alpha <- 0.05
tval <- qt(1-alpha/2,2*n-2)
dev.w <- m.mw-m.fw
```

```
m.sd <- sqrt(sd.mw^2/n+sd.fw^2/n)
l.dev <- round(dev.w-tval*m.sd,2)
u.dev <- round(dev.w+tval*m.sd,2)
```

由此可得总体均值之差的 94% 的置信区间为 (1.28 3.86)。

4). 独立小样本，方差和样本容量皆不等

最后一种情形是总体服从正态分布的小样本抽样，但两个总体的方差 $\sigma_1^2 \neq \sigma_2^2$，且样本容量 $n_1 \neq n_2$。此时，两个样本均值之差经标准化后不再服从自由度为 $n_1 + n_2 - 2$ 的 t 分布，而是近似服从自由度为 v 的 t 分布，其中，自由度 v 的计算公式为:

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}} \tag{6.3.9}$$

在给定的 $1-\alpha$ 条件下，可以得到两个总体均值之差的置信区间为：

$$(\bar{x}_1 - \bar{x}_2) \mp t_{1-\alpha/2}(v)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.3.10}$$

例 6.9. 利用前面的 mtcars 数据集，现假定 4 缸发动机和 8 缸发动机油耗的方差并不相等，估计两种其次油耗均值的 95% 的置信区间。

解：在本例中，首先应当求出总体均值之差经标准化后的自由度:[3]

```
n.f <- length(mpgf$mpg)
n.e <- length(mpge$mpg)
m.f <- mean(mpgf$mpg)
```

[3]在计算自由度是，最后的结果用了四舍五入的方法，用 ceiling() 函数对其取整。

```
m.e <- mean(mpge$mpg)
dif.mpg <- m.f-m.e
sd.f <- sd(mpgf$mpg)
sd.e <- sd(mpge$mpg)
v1 <- (sd.f^2/n.f+sd.e^2/n.e)^2
dev.m <- v1^(1/4)
v2 <- (sd.f^2/n.f)^2/(n.f-1)+(sd.e^2/n.e)^2/(n.e-1)
v <- ceiling(v1/v2)
```

经过计算，可知样本均值差经过标准化后，近似服从自由度为 15 的 t 分布。

```
alpha <- 0.05
tval <- qt(1-alpha/2,v)
l.dev <- round(dif.mpg-tval*dev.m,2)
u.dev <- round(dif.mpg+tval*dev.m,2)
```

因此，两个群体总体均值之差的 95% 置信区间为 (8.32　14.81)。

2. 配对样本

由于抽样误差的存在，使用度量样本估计两个总体均值之差存在潜在的弊端。例如，在教学方法优良性的评估中，若使用独立样本，有可能抽取的一组对象都是学习能力比较强的学生，使得利用该方法教学的测试分数更高，但引起这种差异的原因不在于教学方法的优劣，而在于学习能力的强弱。因此，依据这两个样本得出的结论自然是错误的。

为了克服使用独立样本带来的风险，可以使用配对样本。即用同一组对象进行两次比照实验，使得两次得到的实验结果与实验对象之间配对，可以有效地消除因样本抽样的误差造成的差异。

使用配对样本进行两个总体均值之差的估计时，在大样本条件下，两个总体均值之差 $\mu_d = \mu_1 - \mu_2$ 的 $1-\alpha$ 置信区间为

$$\bar{d} \mp z_{1-\alpha/2}\frac{\delta_d}{\sqrt{n}} \tag{6.3.11}$$

其中，d 表示两个配对样本对于数据的差值；$\bar{d}$ 则表示差值的均值；σ_d 表示差值的标准差，当总体的标准差未知时，可以用样本的差值标准差 s_d 来代替。

在小样本情形下，假定两个总体各观测值的配对差服从正态分布，因此，两个总体均值之差 $\mu_d = \mu_1 - \mu_2$ 在置信水平为 $1-\alpha$ 时的置信区间为

$$\bar{d} \mp t_{1-\alpha/2}\frac{s_d}{\sqrt{n}} \tag{6.3.12}$$

例 6.10. 假定用 10 学生组成一个随机样本，让他们分别采用 A 和 B 两套试卷进行测试，成绩如表所示。是以 95% 的置信程度建立两套试卷平均分之差的置信区间。

解： 首先将数据输入 R 软件

```
paper.a <- c(78,63,72,89,91,49,68,76,85,55)
paper.b <- c(71,44,61,84,74,51,55,60,77,39)
```

现在可以计算配对数据的差值、差值的均值和样本的标准差。

```
diff <- paper.a-paper.b
m.diff <- mean(diff)
sd.diff <- sd(diff)
```

在给定的置信水平下，构造均值差值的置信区间

表 6.3.1: 10 名学生两套试卷的得分情况

学生编号	试卷 A	试卷 B	差值
1	78	71	7
2	63	44	19
3	72	61	11
4	89	84	5
5	91	74	17
6	49	51	-2
7	68	55	13
8	76	60	16
9	85	77	8
10	55	39	16

```
n <-10
alpha <- 0.05
tval <- qt(1-alpha/2,n-1)
ssig <- sd.diff/sqrt(n)
l.diff <- round(m.diff-tval*ssig,2)
u.diff <- round(m.diff+tval*ssig,2)
```

由此可得总体均值之差的 95% 置信区间为 (6.33　15.67)。

6.3.2 两个总体比例之差的区间估计

根据大数定律和中心极限定理，从两个二项分布总体中抽出的两个独立的大样本时，两个样本比例之差的抽样分布服从正态分布。因此，将两个样本的比例之差标准化后则近似服从标准正态分布。即

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1) \tag{6.3.13}$$

在给定的置信水平 $1-\alpha$ 条件下，可以构造相应的置信区间。

但在实际应用中，由于两个总体比例 π_1 和 π_2 通常是未知的，一般用样本比例 p_1 和 p_2 来代替，在大样本条件下，仍然能够得到如式 (6.3.13) 一样的统计量，在 $1-\alpha$ 给定的条件下，可以得到两个总体比例之差的置信区间为

$$(p_1 - p_2) \mp z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n1} + \frac{p_2(1-p_2)}{n_2}} \tag{6.3.14}$$

例 6.11. 利用数据集 bwght，可以研究孕妇抽样对婴儿出生的性别有无影响，在置信水平为 95% 的条件下，估计抽烟孕妇组的出生婴儿的男性所占比例和不抽烟孕妇组的出生婴儿的男性所占比例的差的置信区间。

解： 首先对数据分组

```
data(bwght)
data1 <- subset(bwght[,c("cigs","male")],bwght$cigs==0)
data2 <- subset(bwght[,c("cigs","male")],bwght$cigs>0)
```

现在可以计算各组出生婴儿中男性的比例

```
n1.male <- sum(data1$male==1)
n1 <- length(data1$male)
p1 <- n1.male/n1
n2.male <- sum(data2$male==1)
n2 <- length(data2$male)
p2 <- n2.male/n2
```

利用上面的数据，可以计算出两组出生婴儿中男性所占的比例之差的 95% 的置信区间

```
sd <- sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
alpha <- 0.05
zval <- qnorm(1-alpha/2)
l.p <- round(100*((p1-p2)-zval*sd),3)
u.p <- round(100*((p1-p2)+zval*sd),3)
```

利用数据计算得到，两组出生婴儿中男性所占的比例之差的 95% 的置信区间为 (-1.494　13.107)。在下一章，我们将对这一结果进行深入地讨论。这里只想说明这一结果的实际含义，即孕妇是否抽烟只影响婴儿的健康，而不影响婴儿的性别。

6.3.3 两个总体方差之比的区间估计

比较两个总体的方差问题，也是推断统计学中一个重要的应用。例如，我们想比较两台不同机器生产产品的稳定性，或比较两种不同测量工具的测量精度等。

在大样本条件下，由于两个样本的方差之比服从 $F(n_1-1, n_2-1)$ 分布，即

$$F = \frac{s_1^2}{s_2^2} \bullet \frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1, n_2-1) \tag{6.3.15}$$

因此，在给定的置信水平 $1-\alpha$ 条件下，可以用 F 分布来构造两个总体方差之比 σ_1^2/σ_2^2 的置信区间，即

$$F_{\alpha/2} \leq F \leq F_{1-\alpha/2}$$

将样本统计量带入得

$$F_{\alpha/2} \leq \frac{s_1^2}{s_2^2} \bullet \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\alpha/2} \tag{6.3.16}$$

由此可得两个总体方差之比的 $1-\alpha$ 置信区间的下界和上界满足

$$\frac{s_1^2/s_2^2}{F_{1-\alpha/2}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{\alpha/2}} \tag{6.3.17}$$

例 6.12. 利用 R 软件的 wooldridge 包中的数据集 wage1，计算男性和女性工资的方差之比的 95% 的置信区间。

解： 首先对数据集按照性别分组

```
data.f <- subset(wage1[,c('wage','female')],wage1$female==1)
data.m <- subset(wage1[,c('wage','female')],wage1$female==0)
```

然后，分别求出各自样本容量和方差

```
n.f <- length(data.f$wage)
n.m <- length(data.m$wage)
sqs.f <- sd(data.f$wage)^2
sqs.m <- sd(data.m$wage)^2
```

在给定的置信水平下，可以构造两个总体方差之比的置信区间

```
alpha <- 0.05
F.l <- qf(alpha/2,n.m-1,n.f-1)
F.r <- qf(1-alpha/2,n.m-1,m.f-1)
ratio.mf <- sqs.m/sqs.f
```

```
l.F <- round(ratio.mf/F.r,2)
u.F <- round(ratio.mf/F.l,2)
```

因此，两个总体方差之比的 95% 的置信区间为 (1.41　3.45)。

6.4 确定样本容量

进行参数估计时，在样本容量固定的条件下，估计的精确程度和估计的可靠程度之间不能同时兼得。要提高估计的精确程度，就会降低估计的可靠程度；反之亦然。例如，在一次考试之后，若某同学说自己的统计学的考试分数是 86 分，虽然这种说法的精确程度相当高，但它的可靠程度极低；但如果他说自己的统计学考试分数在 30 分到 90 分之间，这种说法的可靠程度虽然极高，但它的精确程度却相当低，以至于没有实际意义。一种比较能够让人接受的说法可能是他的统计学分数在 70 到 80 之间，因为这种说法兼顾了精确程度和可靠程度。所谓的可靠程度，翻译成统计学的术语就是置信水平，而所谓的精确程度，指的是置信区间的宽度。

在实际应用中，若想保持可靠程度不变的同时，提高精确程度，即在保持置信水平不变的条件下，缩小置信区间，唯一的办法就是增加样本容量。但增加样本容量会受到诸多因素的限制，例如调查的时间、工作量的大小和调查费用，等等。因此，时间统计调查中的样本确定和我们愿意容忍的置信区间的宽度以及与此相联系的置信水平有关。因此，确定样本容量，是抽样估计中面临的一个实际问题。

6.4.1 估计总体均值时的样本容量

在对总体均值进行估计时，假定给定的置信水平 $1-\alpha$，则总体均值的置信区间为

$$\bar{x} \mp z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

其中，我们将 $z_{1-\alpha/2}\sigma/\sqrt{n}$ 称为边际误差。因此，调查中的边际误差由两部分因素决定：一是 $z_{1-\alpha/2}$ 值，另一个是样本容量。由于 $z_{1-\alpha/2}$ 大小有 $1-\alpha$ 决定，因此，当总体标准差 σ 和置信水平 $1-\alpha$ 既定时，假定 E 是调查中可以接受的边际误差，由此可得

$$E = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.4.1}$$

由此可以推导出样本容量的确定公式

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{E}\right)^2 \tag{6.4.2}$$

其中，E 表示调查中可以接受的边际误差。$z_{1-\alpha/2}$ 是一个由 $1-\alpha$ 决定的量，σ 指的是总体的标准差，一般情况下，它是一个未知参数，在实际应用中，它可以用样本的标准差来代替。由式 (6.4.2) 可以看出，在其他条件都不变的情况下，样本容量和置信水平呈正比，即置信水平越高，所需的样本容量就越大；样本容量和总体方差成正比，在相同的置信水平和边际误差下，总体的方差越大，所需的样本容量就越大；样本容量和边际误差成反比，若我们可以接受的边际误差越大，所需的样本容量就越小。

最后，需要特别注意的是，这里计算得到的样本容量在绝大部分情况下都不是一个整数，而是一个小数。这种情况下，需要采用 “进一法” 来保证调查的精度。

例 6.13. wooldridge 包中的数据集 wage1，是对不同人群的小时工资情况进行的调查，假定调查的置信水平为 95%，若边际误差是 3 美元，那么满足要求的样本容量至少是多少？

解：在本例中，由于总体标准差未知，需要用样本标准差替代，因此，首先要计算样本标准差。

```
data(wage1)
sd.wage <- sd(wage1$wag)
```

利用样本标准差，在给定的置信水平下，可以计算出必须的样本容量

```
alpha <- 0.05
zval <- qnorm(1-alpha/2)
error <- 1
n <- (zval*sd.wage/error)^2
n <- ceiling(n)
```

计算结果是样本容量 n 不低于 53。这说明，若对总体的平均小时工资估计时，要求的误差不超过 1 美元，其样本容量至少要达到 53 个。

6.4.2 估计总体比例时的样本容量

在进行总体比例的置信区间估计中，边际误差、置信水平和样本方差之间的关系为

$$E = z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}} \tag{6.4.3}$$

从而得到样本容量的计算公式如下

$$n = \frac{z_{1-\alpha/2^2} \bullet \pi(1-\pi)}{E^2} \tag{6.4.4}$$

其中，E 为事先确定的边际误差，载总体比例的区间估计中，一般要求 E 不能大于 0.10。在实际中，由于总体标准差未知，一般可以用样本标准差去代替它。

例 6.14. 在 wooldridge 包的数据集 bwght 中，有关于婴儿出生性别比例的调查，假定置信水平为 95%，调查可以接受的误差为 5%，求满足条件的样本容量是多少？

解：在本例中，同样不知总体的标准差，需要用样本标准差去代替它。利用 R 软件计算标准差的命令如下：

```
data(bwght)
n.m <- sum(bwght$male==1)
p.m <- n.m/length(bwght$male)
```

因此，可以计算满足置信水平和边际误差的样本容量

```
alpha <- 0.05
zval <- qnorm(1-alpha/2)
error <- 0.05
n <- zval^2*p.m*(1-p.m)/error^2
n.p <- ceiling(n)
```

由程序运行的最后结果可以发现，满足上述条件的样本容量为 384。

6.4.3 估计两个总体均值差时的样本容量

在估计两个总体均值之差的置信区间时，样本容量 (n)、置信水平 $(1-\alpha)$ 和预期的边际误差 (E) 之间的关系如下

$$n_1 = n_2 = \frac{z_{1-\alpha/2}^2 \bullet (\sigma_1^2 + \sigma_2^2)}{E^2} \tag{6.4.5}$$

其中，n_1 和 n_2 是来自两个总体的样本容量；σ_1^2 和 σ_2^2 是两个总体的方差。

例 6.15. 假定两个总体的标准差分别为 $\sigma_1 = 12$，$\sigma_2 = 15$, 若要求区间估计的置信水平为 95%，且误差范围不超过 5，假定 $n_1 = n_2$，估计满足上述条件的最小样本容量。

解：利用 R 软件计算的编程如下：

```
alpha <- 0.05
zval <- qnorm(1-alpha/2)
error<- 5
sig.1 <- 12
sig.2 <- 15
n <- zval^2*(sig.1^2+sig.2^2)/error^2
n.diff <- ceiling(n)
```

因此，所需要的样本容量为 57

6.4.4 估计总体比例差时的样本容量

在给定的边际误差和置信水平 $1-\alpha$ 条件下，估计两个总体比例之差所需要的样本容量为

$$n_1 = n_2 = \frac{z_{1-\alpha/2}^2 \bullet (\pi_1(1-\pi_1) + \pi_2(1-\pi_2))}{E^2} \tag{6.4.6}$$

其中 n_1 和 n_2 就是我们要求的样本容量，但在实际应用时，一般都有 $n_1 = n_2$，另外，一般情况下，总体比例 π_1 和 π_2 是未知的，此时可以用比例值 0.5 来代替它。

例 6.16. 假定 $n_1 = n_2$，边际误差为 $E = 0.05$, 求相应置信水平为 95% 时估计两个总体均值之差的区间估计所需的样本容量。

解：由于在本例中，总体的比例无法确定，因此可以假定 $\pi_1 = \pi_2 = 0.5$，利用 R 软件计算样本容量的基本编程为

```
alpha <-0.05
zval <- qnorm(1-alpha/2)
p1 <- 0.5
p2 <- 0.5
error <- 0.05
n <- zval^2*((p1*(1-p1))+p2*(1-p2))/error^2
n.dp <- ceiling(n)
```

计算的结果表明，该调查的样本容量为 769。

第七章　假设检验

假设检验是推断统计学一个重要的分支，它和参数估计类似，但角度不同。参数估计是利用样本信息推断未知的总体参数，假设检验则先对总体参数进行假设，再利用用样本信息判断该假设是否成立。本章首先介绍和假设检验相关的基本理论，然后再介绍假设检验的实际应用，它主要包括对一个总体参数的假设检验和对两个总体参数的假设检验。

7.1　假设检验的基本问题

7.1.1　原假设和备择假设

在统计学中，一般对假设检验的定义如下：

定义 7.1. 假设检验就是先对总体参数提出某种假设，再利用样本信息判断假设是否成立的过程。

由该定义可以看出，假设检验一般分为两步：第一步是对总体参数的具体数值进行陈述，即通常所受的统计假设。统计假设的提出总是以一定的理由为基础，但这些理由通常又是不完全充分的，因此，需要进行第二步，对假设进行检验，也就是利用样本信息判断第一步所提出的假设是否成立。

在假设检验中，首先需要提出两种假设，即原假设和备择假设。

定义 7.2. 原假设又称零假设，指的是研究者搜集证据想否定的假设。

备择假设又称研究假设，指的是研究者搜集证据想支持的假定。

一般情况下，原假设用 H_0 表示，备择假设用 H_1 表示。

在假设检验中，原假设和备择假设的选择和研究者的目的有关，即使是同一事实，不同的研究者对其进行研究，所提出的原假设和备择假设可能恰好相反。下面的几个案例就说明了不同研究者的原假设和备择假设的建立过程。

例 7.1. 假定企业通过机器生产和包装某种食品，按照规定，每袋食品质量是 500 克，为了对生产过程进行控制，产品质量监督人员定期对这台机器生产的产品进行检验，确定产品是否符合质量要求，试写出该质量监督人员检验产品时的原假设和备择假设。

解：假定这台机器生产和包装的食品的平均质量为 μ，则合格食品的质量为 $\mu = 500$ 克，质量过高或过低都是不合格食品。产品质量监督人员判断产品是否合格，就是判断这两种情形是否出现。根据原假设和备择假设的定义，产品质量监督人员想证明的是 “食品质量不合格”。所以，它的原假设和备择假设分别为：

$$\begin{aligned} H_0: &\quad \mu = 500(\text{食品质量合格}) \\ H_1: &\quad \mu \neq 500(\text{食品质量不合格}) \end{aligned}$$

例 7.2. 接例7.1，请写出消费者协会在对食品质量进行假设检验时可能提出的原假设和备择假设。

解：消费者协会从保护消费者的利益出发，因此，在它看来，每袋包装食品的质量低于 500 克的产品应当是不合格产品，因此，它的研究人员在进行假设检验时，原假设和备择假设可能是

$$\begin{aligned} H_0: &\quad \mu = 500(\text{包装食品是合格的}) \\ H_1: &\quad \mu < 500(\text{包装食品品是不合格的}) \end{aligned}$$

例 7.3. 一家研究机构估计，某市人均收入不低于 8 万元的家庭的比重超过 25%，为了验证这一估计是否正确，该研究机构随机抽取了一个样本进行检验。试写出该检验的原假设和备择假设。

解：很明显，研究机构支持的结论是该市人均收入不低于 8 万元的家庭的比重超过 25%。因此，该研究机构的原假设和备择假设是

$$\begin{aligned} H_0: &\quad \pi \leq 30\%(\text{人均收入不低于 8 万元的家庭比例不超过 25\%}) \\ H_1: &\quad \pi > 25\%(\text{人均收入不低于 8 万元的家庭比例超过 25\%}) \end{aligned}$$

由上面的案例可以看出：

1. 原假设和备择假设是互斥的一个完备事件组。也就是说，在进行假设检验时，原假设和备择假设事实上有且只有一个是正确的；
2. 在假设检验中，**等号总是放在原假设上**。例如，在总体均值的假设检验中，我们的原假设可以是 $H_0: \mu = \mu_0$、$H_0: \mu \geq \mu_0$ 或 $H_0: \mu \leq \mu_0$。相应的备择假设应当是 $H_1: \mu \neq \mu_0$、$H_1: \mu < \mu_0$ 或 $H_1: \mu > \mu_0$，等等。将等号写在原假设上，其实和我们的统计量构造有密切的联系。例如，在上面三种情形下，我们都可以以 μ_0 为中心，构造一个服从正态分布或 t 分布的统计量并求出其临界值。若等号不在原假设上，这一步实际上是无法进行的。当然这也给假设检验带来了一定的麻烦，例如，在对类似于“不低于 25% 的家庭的人均收入在 8 万元以上”的说法进行检验时，我们总会碰到怎样设定原假设和备择假设的困难。通常的处理方法是将备择假设写成 $H_1: \pi > 25\%$，理由也很简单，既然可以否定原假设，那说明人均收入在 8 万以上的家庭所占的比例肯定高于 25%。当然支持不低于 25% 的陈述。
3. 在进行假设检验时，通常首先确定的是备择假设，然后确定原假设。这是因为备择假设是研究者想予以支持或证实的，因此十分清楚且容易确定。因此，只要确定了备择假设，原假设就很容易确定。但是，需要注意的是，由于原假设和备择假设的确定取决于研究者的目的或立场，具

有一定的主观性，因此，即使面对同一个事实，不同的研究者也有可能提出截然相反的原假设和备择假设。

在假设检验中，研究者感兴趣的是备择假设的内容，可以是原假设 H_0 某一特定方向的变化，也可以是 H_0 的没有特定方向的变化。在例7.2中，消费者协会所关心的是包装食品的质量是否低于 500 克；在例7.1中，包装食品质量监督者所关心的是包装食品的质量是不是不等于 500 克；而在例7.3中，研究机构所关心的是人均收入 8 万元以上的家庭所占的比例是否高于 25%。

按照备择假设和原假设之间的方向性，可以将其分为双侧检验和单侧检验。

定义 7.3. 单侧检验又称单尾检验，指的是备择假设具有特定的方向性，即含有 “>” 或 “<” 的假设检验。双侧检验又称双尾检验，指的是没有特定的方向性，并含有 $\neq$ 的假设检验。

表 7.1.1: 假设检验的基本形式

假设	双侧检验	单侧检验	
		左侧检验	右侧检验
原假设	$H_0 : \mu = \mu_0$	$H_0 : \mu > \mu_0$	$H_0 : \mu \leq \mu_0$
备择假设	$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

在单侧检验中，按照研究者感兴趣的方向不同，又可分为左侧检验和右侧检验。如果研究者感兴趣的备择假设的方向为 “<”，则是左侧检验；反之，则是右侧检验。例如，上面的例7.2是一个左侧检验，而例7.3则是一个右侧检验。

在表7.1.1中，给出了对单个总体均值检验时，不同的原假设和备择假设的设定形式。

7.1.2 两类错误和显著性水平

假设检验就是根据样本信息，作出是否拒绝原假设的决策。但由于决策是建立在样本信息基础上的，而样本又是随机的，因此，基于样本信息的决策就有犯错误的可能。

如前所述，原假设和备择假设不能同时成立，我们要么拒绝原假设，要么无法拒绝原假设。一种理想的情况是：当原假设是错误的时候，假设检验的结果拒绝原假设；当原假设是正确的时候，假设检验的结果无法拒绝原假设。但现实的情况是，基于假设检验的决策可能会发生以下两类错误：弃真错误和取伪错误。

定义 7.4. 弃真错误又称第 I 类错误，指的是当原假设为真时拒绝原假设所犯的错误。一般情况下，将犯第 I 类错误的概率记作 α。

定义 7.5. 取伪错误又称第 II 类错误，指的是当原假设为假时没有拒绝原假设所犯的错误。一般情况下，将犯第 II 类错误的概率记作 β

在定义了两类错误后，表假设检验中的决策及其后果主要有以下四种情况：

这里需要注意的是：当拒绝原假设时，我们有可能犯第 I 类错误，但绝不会犯第 II 类错误；反之，当无法拒绝原假设时，我们有可能犯第 II 类错误，但绝不会犯第 I 类错误。因此，在一次假设检验中，绝不会同时犯第 I 类错误和第 II 类错误。不过，这两类错误之间的确存在这样的关系：如果想减小犯第 I 类错误的概率，就必然会增加犯第 I 类错误的概率，反之亦然。若想同时减小犯两类错误的概率，唯一的方法就是增加样本容量。在现实中，由于样本容量的增加总是受到一定的限制。所以，只能在两类错误之间权衡，使得它们被控制在可以接受的范围之内。一般来说，发生哪一类错误的后果更为严重，经应当首先控制此类错误发生的概率。在假设检验中，一般先控制犯第 I 类错误的概率。由第 I 类错误发生的概率，可以定义假设检验中的显著性水平。

定义 7.6. 显著性水平指的是假设检验中犯第 I 类错误的概率。

一般情况下，显著性水平用 α 表示。显著性水平水平代表人们事先指定的犯第 I 错误的最大值。一般而言，显著性水平 α 越小，犯第 II 错误的可能性就越大；显著性水平 α 越大，则犯第 II 类错误的可能性就越小。要使 α 和 β 同时减小，唯一的办法就是增加样本容量。在大数据时代，增大样本容量比过去变得更加容易，因此，可以通过增加样本容量的办法，将 α 和 β 都控制在一定的范围之内。虽然从理论上说，首先要控制的总是引发更严重后果的那一类错误，但在实际应用中，由于犯第 I 错误的概率可以由研究者事先控制，因此在假设检验中，人们往往先控制第 I 类错误的发生概率。

此时，一个更为现实的问题是，我们应当将第 I 类错误，即显著性水平控制在多大范围之内才是合适的。一般而言，人们认为犯第 I 类错误的后果更严重，因此倾向于将显著性水平 α 控制在一个较小的概率水平下。著名的英国统计学家 `Ronald Fisher` 在其研究中，将显著性水平的标准定为 5% 并逐渐被广泛接受。现在，显著性水平 $\alpha = 0.05$ 几乎成为一条不成文的规定。当然，在自然科学的假设检验中，有时会取较小的 α 值，如取 $\alpha = 0.01$ 等；在社会科学的假设检验中，有时会取较大的 α 值，如取 $\alpha = 0.1$ 等。另外，在很多社会科学的假设检验中，人们逐渐倾向于不用显著性水平来作出拒绝或无法拒绝原假设的决策，而是给出相应的伴随概率，让读者自己判断是否拒绝原假设。

最后，需要特别说明的是，确定了显著性水平 α，就等于控制了犯第 I 类错误 α 的最大可能性，也就是说，当拒绝原假设 H_0 时，我们犯弃真错误的概率不会超过 α；但若根据现有的样本信息无法拒绝原假设，我们却无法确切知道我们犯第 II 类错误的概率。因此，在进行假设检验中，我们通常采用 “无法拒绝原假设” 的表述方法，而不是采用 “接受原假设” 的表述方法。虽然该说法实质上被没有给出明确的结论，但它成功地避免了我们犯第 II 类错误。正是应为第 II 类错误 β 的控制相对复杂造成了这一说法的盛行。

7.1.3 检验统计量、拒绝域和临界值

1. 检验统计量

在确定原假设和备择假设后，研究者需要为拒绝原假设提供充分的证据。在实际的假设检验过程中，这些证据往往来自抽取的随机样本。利用样本所观测到的信息，可以为假设检验的决策和判断提供依据。例如，若样本所提供的信息能够充分证明原假设不是真实的，我们就有理由拒绝原假设，并选择备择假设。但在一般情况下，样本提供的信息复杂而纷乱，因此，需要对这些信息进行压缩和提炼，检验统计量便是这一过程的结果。

定义 7.7. 检验统计量是依据样本观测结果得到的，对原假设和备择假设作出决策的统计量。

检验统计量实际上是总体参数的点估计量。例如，样本均值 $\bar{x}$ 就是总体均值 μ 的一个点估计量。在假设检验时，我们考虑的是样本统计量和原假设所陈述的参数值之间的差异程度。但这里需要指出的是，这种差异并不是两者之间的绝对差异，而是发生这种差异的概率。因此，有必要对两者间差异标准化。样本统计量标准化依据的条件有两个：一是假定原假设 H_0 为真，二是样本统计量的分布。在此条件下，可以通过样本统计量的抽样分布，计算出点估计量和原假设中设定的参数值之间的差异的概率。事实上，我们所使用的检验统计量都是经过标准化的统计量。例如，在总体均值的假设检验中，我们利用的标准化的统计量为

$$\text{检验统计量} = \frac{\bar{x} - \mu_0}{\delta_x} \tag{7.1.1}$$

其中，$\bar{x}$ 是样本统计量；μ_0 是对总体参数的假定；δ_x 表示的是总体标准差。由式 (7.1.1) 可以看出，所谓的标准化，就是将点估计量和原假设对总体参数的设定值之间的差异转化为总体标准差的倍数，再利用样本统计量的分布计算出发生这种差异的概率。

2. 拒绝域

假设检验的基本原理就是确定一个准则，利用这一准则和计算得到的样本统计量，就可以作出是否拒绝原假设的决策。由于样本统计量是一个随机变量，

会随着抽取的样本不同而发生变化。因此，在假设检验中，我们必须确定的是当样本统计量取哪些值时我们拒绝原假设，取哪些值时我们无法拒绝原假设。由此可以将拒绝域定义为：

定义 7.8. 拒绝域是所有能够拒绝原假设的检验统计量的取值集合。

由定义可以发现，拒绝域就是显著水平 α，而无法拒绝域则是 $1-\alpha$。在假设检验中，若利用样本观测值计算的样本统计量的取值落入拒绝域内，就拒绝原假设，反之，则无法拒绝原假设。下面的图形则标注了各种检验条件下的拒绝域和无法拒绝域。

（这里应当有三幅图）

3. 临界值

拒绝域的大小和我们事先选定的显著性水平 α 有关。在确定了显著性水平 α 后，就可以根据 α 值的大小，确定拒绝域的具体边界值，而这个边界值就是临界值。

定义 7.9. 临界值是根据给定的显著性水平确定的拒绝域的边界值。

在给定显著性水平 α 和检验统计量的分布后，就可以得到该假设检验的临界值。将检验统计量和临界值进行比较，就可以作出是否拒绝原假设的决策。

当样本容量固定时，拒绝域的面积随显著性水平 α 的减小而减小。也就是说，α 值越小，拒绝原假设所需要的临界值就和原假设的参数值越远。拒绝域的位置则取决于检验时双侧检验还是单侧检验。其中，双侧检验的拒绝域位于抽样分布的量产。在单侧检验中，若备择假设带有 $<$ 号，拒绝域就位于抽样分布的左侧，因此，将这类检验称为左侧检验；同样，若备择假设带有 $>$ 号，拒绝域就位于抽样分布的右侧，因此将这类检验称为右侧检验。

至此，我们可以将假设检验的基本步骤进行总结如下：

1. 设定原假设 H_0 和备择假设 H_1。

2. 确定检验的统计量及其分布。

3. 计算样本统计量的大小。

4. 再给定的显著性水平 α 条件下，确定该检验的拒绝域和临界值。

5. 根据临界值和样本统计量的值之间的关系，作出统计决策。

7.1.4 伴随概率和假设检验

利用伴随概率进行假设检验，是最近比较流行的一种新的假设检验的方法。在标准的假设检验中，显著性水平 α 是事先确定的，也就是说，我们事先确定了拒绝域。一旦该区域确定后，无论检验统计量的值是大还是小，只要它落入拒绝域，就可以作出拒绝原假设的决策，反之，则无法拒绝原假设。显然，这种预先确定显著性水平的办法，控制了在假设检验中犯第 I 类错误的的概率。但美中不足的是，这种方法只能提供检验结论可靠性的一个大致范围，却无法给出样本统计量和原假设之间不一致程度的精确度量。在给定的显著性水平下，所有拒绝原假设的统计决策的可靠性都一样。要测度出统计量对原假设的偏离程度，需要计算检验统计量的伴随概率。

定义 7.10. 伴随概率是指在原假设成立的条件下，检验统计量的观测值大于其由样本决定的计算值的概率。

伴随概率，有时又称 P 值，反映的是样本统计量的值和原假设之间的偏离程度的概率。伴随概率越小，说明样本统计量的值和原假设之间的偏离越大，则伴随概率越小，检验的结果也就越显著。

伴随概率是我们进行假设检验时另外一种决策方法，和确定的显著性水平下的假设检验相比，伴随概率法提供的信息更加明确和具体。下面，我们将在不同的假设检验的类型下说明伴随概率的一般表达式和决策的规则。

- 双侧检验

双侧检验的原假设和备择假设分别为：

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

伴随概率指的是当原假设为真时，检验统计量的取值大于等于按照样本观测值计算出来的样本统计量的值的绝对值的概率的两倍。即

$$P = 2P(z > |z_s||\mu = \mu_0)$$

其中，z 表示的是检验统计量的取值；z_s 表示的样本统计量的取值；μ_0 表示对未知参数的设定值。

- 左侧检验

左侧检验的原假设和备择假设分别是：

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

伴随概率指的是当原假设为真时，检验统计量的取值小于样本统计量取值的概率，即

$$P = P(z < z_s|\mu = \mu_0)$$

- 右侧检验

右侧检验的原假设和备择假设分别为

$$H_0 : \mu \leq \mu_0; \quad \mu > \mu_0$$

伴随概率指的是当原假设为真时，检验统计量的取值大于样本统计量的值的概率，即

$$P = P(z > z_s|\mu = \mu_0)$$

利用伴随概率进行假设检验的决策，规则比较简单，就是将伴随概率和给定的显著性水平 α 相比较，若伴随概率小于给定的显著性水平 α，则拒绝原假设，接受备择假设，若伴随概率大于等于给定的显著性水平 α，则无法拒绝原假设。

利用伴随概率进行假设检验的难点在于 P 值的计算，但随着计算机技术的发展和统计软件的应用，使得 P 值的计算变得十分简单。利用伴随概率进行假设检验的好处在于，可以清楚的知道拒绝原假设犯第 I 类错误的概率。而在传统的假设检验中，我们仅仅知道拒绝原假设可能犯的第 I 类错误的最大概率。所以在现在的假设检验中，利用伴随概率进行假设检验几乎已经可以替代传统的假设检验的方法。在本书的假设检验中，我们仍将按照传统的假设检验方法进行假设检验，但同时也给出其利用伴随概率的方法。

7.2 一个总体参数的假设检验

一个总体参数的假设检验主要包括三种情形：

- 总体均值 μ 的假设检验；
- 总体成数 π 的假设检验；
- 总体方差 σ^2 的假设检验。

7.2.1 总体均值的假设检验

在对总体均值进行假设检验时，采用何种检验统计量，主要取决于是大样本抽样还是小样本抽样，方差已知还是方差未知，以及总体是否服从正态分布等前提条件。总结起来，大致有以下两种情形：

1. 大样本条件下的假设检验

根据大数定律和中心极限定理，在大样本条件下，当总体方差已知时，样本均值 $\bar{x}$ 满足

$$\bar{x} \sim N(\mu, \sigma^2/n)$$

将其标准化后可得 z 统计量

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{7.2.1}$$

例 7.4. 假定数据集 mtcars 是一个随机抽取的样本，根据数据集中的汽车油耗 (mpg) 的信息，检验汽车油耗 (mpg) 和 25 有没有显著性差异。(假定汽车油耗的总体标准差 $\sigma = 8$，显著性水平 $\alpha = 0.05$)。

解：由数据集 mtcars 的观测值的个数可知，这是一个大样本抽样。根据假设检验的内容可知，这是一个双侧检验。其原假设和备择假设分别为

$$H_0 : \mu = 25; \quad H_1 : \mu \neq 25$$

由于总体标准差已知，因此用 z 统计量进行假设检验

```
data(mtcars)
m.mpg <- mean(mtcars$mpg)
n <- length(mtcars$mpg)
sigma <- 8
alpha <- 0.05
mu0 <- 25
cval <- round(qnorm(0.025),2)
zval <- round((m.mpg-mu0)/(sigma/sqrt(n)),2)
```

由此可得该检验的样本统计量为-3.47，在给定的显著性水平 $\alpha = 0.05$ 时，其临界值为-1.96，由此可见，样本统计量的值落入拒绝域区间，因此应当拒绝原假设。若采用伴随概率的方法，可以求得

```
pval <- 2*round(pnorm(zval),5)
```

即其伴随概率为 5.2×10^{-4}，要远远小于给定的显著性水平，同样也应当拒绝原假设。根据伴随概率可知，否定原假设所犯的错误大约是万分之六，这是一个相当小的弃真错误。

在大样本条件下，当总体方差未知时，需要用样本方差取替代总体方差，此时，统计量

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \tag{7.2.2}$$

例 7.5. 接例7.4，假定总体方差未知，在显著性水平 $\alpha = 0.05$ 时，检验平均油耗和 25 之间有无显著性差异。

解：这是一个大样本条件下，总体方差未知的假设检验，因此，需要用 t 统计量进行假设检验。

```
sd.mpg <- sd(mtcars$mpg)
tval <- round((m.mpg-25)/(sd.mpg/sqrt(n)),3)
ctval <- round(qt(alpha/2,n-1),3)
pt <- round(2*pt(tval,n-1),5)
```

在显著性水平 $\alpha = 0.05$ 时，该检验的临界值为-2.04，样本统计量的值为-4.608，因此，应当拒绝原假设。若是利用伴随概率进行假设检验，得到的 P 值为 7×10^{-5}，在显著性水平为 0。05 时，同样应当拒绝原假设。

当然，在本例中，一个更简单的方法时利用 R 软件自带的 *t.test()* 命令进行假设检验，其结果如下：

```
result <- t.test(mtcars$mpg,alternative = 'two.sided',mu=25)
result
```

```
##
##  One Sample t-test
##
## data:  mtcars$mpg
## t = -5, df = 31, p-value = 7e-05
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  17.9 22.3
## sample estimates:
## mean of x
##      20.1
```

利用该命令得到的结果和前面的结果一致。关于 *t.test()* 的详细用法，可以用

```
help('t.test()')
```

来查看。

2. 小样本条件下的假设检验

在小样本情形下，我们总是假定总体服从正态分布，此时检验统计量的选择就和总体方差是否已知有关。若总体方差已知，其检验所用的统计量同式 (7.2.1)；若总体方差未知，则其检验所用的统计量同式 (7.2.2)。

例 7.6. 利用 R 软件中的数据集 women，在显著性水平 $\alpha = 0.05$ 的条件下，判断女性的身高和 6 英寸有无显著性的差异。

解：这是一个典型的小样本检验，其原假设和备择假设分别为：

$$H_0 : \mu = 60; \quad H_1 : \mu \neq 60$$

当总体方差未知时，用样本方差替代总体方差，可得满足下列条件的 t 统计量：

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$$

其检验过程如下：

```
data(women)
n <- length(women$height)
m.height <- mean(women$height)
sd.height <- sd(women$height)
mu0 <- 60
alpha <-0.05
tval <- round((m.height-mu0)/(sd.height/sqrt(n)),3)
ctval <- round(qt(1-alpha/2,n-1),3)
```

当显著性水平 $\alpha = 0.05$ 时，其临界值为 2.145，根据样本得到的 t 统计量的值为 4.33。由于该值位于检验的拒绝域，因此，应当拒绝原假设。

同样，利用伴随概率进行假设检验时，首先需要计算该检验的伴随概率

```
pt <- 2*(1-pt(tval,n-1))
```

其结果为 6.9×10^{-4}，要远小于给定的显著性水平 α，同样可以得到拒绝原假设的结论。

利用 *t.test()* 命令，得到的结果如下：

```
result <- t.test(women$height,mu=60,alternative = 'two.sided')
result
```

```
##
##  One Sample t-test
##
## data:  women$height
## t = 4, df = 14, p-value = 7e-04
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  62.5 67.5
## sample estimates:
## mean of x
##        65
```

其结果和前面的计算结果相同。

7.2.2 总体比例的假设检验

在总体比例的假设检验中，我们仅介绍大样本情形下的总体比例的假设检验。通常用 π_0 表示对总体比例的某一假设值，用 p 表示样本比例。总体比例检验的三种基本形式为：

- 双侧检验：$H_0 : \pi = \pi_0; \quad H_1 : \pi \neq \pi_0$
- 左侧检验：$H_0 : \pi \geq \pi_0; \quad H_1 : \pi < \pi_0$
- 右侧检验：$H_0 : \pi \leq \pi_0; \quad H_1 : \pi > \pi_0$

由大数定律和中心极限定理可得：

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0,1) \tag{7.2.3}$$

由此可见，该统计量实际上仍然将样本比例和总体比例之间的差异进行标准化。

但在实际应用时，总体方差时未知的，因此需要用样本方差取替代它，由此，得到

$$t = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}} \sim t(n-1) \tag{7.2.4}$$

当然，当样本容量足够大时，式 (7.2.4) 中的 t 统计量收敛于式 (7.2.3) 表示的 z 统计量。

例 7.7. 某研究机构认为，出生的婴儿中男孩的比例要高于 50%，wooldridge 包中的 bwght 就是他们随机抽取的一个样本，利用该数据集中的变量 male 的信息，在显著性水平 $\alpha = 0.05$ 的条件下对这一说法进行假设检验。

解：该研究机构想证明的是出生婴儿中男孩的比例超过 50%，所以，该检验的原假设和备择假设分别为：

$$H_0 : \pi = 0.5; \quad H_1 : \pi > 0.5$$

该假设检验的 R 软件编程如下

```
library(wooldridge)
data(bwght)
```

首先要计算样本的比例 p 和样本的标准差

```
n <- length(bwght$male)
num.male <- sum(bwght$male==1)
p.m <- num.male/n
sd.m <- sqrt(p.m*(1-p.m))
```

由此可得，样本比例为 52.09%。由于总体方差未知，因此，需要用样本方差取替代它。在本例中，样本容量为 1388，故而可以用 z 统计量来进行检验。

```
pi0=0.5
zval <- (p.m-pi0)/(sd.m/sqrt(n))
alpha <- 0.05
cval <- qnorm(1-alpha)
```

由此可得样本统计量的取值为 1.558，在显著性水平 $\alpha = 0.05$ 条件下，其临界值为 1.645，因此，样本统计量的取值落入无法拒绝域，因此，无法拒绝原假设。

若利用伴随概率进行假设检验，可以计算出样本统计量取值的 P 值

```
p <- 1-pnorm(zval)
```

可得该检验的伴随概率为 5.96%，要大于给定的显著性水平 $\alpha = 0.05$，同样无法拒绝原假设。

7.2.3 总体方差的假设检验

和总体均值与总体比例的假设检验使用的检验统计量不同，对总体方差的假设检验所用的检验统计量时 χ^2 统计量。另外，在总体方差的假设检验中，无论样本容量 n 是大还是小，都要求总体服从正态分布。

总体方差检验的统计量是

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{7.2.5}$$

用 σ_0^2 表示对总体方差的某种假定，总体方差假设检验的基本形式有三种：

- 双侧检验 $H_0 : \sigma^2 = \sigma_0^2; \quad H_1 : \sigma^2 \neq \sigma_0^2$
- 左侧检验 $H_0 : \sigma^2 \geq \sigma_0^2; \quad H_1 : \sigma^2 < \sigma_0^2$
- 右侧检验 $H_0 : \sigma^2 \leq \sigma_0^2; \quad H_1 : \sigma^2 > \sigma_0^2$

例 7.8. 利用 wooldridge 包中一个名为 bwght 的数据集，在显著性水平 $\alpha = 0.05$ 的条件下，检验 “总体方差是 625” 的陈述。

解：该检验的原假设和备择假设分别为：

$$H_0 : \sigma^2 = 625; \quad \sigma^2 \neq 625$$

由此可见，这是一个典型的双侧检验。

该检验的 R 软件编程如下：

```
library('wooldridge')
data(bwght)
n <- length(bwght$faminc)
```

```
sd.fmc <- sd(bwght$faminc)
sgm0 <- 25
chisq <- (n-1)*sd.fmc^2/sgm0^2
```

得到该检验的卡方统计量为 779.296。当显著性水平 $\alpha = 0.05$ 时，可以得到该检验的临界值

```
alpha <- 0.05
lcrt.chsq <- qchisq(alpha/2,n-1)
ucrt.chsq <- qchisq(1-alpha/2,n-1)
```

根据计算结果可知，样本统计量落入拒绝域，因此拒绝原假设。

如果用伴随概率进行决策，则需要计算相应的 P 值

```
pval <- pchisq(chisq,n-1)
```

从而得到 P 值为 0，要远小于显著性水平 α，因此，拒绝原假设。

7.3 两个总体参数的假设检验

和两个总体参数的区间估计相似，两个总体参数的假设检验包括以下三种情形：

- 两个总体均值之差的假设检验；
- 两个总体比例之差的假设检验；
- 两个总体方差之比的假设检验。

7.3.1 两个总体均值之差的假设检验

在实际中，我们经常需要比较两个总体的差异，例如，男性和女性的工资水平是否有显著的差异；新的教学法是否显著提高学生的考试成绩；新的流水线是否显著降低了次品的数量，等等。这些问题都可以归结为两个总体均值之差 $(\mu_1 - \mu_2)$ 的检验问题。

同样，关于两个总体均值之差的假设检验的形式，大值有三种：

- 双侧检验　$H_0 : \mu_1 - \mu_2 = 0;\quad H_1 : \mu_1 - \mu_2 \neq 0$
- 左侧检验　$H_0 : \mu_1 - \mu_2 \geq 0;\quad H_1 : \mu_1 - \mu_2 < 0$
- 右侧检验　$H_0 : \mu_1 - \mu_2 \leq 0;\quad H_1 : \mu_1 - \mu_2 > 0$

两个总体均值的检验从总体上要考虑以下几个方面的因素：

- 样本匹配形式：是独立的还是匹配的；
- 样本抽样形式：大样本抽样还是小样本抽样；
- 总体方差问题：总体方差是否给定。

1. 独立样本

在独立样本条件下，另个总体均值之差检验的基础是以两个样本的均值之差 $\bar{x}_1 - \bar{x}_2$ 为基础构造检验统计量。对于大样本和小样本两种情形，虽然所构造的统计量略有不同，但其基本思想是一致的。例如，在两个总体参数之差的双侧检验中，其原假设和备择假设分别为：

$$H_0 : \mu_1 - \mu_2 = 0;\quad H_1 : \mu_1 - \mu 2 \neq 0$$

在大样本条件下，两个样本均值之差 $\bar{x}_1 - \bar{x}_2$ 的抽样分布满足

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

因此，在总体方差 σ_1^2 和 σ_2^2 已知的情形下，将 $\bar{x}_1 - \bar{x}_2$ 标准化后可以得到

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{7.3.1}$$

当总体方差 σ_1^2 和 σ_2^2 未知时，需要用样本的方差 s_1^2 和 s_2^2 去替代总体方差，由此得到统计量

$$r = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(n_1 + n_2 - 2) \tag{7.3.2}$$

当样本容量足够大时，可以用式 (7.3.1) 中的 z 统计量替代式 (7.3.2) 中的 t 统计量。

例 7.9. 在 wooldridge 包中的数据集 wage1 中，给出了男性和女性的小时工资的抽样数据，请根据该数据集，在显著性水平 $\alpha = 0.05$ 时，能否认为男性的平均小时工资显著高于女性的平均小时工资。

解：这是一个典型的右侧检验，因此，该检验的原假设和备择假设为

$$H_0 : \mu_1 - \mu_2 = 0; \quad H_1 : \mu_1 - \mu_2 > 0$$

调用数据集 wage1，并按照性别对其分组求平均小时工资及其标准差

```
library(wooldridge)
data(wage1)
n <- length(wage1$wage)
sub.m <- subset(wage1[,c('wage','female')],wage1$female==0)
```

```
sub.f <- subset(wage1[,c('wage','female')],wage1$female==1)
m.wagem <- mean(sub.m$wage)
sd.wagem <- sd(sub.m$wage)
m.wagef <- mean(sub.f$wage)
sd.wagef <- sd(sub.f$wage)
```

由样本数据可知，男性的平均小时工资为 7.099，标准差为 4.161；女性的平均小时工资为 4.588，标准差为 2.529。由于样本容量为 526，已经足够大，因此，可以用 z 统计量来替代 t 统计量。

```
n.m <- length(sub.m$wage)
n.f <- length(sub.f$wage)
ds <- m.wagem-m.wagef
sd.std <- sqrt(sd.wagem^2/n.m+sd.wagef^2/n.f)
zval <- ds/sd.std
alpha <- 0.05
crtv <- qnorm(1-alpha)
```

该检验的样本统计量为 8.44。当显著性水平 $\alpha = 0.05$ 时，该检验的临界值为 1.645，因此，应当拒绝原假设。即男性和女性的小时工资水平由显著性差异。

本例中，若采用伴随概率进行假设检验的决策，可以计算出样本统计量的伴随概率

```
pval <- 1-pnorm(zval)
```

由此可得该检验的伴随概率为 0，要远小于给定的显著性水平，同样应当拒绝原假设。

当然，在本例中，我们也可以直接用 *t.test()* 命令来进行假设检验，其结果如下

```
result<- t.test(sub.m$wage,sub.f$wage,paired=F,
                var.equal=F,mu=0,alternative = 'greater')
```

```
##   Statistic Degree of Freedom Probability
##        8.44               456    2.12e-16
```

在小样本抽样条件下，对两个总体均值之差进行假设检验时，首先需要假定两个总体都服从正态分布。一般而言，主要有以下四种情形：

- 总体方差 σ_1^2 和 σ_2^2 已知

此时无论样本容量时多大，两个样本均值之差都服从正态分布。可以用式 (7.3.1) 中的 z 统计量来进行假设检验。

- 总体方差 σ_1^2 和 σ_2^2 未知但相等

在两个总体方差 σ_1^2 和 σ_2^2 未知但相等时，需要用样本方差 s_1^2 和 s_2^2 来估计总体方差。由于假定 $\sigma_1^2 = \sigma_2^2$，可以将两组样本数据混合起来估计总体方差，即

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \tag{7.3.3}$$

将两个样本均值之差标准化后，得到

$$t < -\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2) \tag{7.3.4}$$

可用该统计量来进行两个总体均值之差的假设检验。

例 7.10. 利用 R 软件的 mtcars 数据集，在显著性水平 $\alpha = 0.05$ 时判断 4 缸发动机和 8 缸发动机的油耗有无明显的差别。(假定总体服从正态分布，两种发动机的油耗的方差相同且未知)

解：假定 4 缸发动机的平均油耗时 μ_1，8 缸发动机的油耗是 μ_2，这里需要注意的是油耗的单位是英里/加仑。因此，该检验的原假设和备择假设分别为

$$H_0 : \mu_1 - \mu_2 = 0; \quad H_1 : \mu_1 - \mu_2 > 0$$

首先调用数据集并按照发动机气缸数分组，分别计算各自的油耗均值和标准差。

```
data(mtcars)
fcyld <- subset(mtcars[,c('mpg','cyl')],mtcars$cyl==4)
ecyld <- subset(mtcars[,c('mpg','cyl')],mtcars$cyl==8)
m.fmpg <- mean(fcyld$mpg)
sd.fmpg <- sd(fcyld$mpg)
m.empg <- mean(ecyld$mpg)
sd.empg <- sd(ecyld$mpg)
```

由于假定总体服从正态分布，方差未知但相等，因此，首先需要计算样本方差

```
n.f <- length(fcyld$mpg)
n.e <- length(ecyld$mpg)
sq.sp <- ((n.f-1)*sd.fmpg^2+(n.e-1)*sd.empg^2)/(n.f+n.e-2)
```

因此，可计算 t 统计量的大小

```
ds <- m.fmpg-m.empg
tval <- ds/sqrt(sq.sp*(1/n.f+1/n.e))
```

由此可得样本统计量的取值为 8.102。

在给定的显著性水平下，求该检验的临界值

```
alpha <- 0.05
crt <- qt(1-alpha,n.f+n.e-2)
```

因此，可到到该检验的临界值为 1.714。通过比较二者之间的大小，可以作出拒绝原假设的决策，即 4 缸汽车比 8 缸汽车更省油。

当然，本例也可以用 R 软件自带的 *t.test()* 命令来进行假设检验，留给读者自行研究该命令的使用方法，这里不再赘述。

- 总体方差未知且不相等，但样本容量相等

若总体方差 σ_1^2 和 σ_2^2 未知，且 $\sigma_1^2 \neq \sigma_2^2$，但两个独立样本的样本容量相同，即 $n_1 = n_2 = n$，此时，将两个样本均值之差进行标准化后得到统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2 + s_2^2}{n}}} \sim t(2n-2) \qquad (7.3.5)$$

例 7.11. 利用例 (7.9) 中的两个独立样本数据，从数据集 sub.m 和 sub.f 中各抽取 20 个观测值组成新的数据集，利用该数据集来检验男性和女性的工资水平有无显著性差异。(假定总体服从正态分布，方差不等，显著性水平为 $\alpha = 0.05$)

解： 首先从两个数据集中抽取 20 个观测值，组成新的数据集

```
set.seed(123)
wagem <- sub.m[sample(1:nrow(sub.m),size=20,replace=T),]
wagef <- sub.f[sample(1:nrow(sub.f),size=20,replace=T),]
```

利用该数据集检验男性和女性的平均工资水平是否有显著差异时，其原假设和备择假设分别为

$$H_0 : \mu_1 - \mu_2 = 0; \quad H_1 : \mu_1 - \mu_2 \neq 0$$

其检验程序如下

```
n <- nrow(wagem)
m.wm <- mean(wagem$wage)
sd.wm <- sd(wagem$wage)
m.wf <- mean(wagef$wage)
sd.wf <- sd(wagef$wage)
dw <- m.wm-m.wf
t <- dw/sqrt((sd.wm^2+sd.wf^2)/(2*n))
alpha <- 0.05
crt <- qt(1-alpha,2*n-2)
```

由此可得，该检验的临界值为 1.686，样本统计量为 4.217，因此，应当拒绝原假设，即两者工资水平由显著的差异。

- 总体方差未知且不相等

当总体服从正态分布，两个总体的方差不相等，且样本容量也不相等，即 $\sigma_1^2 \neq \sigma_2^2$ 且 $n_1 \neq n_2$ 时，两个样本均值之差经标准化后不再服从自由度为 n_1+n_2-2 的 t 分布。此时检验统计量满足

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(v) \tag{7.3.6}$$

其中，自由度 v 满足

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \tag{7.3.7}$$

若利用上式计算得到的不是整数，则采用进一法取整。

例 7.12. 在例 (7.10) 中，假定总体的方差未知且不相等，在显著性水平 $\alpha = 0.05$ 条件下，判断两种发动机的油耗是否显著的不等。

解：由于总体的方差未知且不相等，样本容量也不相同，在总体服从正态分布的条件下，首先计算自由度

```
df1 <- (sd.fmpg^2/n.f+sd.empg^2/n.e)^2
df2 <- (sd.fmpg^2/n.f)^2/(n.f-1)+(sd.empg^2/n.e)^2/(n.e-1)
df <- ceiling(df1/df2)
```

再计算检验的统计量和检验的临界值

```
ds <- m.fmpg-m.empg
sq.sd <- sqrt(df1)
tval <- ds/sqrt(sq.sd)
alpha <- 0.05
crt <- qt(1-alpha,df)
```

由此可得该检验的临界值为 1.753，统计量为 7.597，因此拒绝原假设。

2. 配对样本

为了避免因使用独立样本时因个体差异所带来的信息干扰，一个有效的解决办法时使用配对样本。在介绍使用配对样本进行假设检验之前，首先要对下面的运算有一个大致的了解。

$d_i \quad (i = 1, 2, \cdots, n)$ 表示的是第 i 个配对样本数据的差值。

$\bar{d}$ 表示的是配对样本数据差值的均值，即

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

s_d^2 表示配对样本数据差值的方差，即

$$s_d^2 = \frac{\sum_{i=1}^{2} (d_i - \bar{d})^2}{n-1}$$

在样本容量足够的的情形下，配对样本的差值经标准化后服从正态分布，即

$$z = \frac{\bar{d} - (\mu_1 - \mu_2)}{s_d/\sqrt{n}} \tag{7.3.8}$$

在大样本条件下，由于是用样本的方差去替代两个总体的方差，因此可得

$$z = \frac{\bar{d} - (\mu_1 - \mu_2)}{s_d/\sqrt{n}} \sim t(n-1) \tag{7.3.9}$$

同理，若两个总体服从正态分布，且配对样本的配对差所构成的总体的方差是未知的，此时可以得到式 (7.3.9) 所示的 t 统计量。

利用配对样本进行假设检验，同样可以分为双侧检验、左侧检验和右侧检验三种情形，这里也不再进行详细的叙述。

例 7.13. 某饮料公司开发出一种新的饮品，为了比较消费者对新老两种饮品的偏爱程度，该公司随机抽选一组消费者 (10 人)，每个消费者先后随机的品尝两种饮品，并对其进行评分，评分结果如表 (7.3.1) 所示。在显著性水平 $\alpha = 0.05$ 的条件下，判断消费者对两种饮品的评分是否存在显著差异。

表 7.3.1: 两种饮品的消费者评分的配对样本数据

消费者编号	1	2	3	4	5	6	7	8	9	10
旧饮品评分	5	4	7	3	5	6	5	8	6	7
新饮品评分	6	6	7	7	4	5	6	9	7	6

解： 该检验的原假设和备择假设分别为

$$H_0 : \mu_1 - \mu_2 = 0; \quad \mu_1 - \mu_2 \neq 0$$

该假设检验的程序如下

```
drink.o <- c(5,4,7,3,5,6,5,8,6,7)
drink.n <- c(6,6,7,7,4,5,6,9,7,6)
ds <- drink.o-drink.n
m.d <- mean(ds)
sd.d <- sd(ds)
n <- length(drink.o)
tval <- m.d/(sd.d/sqrt(n))
alpha <- 0.05
crt<- qt(1-alpha/2,n-1)
```

由此可得该检验的临界值为 2.262，样本统计量为-1.413。因此无法拒绝原假设。即消费者对两种饮品的评分没有显著的差异。

7.3.2 两个总体比例之差的假设检验

我们只考虑样本容量足够大的情形下，两个总体比例之差的假设检验问题。所谓样本容量足够大，指的是 n_1p_1、$n_1(1-p_1)$、n_2p_2 和 $n_2(1-p_2)$ 都大于等于 5。其中，n_1、p_1 和 n_2、p_2 分别表示两个样本的样本容量和样本比例。

由两个样本比例之差的抽样分布，可以得到检验两个总体比例之差的检验统计量

$$z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sigma_{p_1-p_2}} \tag{7.3.10}$$

其中，$\sigma_{p_1-p_2}$ 表示的是两个样本比例之差的抽样分布的标准差。当总体比例已知时，有

$$\sigma_{p_1-p_2}=\sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$$

但在实际应用中，两个总体的比例 π_1 和 π_2 是未知的，需要用样本比例 p_1 和 p_2 取估计它。不过怎么估计总体标准差和原假设有关。

一是在原假设成立，即 $\pi_1-\pi_2=0$ 的情形下，总体标准差的最佳估计量时将两个样本混合起来，得到合并后的样本比例 p 满足

$$p=\frac{n_1p_1+n_2p_2}{n_1+n_2} \tag{7.3.11}$$

再利用混合后的 p 去估计总体标准差，可得

$$\sigma_{p_1-p_2}=\sqrt{\frac{p(1-p)}{n_1}+\frac{p(1-p)}{n_2}}=\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{7.3.12}$$

将式 (7.3.12) 代入式 (7.3.10) 可得 z 统计量

$$z=\frac{p_1-p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \tag{7.3.13}$$

当我们要检验的原假设为 $H_0:\pi_1-\pi_2=d_0(d_0\neq 0)$ 时，可直接用样本比例 p_1 和 p_2 作为总体比例 π_1 和 π_2 的估计量，从而得到两个样本比例之差的抽样分布的标准差 $\sigma_{p_1-p_2}$ 的估计量为

$$\hat{\sigma}_{p_1-p_2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \tag{7.3.14}$$

例 7.14. 吸烟会不会影响婴儿的性别，这是一个有趣的话题。利用 wooldridge 包中一个名为 bwght 的数据集，分别计算吸烟孕妇和不吸烟孕妇生产的婴儿的男性所占的比重，并在显著性水平为 $\alpha=0.05$ 的条件下判断两组婴儿的性别有无显著的差异。

解： 首先读入数据集 bwght，并分组计算出生婴儿中男性所占的比重。

```
library(wooldridge)
data(bwght)
data.cig <- subset(bwght[,c('male','cigs')],bwght$cigs>0)
data.nocig <- subset(bwght[,c("male",'cigs')],bwght$cigs==0)
n.c <- nrow(data.cig)
n.nc <- nrow(data.nocig)
p1.m <- sum(data.cig$male==1)/n.c
p2.m <- sum(data.nocig$male==1)/n.nc
```

该假设检验的原假设和备择假设分别为

$$H_0:\pi_1-\pi_2=0;\quad \pi_1-\pi_2\neq 0$$

在原假设为真的情形下，估计 $\sigma_{p_1-p_2}$ 的标准差

```
p <- sum(bwght$ma==1)/(n.c+n.nc)
```

由此可得该检验的统计量

```
dp <- p1.m-p2.m
sigma <- sqrt(p*(1-p)*(1/n.c+1/n.nc))
zval <- dp/sigma
```

当显著性水平为 $\alpha = 0.05$ 时，其临界值为

```
alpha <- 0.05
crz <- qnorm(alpha/2)
```

由此可得，该检验的临界值为-1.96，统计量的取值为-1.558，因此，无法拒绝原假设。即孕妇抽烟对出生婴儿的性别时没有影响的。

7.3.3 两个总体方差之比的假设检验

在对两个总体的方差进行比较时，我们总是假定总体服从正态分布。因为，当总体服从正态分布时，从两个正态总体中分别独立地抽取两个样本时，方差之比 σ_1^2/σ_2^2 的估计量的抽样分布服从 F 分布。因此，在实践应用中，当对两个总体的方差进行比较时，最合适的形式是两个总体方差之比，即 σ_1^2/σ_2^2 或 σ_2^2/σ_1^2，且经常将原假设和备择假设的基本形式设定为两个总体方差与 1 之间的比较关系。

在对两个总体方差之比进行假设检验时，按照检验的类型不同，原假设和备择假设有三种不同的形式。

- 双侧检验

在对两个总体方差之比进行双侧检验时，其原假设和备择假设分别为

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1; \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

- 左侧检验

左侧检验的原假设和备择假设分别为

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1; \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$$

- 右侧检验

右侧检验的原假设和备择假设分别为

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1; \quad H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

由于两个样本的方差之比 s_1^2/s_2^2 是两个总体方差之比 σ_1^2/σ_2^2 的理想估计量。当从两个正态总体中随机抽取样本容量分别为 n_1 和 n_2 的两个独立样本时，统计量

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1) \tag{7.3.15}$$

可以作为两个总体方差之比的检验统计量。在原假设成立的条件下，可以将式 (7.3.15) 改写成

$$F = \frac{s_1^2}{s_2^2} \quad \left(\text{或 } F = \frac{s_2^2}{s_1^2}\right) \tag{7.3.16}$$

这里需要注意的是，在进行方差之比的双侧检验时，我们总是用较大的方差除以较小的方差，这样做的目的是让拒绝域总是在抽样分布的右侧。而在单侧检验中，我们也可以将任何一个单侧检验问题转换称为右侧检验。因此，任何两个总体方差之比的单侧检验，都可以具有相同的拒绝域。当然，我们也可以用较小的样本方差去除以较大的样本方差，只不过，此时的拒绝域在左侧。在利用 R 软件进行假设检验时，较大的方差在分子，与其说是为了简便，不如说是一种习惯。

例 7.15. 利用数据集 bwght 中的信息，取显著性水平 $\alpha = 0.05$，判断出生的男婴和女婴的体重的方差是否显著不同。

解： 首先对数据集 bwght 进行分组计算男婴和女婴的体重的方差

```
bw.m <- subset(bwght[,c('bwght','male')],male==1)
bw.f <- subset(bwght[,c('bwght','male')],male==0)
sqsd.m <- sd(bw.m$bwght)^2
sqsd.f <- sd(bw.f$bwght)^2
```

该检验的原假设和备择假设分别为

$$H_0 : \sigma_f^2/\sigma_m^2 = 1; \quad H_1 : \sigma_f^2/\sigma_m^2 \neq 1$$

现计算样本统计量和检验的临界值

```
Fval <- sqsd.f/sqsd.m
n.m <- nrow(bw.m)
n.f <- nrow(bw.f)
alpha <- 0.05
fcrt <- qf(1-alpha/2,n.f-1,n.m-1)
```

由计算结果可以发现，当显著性水平 $\alpha = 0.05$ 时，该检验的临界值为 1.161，样本统计量为 1.004，因为样本统计量落入无法拒绝域，因此，无法拒绝原假设。

例 7.16. 利用数据集 wage1 中的信息，取显著性水平 $\alpha = 0.05$，判断男性的工资水平的房和女性工资水平的方差是否显著不同。

解： 首先对数据按照性别分组，分别计算两组的方差

```
data(wage1)
wage.m <- subset(wage1[,c('wage','female')],female==0)
wage.f <- subset(wage1[,c('wage','female')],female==1)
sqsd.m <- sd(wage.m$wage)^2
sqsd.f <- sd(wage.f$wage)^2
```

该检验的原假设和备择假设分别为

$$H_0 : \sigma_m^2/\sigma_f^2 = 1; \quad \sigma_m^2/\sigma_f^2 \neq 1$$

在原假设为真时，可以计算样本统计量和该检验的临界值。

```
Fval <- sqsd.m/sqsd.f
n.m <- nrow(wage.m)
n.f <- nrow(wage.f)
alpha <- 0.05
Fcrt <- qf(1-alpha/2,n.m-1,n.f-1)
```

因为该检验的临界值为 1.276，样本统计量为 2.706，样本统计量落入拒绝域，因此应拒绝原假设。即男性工资水平的方差和女性工资水平方差具有显著性的差异。

练习题

第八章　回归分析

8.1　相关分析

8.1.1　变量之间的关系

在经济学科中，经常需要对经济变量之间的关系进行分析。例如，在企业生产中，最重要的是进行投入-产出分析，由此产生了柯布-道格拉斯生产函数；在现实生活照，最重要的是进行收入-消费之间的关系分析，凯恩斯的消费函数应运而生；随着大数据时代的来临，相关分析更是大行其道，几乎充斥着经济学的每一个领域。在实践中人们发现，变量之间的关系大致有三种情况：

- 没有关系

很多变量之间都没有关系，例如，广告费的支出和股市的指数，几乎是八竿子也打不着的事情。其实，现实中绝大部分变量之间的确没有关系，在经济领域如此，在其他领域也概莫能外。

- 函数关系

两个变量之间关系的第二种情形时函数关系。函数关系是两个变量之间关系的一种特殊形态。假定有两个变量 x 和 y，若变量 y 随着变量 x 的变化而变化，且完全依赖于 x，也就是说当 x 取某个数值时，y 取相应的值，则称 y

时 x 的函数，记作 $y = f(x)$，其中 x 称为自变量，y 称为因变量。下面给出几个常见的函数形式。

1. 圆的半径和圆的面积之间的关系。圆的面积 S 和圆的半径 r 之间的关系的形式为

$$S = \pi R^2$$

2. 某种产品的销售额和销售量之间的关系。在完全竞争市场上，假定产品的销售数量为 Q，销售价格是一个外生的变量 P，由此可得销售额 TR 和销售量之间的关系为

$$TR = PQ$$

变量之间的函数关系其实是两个变量之间关系的一种特殊情形，这种特殊性主要表现为变量之间的一一对应关系。但在实际问题中，变量之间的关系往往表现得更复杂。例如，人的体重和身高之间的关系。首先，我们知道，人的体重和身高之间有关系，即随着人的身高的增加，一般而言其体重也会增加，说明人的体重和身高之间有关系。但人的体重和身高之间的关系并不是一个一一对应的关系。也就是说，即使两个人的身高相同，但也不见得他们的体重相等。出现这种现象的根本原因在于，除了人的身高外，影响人的体重的因素有很多，它们共同造成了变量之间关系的不确定性。

定义 8.1. 相关关系 (correlation) 是变量之间存在的不确定的数量关系。

在实际中，有很多存在相关关系的例子。例如，子女的身高 y 和父母的身高 x 有较为密切的关系。一般而言，父母的身高越高，其子女的身高通常也比较高，父母身高较低时，其子女的身高通常也比较低。但实际情况并不完全是这样，因为它们之间并不是完全确定的关系，而是一种典型的相关关系。

另一个存在相关关系的经典案例是一个人的收入水平 y 和他的受教育程度 x 之间的关系。一般而言，一个人的受教育程度越高，其收入水平也越高，但收

入水平和受教育程度之间并非函数关系，两个受教育程度一样的人，其工资水平可能不同，因此，可以断定受收入水平和受教育程度之间存在相关关系。

从上面几个例子可以看出，相关关系有自己的特点：

- 相关关系不是函数关系。因为在函数关系中，两个变量的取值存在一一对应关系，这一关系在相关关系中几乎不存在。
- 相关关系要通过大量的观测才能被发现。由于相关关系中存在不确定性，因此，它无法用函数关系进行描述。虽然相关关系也可以揭示一定的规律性，但这种规律性必须通过大量的观测才能发现。

8.1.2 相关关系的测度

相关关系的测度包括两个方面：一是有无相关关系，二是相关关系的密切程度。利用散点图可以判断两个变量之间有无相关关系；利用相关系数可以判断相关的密切程度。

1. 散点图

散点图时判断两个变量之间有无相关关系最直观的方法。其具体的做法如下：

- 以横坐标表示自变量 x，纵坐标 y 表示因变量。
- 将每一个 x 和与之相对应的 y 转化为坐标轴中的散点。

由坐标轴和散点形成的二维数据图则称为散点图。

例 8.1. 利用数据集 mtcars 中的数据，作出变量 mpg 和变量 wt 之间的散点图。

解：利用 R 软件作出散点图的程序如下：

```
data(mtcars)
plot(mtcars$wt,mtcars$mpg,col="red",
     xlab="Weight",ylab="Mile per Gallon")
```

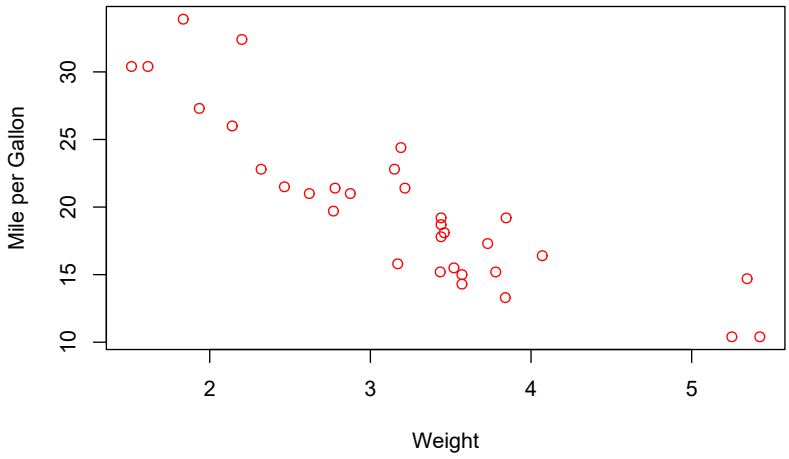


图 8.1: 汽车重量和汽车油耗的散点图

若两个变量之间的散点图近似地表现为一条直线，则这两个变量之间线性相关；若两个变量之间的关系近似地表现为一条曲线，则这两个变量之间非线性相关。在本科阶段的统计学中，只研究线性相关。在线性相关中，若两个变量的变动方向相同，即一个变量的数值增加，另一个变量的数值也随之增加，或者一个变量的数值减小，另一个变量的数值也随之减小，则这两个变量之间正相关；反之，若两个变量的变动方向相反，即一个变量的数值增加，另一个变量的数值随之减小，或者一个变量的数值减小，另一个变量的数值随之增加，则这两个变量之间负相关。由散点图 (8.1) 可知，每加仑油的行驶里程和汽车的重量之间负相关。

2. 相关系数

利用散点图描述两个变量之间的相关关系，最大的优点是生动、直观。他可以迅速判断两个变量之间有无线性相关关系，并对相关关系的形式进行大致地描述。但散点图的缺点在于无法精确反映两个变量之间的线性相关程度。要精确测度两个变量之间的相关关系，需要计算两个变量之间的线性相关系数。

定义 8.2. 线性相关系数 (Correlation coefficients) 是衡量两个变量之间线性相关程度的参数或统计量。

若线性相关系数是利用总体数据计算的，则称之为总体相关系数，其计算公式为

$$\rho = \frac{Cov(X,Y)}{SD(X)SD(Y)} \tag{8.1.1}$$

其中，$Cov(X,Y)$ 指的是两个变量的协方差；$SD(X)$ 和 $SD(Y)$ 分别表示两个总体的标准差。

可以证明，总体相关系数 ρ 满足

$$-1 \leq \rho \leq 1$$

若 $0 \leq \rho \leq 1$，则两个变量之间正相关；若 $-1 \leq \rho \leq 0$，则两个变量之间负相关。其中，若 $\rho = \pm 1$，则两个变量之间完全相关，若 $\rho = 0$ 则表示两个变量之间无线性相关关系。当然，两个变量之间无线性相关关系，并不表示两个变量之间无关。因为，两个变量之间可能存在某种非线性关系。根据计算得到的总体相关系数，可以对两变量之间的线性相关的密切程度进行简单的判断：

- 若 $|\rho| \geq 0.8$，则两个变量之间高度相关；
- 若 $0.5 \leq |\rho| < 0.8$，则两个变量之间中度相关；

- 若 $0.3 \leq |\rho| < 0.5$，则两个变量之间低度相关；
- 若 $|\rho| < 0.3$，则两个变量之间无线性相关。

一般情况下，我们得到的线性相关系数都不是利用总体数据计算的。

若线性相关系数是利用样本数据计算的，则称之为样本相关系数，其计算公式为

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{8.1.2}$$

8.1.3 相关系数的显著性检验

一般情况下，总体相关系数 ρ 是未知的，需要利用样本数据计算样本相关系数 r，作为总体相关系数的估计量。因此，可以将 ρ 视作一个未知但不变的总体参数，将 r 视作一个随机变量。由于 r 是一个随机变量，因此，它受抽样波动的影响。所以，在利用样本相关系数判断两个变量的相关程度时，需要进行样本相关系数的显著性检验。

对样本相关系数进行显著性检验，其基本步骤如下：

- 提出原假设和备择假设。该检验的原假设和备择假设分别为

$$H_0 : \rho = 0; \qquad H_1 : \rho \neq 0$$

- 确定检验的统计量及其分布

$$t = \sqrt{\frac{r^2}{1 - r^2}(n-2)} \sim t(n-2) \tag{8.1.3}$$

- 计算样本统计量
- 根据显著性水平 α(一般情况下，$\alpha = 0.05$)，确定检验的临界值
- 比较样本统计量和临界值，作出检验的统计决策。

例 8.2. 利用数据集 mtcars 中的 mpg 和 wt 的数据信息，计算两者间的线性相关系数，并在显著性水平 $\alpha = 0.05$ 的条件下，对该样本相关系数进行显著性检验。

解：利用数据集，计算变量 mpg 和 wt 之间的线性相关系数

```
data(mtcars)
r.mw <- cor(mtcars$mpg,mtcars$wt)
```

在原假设为真时，计算样本统计量

```
n <- nrow(mtcars)
tval <- r.mw*sqrt((n-2)/(1-r.mw^2))
```

计算当显著性水平 $\alpha = 0.05$ 时，该检验的临界值

```
alpha <- 0.05
crt <- qt(0.025,n-2)
```

由于该检验的临界值为-2.042，样本统计量的取值为-9.559，因此，可以拒绝原假设。即汽车的油耗与汽车的重量之间存在显著的相关关系。

需要说明的是，在对样本回归系数 r 进行显著性检验时，需要考察 r 的抽样分布。当总体相关系数 $|\rho|$ 很小或接近于 0 时，随着 n 的增加，r 的抽样分

布趋于正态分布。但当 $|\rho|$ 接近于 1 时，除非抽样的样本容量足够大，否则 r 的抽样分布呈现出一定的偏态。

总之，当 ρ 的取值越接近 1，则 r 呈左偏分布；当 ρ 的取值越接近-1，则 r 呈右偏分布。只有当 ρ 接近于 0，且在大样本抽样下，才能认为 r 接近正态分布。

8.1.4 回归分析

8.2 简单线性回归模型

只有一个解释变量的线性回归模型被称为简单线性回归模型。在实际应用中，简单线性回归模型是我们研究的出发点和基础。囿于篇幅和对本科统计学的要求，我们主要学习简单线性回归模型的设定、基本假定、参数估计、显著性检验、模型诊断和应用等方面的基础知识。

8.2.1 总体回归模型和样本回归模型

简单线性回归模型的总体回归模型可以表示为

$$Y_i = \beta_0 + \beta_2 X_i + u_i \quad (i = 1, 2, \cdots, N) \tag{8.2.1}$$

其中，Y_i 表示的是因变量及其取值；X_i 则表示自变量及其取值；β_0 和 β_2 是两个未知参数。u_i 表示的是随机扰动项。式 (8.2.1) 反映的是两个变量总体之间的关系，因此，该式又被称为总体回归模型。

由总体回归模型可以看出，简单线性回归模型由两部分组成：

- 确定性部分

在式 (8.2.1) 中，确定性部分是 $\beta_0 + \beta_2 X_i$，它刻画了变量 Y_i 和 X_i 之间的内在联系。

- 不确定性部分

在式 (8.2.1) 是模型的不确定性部分，即所谓的随机扰动部分。正是由于随机扰动部分的存在，导致被解释变量 Y_i 和解释变量之间的不确定性关系。

在简单线性回归模型中，加入随机扰动项 u_i 的原因在于：

1. 随机扰动项是已知的所有不重要因素影响的代表。影响因变量的因素有很多，但在建模时，我们要暂时忽略其他因素的影响，建立简单线性回归模型，因此，那些被忽略的因素，就构成了随机扰动项的重要组成部分。例如，在研究收入水平和消费支出之间关系时，建立的模型是

$$Consmpt_i = \beta_0 + \beta_2 Income_i + u_i$$

根据现有的认知，影响人们消费支出的因素主要是收入水平，但并不意味着其它因素不影响消费支出，只不过从简化的角度考虑，我们抓住最核心的影响因素建立模型，而将其它影响消费支出的因素—— 过去收入、预期收入和其他人的支出等——并入随机扰动项。

2. 随机扰动项是所有已知但无法取得数据的影响因素的代表。在上面的简单线性回归模型中，一个公认的对消费支出有重要影响的因素是家庭财产，它其实也是一个影响消费支出的重要因素。但在实际应用中，一般很难获得家庭财产数据，因此无法将该因素放入回归模型中，只能并入随机扰动项。

3. 随机扰动项是所有未知影响因素的代表。由于认识的相对局限性，我们可能无法穷尽所有影响消费的因素，因此，只能将其并入随机扰动项。

4. 随机扰动项是人的行为随机性的反映。人的决策行为本身就具有极强的内在的随机性，也就是说，即使面对同一决策，外部环境也没有变化，人的决策行为也有可能发生变化。在消费支出模型中，人的消费决策本身就具有内在的随机性，因此，只能将这种随机性并入随机扰动项。

由此可见，在简单线性回归模型中，随机扰动项包含的内容相当丰富。而且它们之间可能会相互抵消，因此，我们一般假定随机扰动项的数学期望值为 0。在这一假定下面，总体回归模型的随机形式可以写成其数学期望形式

$$E(Y_i) = \beta_0 + \beta_2 X_i \tag{8.2.2}$$

利用式 (8.2.1) 和式 (8.2.2) 对社会经济线性进行回归分析时，一个重要的任务就是得到未知参数 β_0 和 β_2 的取值。由于我们无法对总体进行全面的调查，一个可行的办法就是利用随机抽样的方法，从总体中抽取一部分样本，利用样本得到总体参数 β_0 和 β_2 的样本值，再根据样本值取推断总体参数的可能取值。样本回归模型可以表示为

$$Y_i = \hat{\beta}_0 + \hat{\beta}_2 X_i + e_i \quad (i = 1, 2, \cdots, n) \tag{8.2.3}$$

其中，$\hat{\beta}_0$ 和 $\hat{\beta}_2$ 是未知参数 β_0 和 β_2 的估计量或估计值；e_i 表示残差，它是总体回归模型中随机扰动项 u_i 的一个估计。

有时，样本回归模型又可以写成

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_i \tag{8.2.4}$$

8.2.2 简单线性回归模型的基本假定

得到样本回归模型后，现在的核心问题是如何利用样本信息得到参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_2$。在进行估计之前，需要对总体回归模型进行一定的假定。

1. 对模型的假定

对模型的假定主要包括：

- 模型的设定是正确的。
- 模型中因变量和自变量之间的关系是线性的。

2. 对解释变量的假定

在简单线性回归模型中，假定解释变量是有变化的。

3. 对随机扰动项的假定

对随机扰动项的假定包括：

- 随机扰动项 u_i 是一个随机变量，其数学期望值等于 0，即 $E(u_i) = 0$；
- 对于任意的 X_i，随机扰动项的方差相等，即 $Var(u_i) = \sigma^2$；
- 随机扰动项之间相互独立，即 $Cov(u_i, u_j) = 0 \quad i \neq j$；
- 随机扰动项和解释变量之间不相关，即 $Cov(x_i, u_i) \neq 0$。

若总体回归模型同时满足上述条件，可以用普通最小二乘法 (OLS) 对模型的参数估计量进行估计。

8.3 简单线性回归模型的参数估计

当总体回归模型满足一定的假定时，可以利用普通最小二乘法进行参数估计。在正式估计参数估计量之前，有必要了解普通最小二乘法的基本原理。

8.3.1 普通最小二乘法的基本原理

假定样本回归模型为

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_i \tag{8.3.1}$$

其中，$\hat{Y}_i$ 表示根据样本回归模型得到的预测值，比较式 (8.2.3) 和式 (8.3.1) 可得：

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i \tag{8.3.2}$$

普通最小二乘法的实质是确定参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_2$ 的取值，使得残差平方和

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_2 X_i)^2 \tag{8.3.3}$$

最小。

因此，普通最小二乘法的实质，是确定参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_2$ 的取值，使函数 $f(\hat{\beta}_0, \hat{\beta}_2)$ 取得最小值

$$f(\hat{\beta}_0, \hat{\beta}_2) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_2 X_i)^2 \tag{8.3.4}$$

这就是普通最小二乘法的含义。

8.3.2 普通最小二乘估计量

根据函数取极值的一阶条件，可得

$$\begin{aligned} \frac{\partial f}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_2 X_i) = 0 \\ \frac{\partial f}{\partial \hat{\beta}_2} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_2 X_i) X_i = 0 \end{aligned} \tag{8.3.5}$$

经整理得方程组

$$\begin{aligned} n\hat{\beta}_0 + \sum X_i \hat{\beta}_2 &= \sum Y_i \\ \sum X_i \hat{\beta}_0 + \sum X_i^2 \hat{\beta}_2 &= \sum X_i Y_i \end{aligned} \tag{8.3.6}$$

解得

$$\begin{aligned}\hat{\beta}_2 =& \frac{n\sum X_iY_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i^2)} \\ \hat{\beta}_0 =& \bar{Y} - \hat{\beta}_0\bar{X}\end{aligned} \tag{8.3.7}$$

上式经过化简后，可写成

$$\begin{aligned}\hat{\beta}_2 =& \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ \hat{\beta}_0 =& \bar{Y} - \hat{\beta}_0\bar{X}\end{aligned} \tag{8.3.8}$$

比较式 (8.3.8) 和式 (8.1.2) 可知，参数估计量 $\hat{\beta}_2$ 和样本相关系数 r 的符号相同。

从普通最小二乘估计量的推导过程可以看出，简单线性回归模型具有以下三个方面的特征：

1. 样本回归模型所代表的一条直线，过点 $(\bar{X}, \bar{Y})$。

由式 (8.3.7) 的第二个式子可知，点 $(\bar{X}, \bar{Y})$ 的坐标满足

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_2\bar{X}$$

因此，点 $(\bar{X}, \bar{Y})$ 的坐标满足线性回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_i$$

因此，样本回归模型过点 $(\bar{X}, \bar{Y})$。利用这一特点，可将样本回归模型写成离差形式

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2(X_i - \bar{X}) \tag{8.3.9}$$

2. 残差和 $\sum e_i$ 等于 0。

由式 (8.3.5) 中的第一个式子，可知

$$\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i) = 0$$

即残差的和等于 0。

3. 残差 e_i 和解释变量 X_i 之间不相关。

由式 (8.3.5) 中的第二个式子可得

$$\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i)X_i = 0$$

即

$$\sum e_i X_i = 0$$

因此，残差和解释变量的协方差

$$Cov(e_i, X_i) = \sum (e_i - \bar{e})(X_i - \bar{X}) = \sum e_i X_i - \bar{X} \sum e_i = 0$$

所以，残差和解释变量之间不相关。

4. 残差和被解释变量的预测值之间无关。

容易证明

$$Cov(e_i, \hat{Y}_i) = \sum e_i \hat{Y}_i = \sum e_i(\hat{\beta}_0 + \hat{\beta}_2 X_i) = 0 \tag{8.3.10}$$

因此，残差和解释变量之间不相关。

例 8.3. 利用数据集 mtcars 中的数据，计算样本回归模型

$$MPG_i = \hat{\beta}_0 + \hat{\beta}_2 Weight_i + e_i$$

中的参数估计量的值。

解：计算参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_2$ 值的 R 软件程序如下

```
data(mtcars)
n <- nrow(mtcars)
cov.mw <- cov(mtcars$mpg,mtcars$wt)
sd.w <- sd(mtcars$wt)
hatb2 <- cov.mw/(sd.w^2)
m.m <- mean(mtcars$mpg)
m.w <- mean(mtcars$wt)
hatb0 <- m.m-hatb2*m.w
```

由此可得该线性回归模型的表达式为

$$\widehat{MPG_i} = 37.285\text{+-}5.344\ Weight_i$$

在本例中，估计参数估计量的另一种方法时利用 R 软件自带的 *lm()* 命令

```
mod <- lm(mpg~wt,data=mtcars)
smod <- summary(mod)
```

此时，R 软件将回归模型的信息全部存储于一个名为 mod 的对象中，利用 *summary()* 提取其中的信息存储于一个名为 smod 的对象中。关于 *lm()* 命令和 *summary()* 命令的具体使用方法，可以用

```
help('lm')
help('summary')
```

命令查看其帮助文件。

表 8.3.1: 简单线性回归模型的结果

term	estimate	std.error	statistic	p.value
Intercept	37.29	1.878	19.86	0
Weight	-5.34	0.559	-9.56	0

利用利用 **broom** 包中的 *tidy()* 命令和 **knitr** 包中的 *kable()* 命令可以将结果以三线表的形式输出。由于这两个包都不是 R 软件自带的包，因此，在调用该包之前，需要用

```
install.packages('broom')
install.packages('knitr')
```

命令来安装它们，并调用它们

```
library(knitr)
library(broom)
tmod <- tidy(smod)
tmod$term<- c("Intercept","Weight")
```

利用 **broom** 包和 **knitr** 包，将估计的最终结果输出，如表 (8.3.1) 所示：[1]

```
knitr::kable(tmod,digits=4,booktabs=T,
             caption = "简单线性回归模型的结果")
```

[1]在本书中，回归结果的输出不是重点，若没有特殊需要，此处的输出可以看成其标准形式。后面有相同的输出时，就不再展示其源代码。

8.4 简单线性回归模型的评价

简单线性回归模型的评价主要包括三个方面的内容：

- 模型的拟合优度
- 模型的显著性检验
- 参数的显著性检验

8.4.1 模型的拟合优度

样本回归模型

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_i$$

是一条直线，可以称之为样本回归直线。它在一定程度上描述了因变量 Y_i 和自变量 X_i 之间的数量关系。根据线性回归直线，当自变量的取值确定后，可以用来估计或预测因变量的取值。但这里需要注意的是，由于总体回归模型中存在随机扰动项，抽样的时候存在抽样误差，因此，依据样本回归模型得到的预测值和抽取的样本观测值之间不可能完全相等，二者之间的差异就是残差 e_i。因此，回归模型好不好，就取决于残差 e_i。一个直观的感觉是，若样本观测值与预测值之间的误差越小，则模型的拟合程度就越高，反之，则模型的拟合程度就越低。我们将样本回归直线和样本观测值之间的接近程度称为模型的拟合优度。模型的拟合优度可以用判定系数来度量。

1. 判定系数

判定系数是说明样本回归模型拟合优度的一个统计量。它从模型对因变量变化的解释程度入手，来说明模型的拟合优度。

由前面的推导过程可知，样本回归模型是一条直线。该直线始终过被解释变量和解释变量的样本均值组成的点 $(\bar{X}, \bar{Y})$，如图 (8.2) 所示

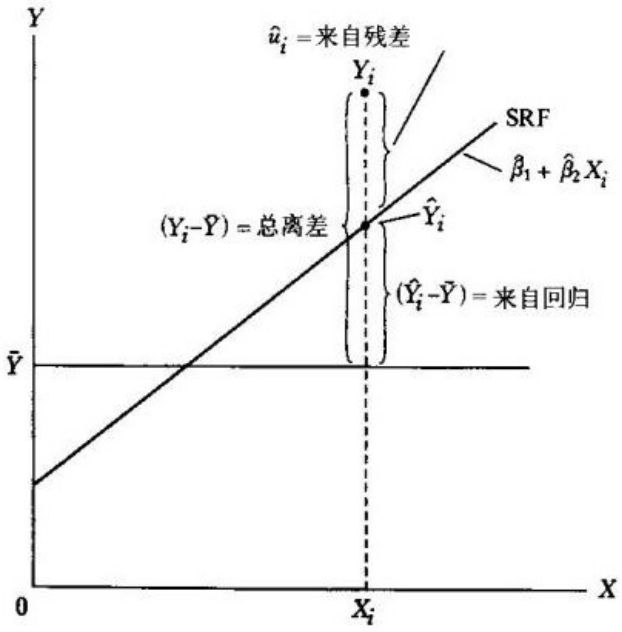


图 8.2: 总离差分解图

那么，一个很自然的问题是，因变量为什么会发生变化，和点 $(\bar{X}, \bar{Y})$ 发生偏离。其根本原因有两个：

- 自变量 X 的变化引起了因变量的变化；
- 自变量以外的所有因素变化引起的。

但是，我们不知道因变量变化中，有多少是自变量变化引起的，有多少是其他因素造成的，因此，需要对因变量的变化进行分解。其基本的思路是：假定一开始位于 $(\bar{X}, \bar{Y})$ 处，现在位于 (X_i, Y_i) 处，因变量的变化可以分解为两部分：

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i \tag{8.4.1}$$

可以发现，$\hat{Y}_i - \bar{Y}$ 表示的是因自变量变化而导致的因变量变化，即可以用样

本回归模型解释的部分；$Y_i - \hat{Y}_i$ 表示的是因其他因素变化而导致的变化，即无法用样本回归模型解释的部分。

将式 (8.4.1) 两边平方，可得

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \tag{8.4.2}$$

对每一点都作相同的处理，并求和，可得

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 = &\sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \\ &+ 2\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \end{aligned} \tag{8.4.3}$$

又因为

$$\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum \hat{Y}_i e_i = \sum (\hat{\beta}_0 + \hat{\beta}_2 X_i) e_i$$

由式 (8.3.10) 可知，上式等于 0。因此，式 (8.4.3) 可以写成

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \tag{8.4.4}$$

其中，$\sum (Y_i - \bar{Y})^2$ 被称为总离差平方和 (SST)，表示的是被解释变量的总的变化；$\sum (\hat{Y}_i - \bar{Y})^2$ 被称为解释平方和 (SSE)，表示的是可以用模型解释的因变量的变化；$\sum (Y_i - \hat{Y}_i)^2$ 被称为残差平方和 (SSR)，表示的是无法用模型解释的因变量的变化。因此，式 (8.4.4) 可以写成

$$SST = SSE + SSR \tag{8.4.5}$$

由图 (8.2) 可以发现,样本回归模型对数据的拟合程度取决于解释平方和 SSE 占总离差平方和的比重。若解释平方和占总离差平方和的比重越大，则模型的解释能力越强。因此，我们可以用解释平方和占总离差平方和的比重来判断模型对数据拟合的优劣。

定义 8.3. 判定系数 (Coefficient of Determination) 指的是解释平方和占总离差平方和的比重。

一般用 R^2 表示判定系数，它的计算公式为。

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \tag{8.4.6}$$

判定系提供了模型对现实的解释能力的判断标准。两种极端情况是：若所有的点都落在直线上，即残差平方和 $SSR = 0$，则模型的拟合优度 $R^2 = 1$，说明利用解释变量 X 的变化可以解释 Y 的所有变化，即模型是完全拟合的；反之，若 $SSE = 0$，则模型的拟合优度 $R^2 = 0$，说明解释变量的变化对被解释变量的变化毫无解释作用。一般情况下，拟合优度 R^2 的取值范围总在 0 到 1 之间。R^2 越接近 1，表明解释平方和占总离差平方和的比重越大，模型的解释能力越强；反之，R^2 越接近 0，表示解释平方和占总离差平方和的比重越小，模型的解释能力越弱。

由式 (8.3.9) 可知，解释平方和可以化简为

$$SSE = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2 \sum (X_i - \bar{X})(Y_i - \bar{y}) \tag{8.4.7}$$

因此，有

$$\begin{aligned} R^2 =& \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\hat{\beta}_2 \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \\ =& \left[\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \right] = r^2 \end{aligned} \tag{8.4.8}$$

因此，判定系数在数值上和样本相关系数相等。但需要注意的是，两者的意义有着根本的区别。

例 8.4. 根据例 (8.3) 的计算结果，计算该模型的拟合优度。

解： 利用 R 软件计算模型的拟合优度的程序如下

```
mean.mpg <-mean(mtcars$mpg)
SST.mpg <- sum((mtcars$mpg-mean.mpg)^2)
hat.mpg <- hatb0+hatb2*mtcars$wt
SSE.mpg <- sum((hat.mpg-mean.mpg)^2)
sqr <- SSE.mpg/SST.mpg
```

由此可得，该模型的拟合优度为 0.753。即该模型可以解释因变量变化的 75.28%。由此可见，该模型具有较强的解释能力。

当然，由于我们已经利用 *lm()* 命令估计出模型的参数，并利用 *summary()* 命令将其概要信息存入一个名为 **smod** 的对象中一个名为 *r.squared* 的变量中，因此，可以直接从其中调用样本回归模型拟合优度的信息

```
smod$r.squared
```

这里不再展示其结果，留给读者自行验证。

2. 估计的标准误

判断系数从正面衡量模型对数据的拟合程度，而估计的标准误则从反面衡量模型对数据的拟合程度。

定义 8.4. 估计的标准误 (Standard Error of Estimate) 是估计量的标准误差的简称，指的是残差平方和的均值的平方根。

一般用 s_y 表示估计的标准误。估计的标准误的计算公式为

$$s_y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-2}} = \sqrt{\frac{SSR}{n-2}} = \sqrt{MSE} \tag{8.4.9}$$

估计的标准误可以看成是在排除了解释变量的变化对被解释变量的变化的影响后，被解释变量的随机波动性，它实际反映的是利用回归模型对被解释变量 y 进行预测时预测误差的大小。

由估计的标准误可以看成，该统计量其实是总体回归模型中随机扰动项的标准差的一个估计量，即有

$$\hat{\sigma_u^2} = \frac{\sum(Y_i - \bar{Y})^2}{n-2} \tag{8.4.10}$$

例 8.5. 根据前面各例的结果，估计该模型的标准误。

解： 利用 R 软件计算标准误的程序为

```
res <- mtcars$mpg-hat.mpg
n <- nrow(mtcars)
std.mpg <- sum(res^2)/(n-2)
```

得到结果为 9.277。其含义是利用该样本回归模型对每加仑汽油的行驶里程数进行预测时，平均的误差为 9.277 英里。

本例中计算估计的标准误的另一种方法时调用 **smod** 中的残差，利用残差计算估计的标准误。

```
res2 <- smod$residuals
std.mpg2 <- sum(res2^2)/(n-2)
```

可以得到和前面编程计算一样的结果，这里也不再给出，请读者自行验证。

8.5 简单线性回归模型的检验

回归分析的目的在于应用样本回归模型，用自变量 X_i 的变化来预测因变量 Y_i 的变化。但我们在得到样本回归模型后，在应用它进行预测之前，还需要进行显著性检验。它主要包括几个方面的内容：

- 模型的显著性检验
- 参数的显著性检验
- 残差的正态性检验
- 离群点、强影响点和高杠杆值点检验

在简单线性回归模型中，二者的几乎没有区别，但在多元线性回归模型中，二者的差别却相当大。若用一句话来概括两者的差异，可以表述为：模型的显著性检验从模型设定的角度进行检验；参数的显著性检验从参数估计量扰动的角度进行检验。

8.5.1 模型的显著性检验

模型的显著性检验从模型设定的角度来考察建模的合理性。在总体回归模型

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$

中，假定解释变量 X_i 前面的系数 β_2 等于 0。因此，模型的显著性检验中的原假设和备择假设分别为

$$H_0 : \beta_2 = 0; \quad H_1 : \beta_2 \neq 0$$

从模型的设定角度看，若原假设成立，则被解释变量 Y_i 的变化和解释变量 X_i 没有任何关系，完全由随机扰动项决定。即模型可以写成

$$Y_i = \beta_0 + u_i \tag{8.5.1}$$

我们将式 (8.5.1) 称为约束模型，因为它规定解释变量前面的系数 β_2 的取值为 0。在此情况下，未知参数 β_0 的最佳估计量是被解释变量 Y_i 的样本均值。即有样本回归模型：

$$Y_i = \bar{Y} + e_i$$

由此可得约束模型的总离差平方和于残差平方和之间的关系为：

$$SSR_R = \sum e_i^2 = \sum (Y_i - \bar{Y})^2 = SST \tag{8.5.2}$$

如果约束模型的约束是正确的，那么我们利用无约束模型

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$

估计总体参数时，得到的残差平方和为

$$SSR_U = SST - SSE$$

易证

$$SSR_R \sim \chi^2(n-1) \quad SSR_u \sim \chi^2(n-2)$$

根据 χ^2 分布的可列可加性，可得该检验的统计量

$$F = \frac{(SSR_R - SSR_U)/[(n-1)-(n-2)]}{SSR_U/(n-2)} \sim F(1, n-2)$$

在约束模型中，由于 $SSR_R = SST$，上式可化为

$$F = \frac{SSE/1}{SSR_U/(n-2)} \sim F(1, n-2) \tag{8.5.3}$$

在给定的显著性水平 α 下，可得其临界值为 $F_{1-\alpha}(1, n-2)$，若样本统计量大于其临界值，则拒绝原假设，认为模型是显著的，即该模型选用 X_i 作解释变量是合理的；反之，则无法拒绝原假设，认为模型是不显著的，即该模型选用 X_i 作解释变量是不合理的。

例 8.6. 利用例 (8.4) 估计的结果，对模型进行显著性检验。

解： 首先，计算 F 统计量

```
n <- nrow(mtcars)
alpha <- 0.05
SSR.r <- SST.mpg
SSR.u <- SST.mpg-SSE.mpg
Fval <- (SSR.r-SSR.u)/(SSR.u/(n-2))
```

由此可得，该检验的 F 统计量的值为 91.375。当显著性水平 $\alpha = 0.05$ 时，可以计算该检验的临界值

```
crtF <- qf(1-alpha,df1=1,df2=n-2)
```

由于该检验的临界值为 4.171，因此可以拒绝原假设，即该模型的设定是合理的。

当然，在本例中，也可以直接调用 **smod** 中的 F 统计量的值，其结果和上面编程结果一致。

```
smod$fstatistic
```

```
## value numdf dendf
##  91.4   1.0  30.0
```

8.5.2 参数的显著性检验

参数的显著性检验主要是检验自变量对因变量的影响是否显著的问题。在简单线性回归模型 $Y_i = \beta_0 + \beta_2 X_i + u_i$ 中，若参数 $\beta_2 = 0$，说明解释变量 X_i 的

变化对被解释变量 Y_i 没有影响，即两个变量之间没有线性相关关系。由于总体参数是未知的，我们得到的是利用样本观测值计算出来的参数估计量。由于存在抽样误差，即使参数的真实值为 0，通过抽样得到的参数估计量 $\hat{\beta}_2$ 也有可能不为 0。所以，在我们得到参数估计量不为 0 的时候，并不意味着参数真实值一定不为 0，必须要进行参数的显著性检验。该检验的原假设和备择假设分别为

$$\beta_2 = 0 \quad \beta_2 \neq 0$$

由于样本估计量 $\hat{\beta}_2$ 是通过样本观测值计算得到的，因此，该估计量的取值会随着样本的变化而变化，是一个随机变量。根据大数定律和中心极限定理，当总体回归模型满足经典假定时，在大样本抽样条件下，样本统计量 $\hat{\beta}_2$ 满足

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}\right) \tag{8.5.4}$$

其中，σ_u^2 表示的是随机扰动项 u_i 的方差。当随机扰动项的方差已知时，可以得到

$$z = \frac{\hat{\beta}_2 - \beta_2}{se\left(\hat{\beta}_2\right)} \sim N(0,1) \tag{8.5.5}$$

其中，$se\left(\hat{\beta}_2\right)$ 表示的是参数估计量 $\hat{\beta}_2$ 的标准差。即

$$se\left(\hat{\beta}_2\right) = \sqrt{\frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}} = \frac{\sigma_u}{\sqrt{\sum (X_i - \bar{X})^2}}$$

但在现实中，由于随机扰动项的方差是未知的，因此，需要用残差对其进行估计。随机扰动项的方差的无偏估计量为：

$$\hat{\sigma_u^2} = \frac{\sum e_i^2}{n-2} \tag{8.5.6}$$

将其代入式 (8.5.5) 得

$$t = \frac{\hat{\beta}_2 - \beta_2}{\hat{se}(\hat{\beta}_2)} \sim t(n-2) \tag{8.5.7}$$

当原假设成立时，可以得到 t 统计量

$$t = \frac{\hat{\beta}_2}{\hat{se}(\hat{\beta}_2)} \sim t(n-2) \tag{8.5.8}$$

利用该式可以计算得到样本统计量的大小。在显著性水平为 α 时，得到该检验的临界值。最后，比较样本统计量和临界值的大小，作出统计决策。

例 8.7. 根据前面各例的计算结果，当显著性水平 $\alpha = 0.05$ 时，对变量 Weight 前面的系数进行显著性检验。

解： 该检验的原假设和备择假设分别为

$$\beta_2 = 0; \quad \beta_2 \neq 0$$

首先计算检验的样本统计量

```
res <- mtcars$mpg-hat.mpg
hat.sigmasq <- sum(res^2)/(n-2)
SST.wt <- sum((mtcars$wt-mean(mtcars$wt))^2)
hat.se <- sqrt(hat.sigmasq/SST.wt)
tval <- hatb2/hat.se
```

当显著性水平 $\alpha = 0.05$ 时，可以得到该检验的临界值

```
alpha <- 0.05
crtval <- qt(alpha/2,n-2)
```

由于该检验的临界值为-2.042，样本统计量为-9.559，落入拒绝域。因此，应当拒绝原假设，即变量 **Weight** 对汽车的油耗有显著性影响。

在本例中，由于 **smod** 中有显著性检验的统计量信息，因此可直接调用该信息

```
tval2 <- smod$coefficients[2,3]
tval2
## [1] -9.56
```

其结果和编程结果一致。

8.5.3 残差的正态性检验

在回归模型

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$

中，我们曾假定随机扰动项 u 满足下列条件。

- 随机扰动项的数学期望是 0，即 $E(u_i|X_i) = 0$
- 随机扰动项的方差相等，即 $Var(u_i|X_i) = \sigma^2$
- 随机扰动项之间不相关，即 $Cov(u_i, u_j) = 0$
- 随机扰动项与解释变量间不相关，即 $Cov(u_i, X_i) = 0$

当模型满足上述假定时，我们将之称为经典的线性回归模型。

当模型满足上述假定时，在大样本抽样条件下，模型的残差服从正态分布。我们可以利用它们进行模型的显著性检验和参数的显著性检验等。因此，当样本容量足够大时，随机扰动项的正态分布假定就不是必要的。但在样本容量不够大时，若没有随机扰动项的正态分布的保证，则必须进行残差的正态性检验。

这里首先给出残差的定义：

定义 8.5. 残差 (Residuals) 指的是因变量的观测值和根据样本回归模型得到预测值之间的差异。

残差的正态性检验包括两类方法：

- 基于经验的残差图法
- 基于假设检验的雅克-贝拉检验

在样本回归模型中，参数估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_2$ 是利用样本观测值，采用普通最小二乘法得到参数 β_0 和 β_2 的估计量。样本回归模型可以写成

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_2 X_i \tag{8.5.9}$$

根据残差的定义，可将第 i 个观测值的残差写成

$$e_i = Y_i - \hat{Y}_i \tag{8.5.10}$$

分析残差是否服从正态分布，可以通过残差图来完成。常用的残差图包括关于解释变量 X 的残差图和标准化残差图等。

残差分析中，关于解释变量的残差图是最常用的残差图。若模型对变量之间的关系的设定是合理的，在同方差假定下，残差图中所有的点都应当在其 2 倍标准差范围之内。首先作出残差关于解释变量的散点图。

```
plot(smod$residuals~mtcars$wt,ylim=c(-7,7),
     xlab="Weight of Car ",ylab =" Residuals of the Model")
abline(h=c(-2*smod$sigma,2*smod$sigma),col="red",lty=2)
```

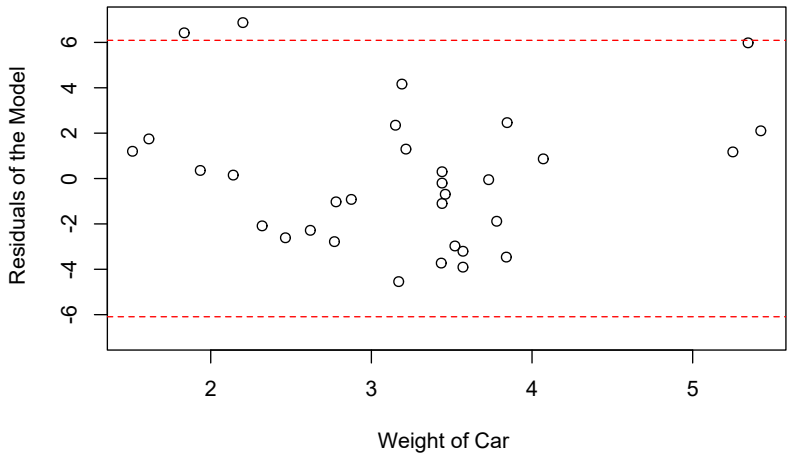


图 8.3: 汽车重量和行驶里程的残差图

由残差和解释变量的散点图可以发现，除了两个特殊的点外，其他残差点基本上都位于中间的一条水平带内，因此，我们所选的模型是合理的，且不存在所谓的异方差问题。由于该数据是横截面数据，因此，我们假定其不存在自相关性，另外，从关于解释变量的残差图还可以看出，解释变量和残差之间也应当是无关的。

标准化残差图为判断残差是否服从正态分布提供了直观的手段。

定义 8.6. 标准化残差 (Standardized Residual) 是指用残差除以它的标准差后得到的数值。

一般情况下，标准化残差用 z_{e_i} 表示。

根据定义，第 i 个观测值的标准化残差可以写成

$$z_{e_i} = \frac{e_i}{s_{e_i}} = \frac{Y_i - \hat{Y}_i}{s_{e_i}} \tag{8.5.11}$$

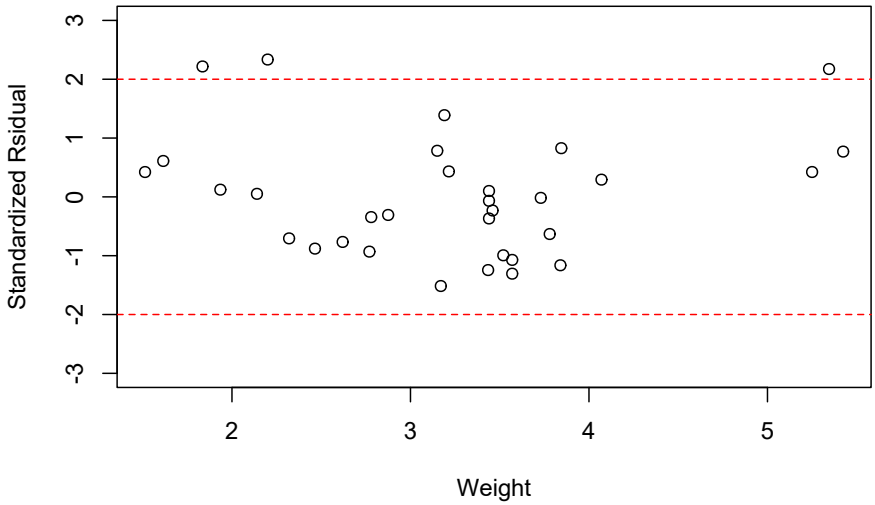
其中，残差的标准差满足

$$s_{e_i} = \sqrt{1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)} \sigma_u \tag{8.5.12}$$

利用标准化残差，可以很轻易地判断残差是否服从正态分布，若残差服从正态分布，则标准化残差服从标准正态分布，因此，在其标准化残差图中，大约由 95% 的标准化残差位于区间-2 到 2 之间。为此，我们可以将前面估计的模型中的残差标准化，得到标准化残差图

```
res <- smod$residuals
s.u <- smod$sigma
m.wt <- mean(mtcars$wt)
sq.devwt <- (mtcars$wt-m.wt)^2
coef.wt <- sq.devwt/SST.wt
res.coef <- sqrt(1-(1/n+coef.wt))
s.res <- s.u*res.coef
zs <- res/s.res
```

```
plot(mtcars$wt,zs,ylim=c(-3,3),xlab = "Weight",
     ylab="Standardized Rsidual")
abline(h=c(-2,2),col="red",lty=2)
```



由标准化的残差图可以发现，除了三点在区间 $[-2, 2]$ 以外，其他的都在区间内。说明它服从正态分布。

这里需要注意的是，在 R 中，可以用 *rstandard()* 命令来得到标准化残差。

```
std.res <- rstandard(mod)
plot(mtcars$wt,std.res)
```

可以得到相同的图形。

虽然残差图和标准化残差图都对残差是否服从正态分布给出了比较直观的展示，但它们却无法回答残差是否服从正态分布。判断残差是否服从正态分布，最好的办法就是对其进行正态性检验。而雅克- 贝拉检验是判断残差是否服从正态分布的一种方法。

雅克-贝拉检验的原假设和备择假设分别为：

H_0 : 样本数据服从正态分布; H_1 : 样本数据不服从正态分布

在原假设成立的条件下，统计量

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \sim \chi^2(2) \tag{8.5.13}$$

其中，n 表示样本容量，S 表示样本数据的偏度，K 表示样本数据的峰度。

在式 (8.5.13) 中，偏度和峰度的计算如下

$$S = \frac{\frac{1}{n}\sum (X_i - \bar{X})^3}{\left(\frac{1}{n}\sum (X_i - \bar{X})^2\right)^{\frac{3}{2}}} \tag{8.5.14}$$

$$K = \frac{\frac{1}{n}\sum (X_i - \bar{X})^4}{\left(\frac{1}{n}\sum (X_i - \bar{X})^2\right)^{2}} \tag{8.5.15}$$

例 8.8. 根据前面估计的结果，在显著性水平 $\alpha = 0.05$ 时，对模型的残差进行正态性检验。

解： 首先计算模型残差的偏度和峰度

```
fsum <- sum(smod$residuals^4)
tsum <- sum(smod$residuals^3)
sqsum <- sum(smod$residuals^2)
s <- (tsum/n)/(sqsum/n)^{3/2}
k <- (fsum/n)/(sqsum/n)^2
```

因此，可得 JB 统计量

```
JB <- n/6*(s^2+(k-3)^2/4)
```

当显著性水平 $\alpha = 0.05$ 时，该检验的临界值为

```
alpha <- 0.05
crt.l <- qchisq(alpha/2,df=2)
crt.u <- qchisq(1-alpha/2,df=2)
cbind(crt.l,JB,crt.u)
##       crt.l  JB crt.u
## [1,] 0.0506 2.4  7.38
```

样本统计量 JB 的值为 2.399，落入无法拒绝域区间，因此，无法拒绝原假设，即认为残差服从正态分布。

当然，我们也可以利用 **tseries** 包中的 *jarque.bera.test()* 来对样本数据进行雅克-贝拉检验

```
library(tseries)
jarque.bera.test(smod$residuals)
##
##  Jarque Bera Test
##
## data:  smod$residuals
## X-squared = 2, df = 2, p-value = 0.3
```

可以得到相同的结果。但由于 **tseries** 包并非 R 软件自带的包，因此在调用该包前，需要读者自行安装。

8.5.4 离群值、强影响点和高杠杆值点检验

离群值点检验

在一元线性回归模型中，可以通过散点图来观测异常值。在散点图中，若一个点于其他点所呈现的趋势不相吻合，这个点就可能是异常点。当然，异常点有其判断标准，一般来说，若某一观测值具有较大的残差，或其标准化残差大于 2 或小于-2，则认为该点事异常值。因此，在标准化的残差图中，有三点都是异常值点。R 软件的 **car** 包提供了判断异常值点的命令 *outlierTest()*，利用该命令，可以求得回归模型中的异常值点

```
library(car)
outlierTest(mod)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##          rstudent unadjusted p-value Bonferroni p
## Fiat 128     2.54             0.0168        0.537
```

这里需要注意的是，产生异常值的原因比较复杂，既有可能是因人为的错误引起的，也可能是因为模型的设定存在问题导致的，当然，也有可能是由随机因素导致的。因此，所采用的方法也有所不同。大致的原则是，若异常值是因人为的错误导致，则需要修正数据，若数据无法修正，则必须删除数据；但若异常值由随机因素导致的，则它是一个有效的观测值，应当保留在数据集中。

强影响点检验

定义 8.7. 强影响点是对模型的参数估计值的影响比例失衡的点。

在简单线性回归模型中，强影响点可以从回归模型中显示出来。一个有显著影响的

值可能是个异常值，即被解释变量 Y 远远偏离了散点图中的趋势线；也有可能是远离自变量 X 均值的点；或者是两者之间组合而成的观测值。是否需要修改、删除或保留，必须慎重对待。如果该点是一个有效的观测值，则应当被保留。因为该观测值可以很好地帮助我们分析所设定的模型是否合理。

一般用 Cook's D 统计量来判断强影响点，其基本思路是：若一个观测值的 D 满足

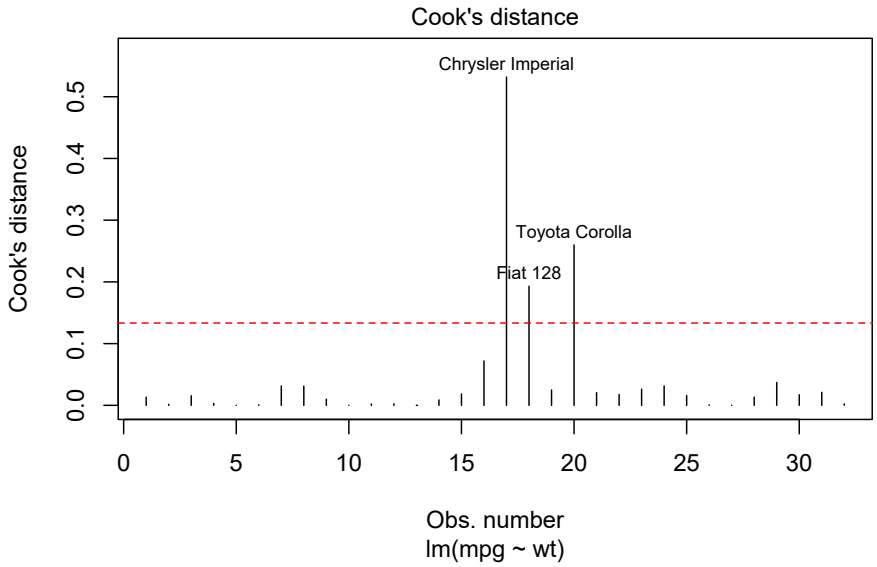
$$D > \frac{4}{n-k-1} \tag{8.5.16}$$

则认为该点是强影响点。其中，D 表示 Cook's D 统计量，它的计算过程十分复杂，R 软件计算该值的命令是 *cooks.distance()*，n 表示样本容量，k 表示模型中待估计参数的个数。下面，我们可以利用该命令计算 mtcars 回归模型中各点的 D 值。

```
D.cook<- cooks.distance(mod)
k<- 1
n <- nrow(mtcars)
crt.D <- 4/(n-k-1)
D.cook[D.cook>crt.D]
## Chrysler Imperial          Fiat 128    Toyota Corolla
##             0.532             0.193             0.260
```

由此可见，该模型中有三个强影响点，和我们前面的标准化残差图看到的结果相一致。其更直观的表现是作强影响点图

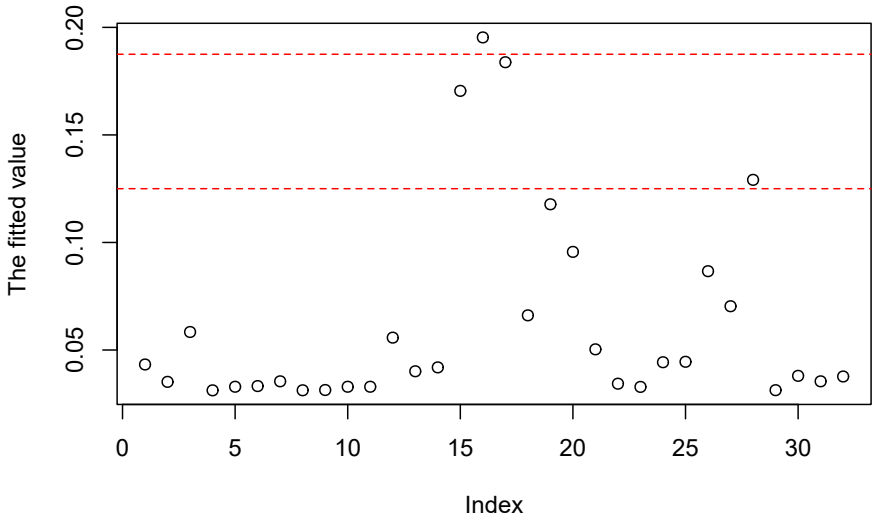
```
plot(mod,which=4,cook.levels = crt.D)
abline(h=crt.D,lty=2,col='red')
```

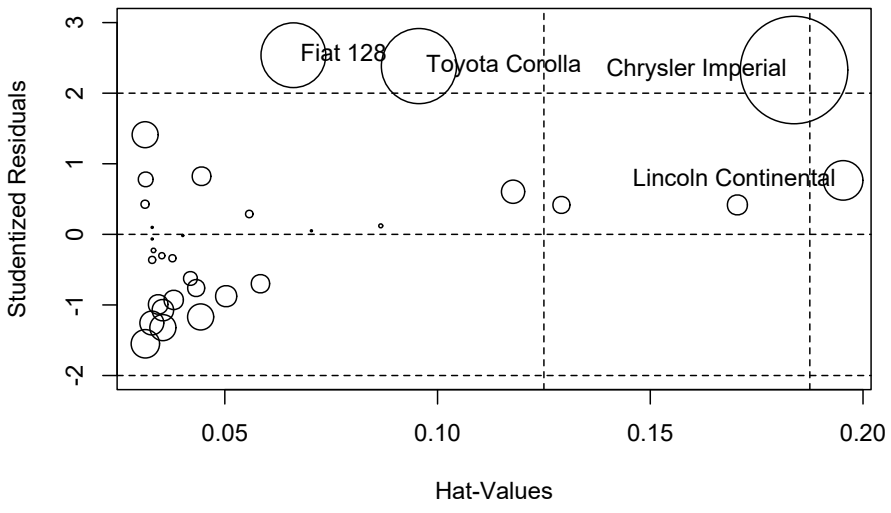


有图可以发现，该模型中三个强影响点分别是第 17 个、18 个和 20 个观测值。

高杠杆值点检验

```
hat.plot <- function(fit){
  p <- length(coefficients(mod))
  n <- length(fitted(mod))
  plot(hatvalues(fit),ylab='The fitted value')
  abline(h=c(2,3)*p/n,col='red',lty=2)
}
hat.plot(mod)
```


The fitted value
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
Index


Studentized Residuals
-2
-1
0
1
2
3
Fiat 128
Toyota Corolla
Chrysler Imperial
Lincoln Continental
0.05
0.10
0.15
0.20
Hat-Values

```
##                      StudRes    Hat  CookD
## Lincoln Continental    0.764 0.1953 0.0719
## Chrysler Imperial      2.328 0.1838 0.5319
## Fiat 128               2.538 0.0661 0.1930
## Toyota Corolla         2.384 0.0956 0.2599
```

8.6 简单线性回归模型的应用

样本回归模型经检验后，可以用来进行预测，即利用自变量的取值来预测因变量的取值。这既是回归分析的目的，更是回归分析实际应用性的体现。利用简单线性回归模型进行预测，从方法上看，主要有点估计和区间估计，从内容上看，则包括对均值的预测和对个别值的预测。

8.6.1 点估计

利用样本回归模型，给定自变量的一个特定取值，可以求出因变量一个估计值，这就是点估计。一般而言，点估计分为平均值的点估计和个别值的点估计两种。若利用解释变量的特定值，求出被解释变量的平均值就是一个估计值，称为平均值的点估计。若利用解释变量的特定值，求出被解释变量的个别值的一个估计值，就是个别值的点估计。

在点估计条件下，对于同一个解释变量的取值，对均值的点估计和对个别值的点估计相等，但两者的区间不同。

例 8.9. 利用 mtcars 数据集，建立油耗和重量之间的简单线性回归模型，并预测当 wt=3 时汽车油耗的点估计值。

解：首先估计简单线性回归模型的参数，然后利用参数估计量计算被解释变量的值。

```
mod <- lm(mpg~wt,data=mtcars)
b0 <- coef(mod)[1]
b2 <- coef(mod)[2]
weight <-3
yhat <- b0+b2*weight
```

计算结果表明，当汽车的重量为 3000(lbs) 时，汽车的油耗是 21.252。

8.6.2 区间估计

利用样本回归模型，对于自变量的一个特定值 x_0，求出因变量的估计值的区间，就是区间估计。区间估计有两种情形：

- 估计因变量个别值的置信区间；
- 估计因变量的均值的置信区间；

1. 因变量个别值的区间估计

假定当解释变量为 X_0 时，其真实值和预测值分别为 Y_0 和 $\hat{Y}_0$，因此有

$$e_0 = Y_0 - \hat{Y}_0 \tag{8.6.1}$$

其中，e_0 是真实值和预测值之间的差异，即预测误差。由于预测值 $\hat{Y}_i$ 是一个随机变量，在大样本或总体服从正态分布情形下，预测误差 e_0 服从正态分布。可以证明，预测误差 e_0 满足[2]

$$e_0 \sim N\left[0, \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right)\sigma_u^2\right] \tag{8.6.2}$$

[2]关于预测误差的标准差推导过程，见附录。

其中，σ_u^2 是总体方差，一般情况下是未知的，因此，可以用残差平方和取估计它。即有

$$\hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2} \tag{8.6.3}$$

因此，可以构造一个服从自由度为 $(n-2)$ 的 t 分布

$$t = \frac{Y_0 - \hat{Y}_0}{\hat{\sigma_{e_0}}} \tag{8.6.4}$$

因此，在给定的显著性水平 α 后，可以求得个别值的置信区间为：

$$[\hat{Y}_0 - t_{1-\alpha/2}\hat{\sigma}_{e_0},\ \hat{Y}_0 + t_{1-\alpha/2}\hat{\sigma}_{e_0}] \tag{8.6.5}$$

由式式 (8.6.5) 可知，当给定显著性水平后，预测区间的大小则由 e_0 的标准差 $\hat{\sigma_e}$ 决定，由式 (8.6.2) 可知，$\hat{\sigma}_2$ 由给定的解释变量的值及其均值之间的距离有关：X_0 越接近样本区间内解释变量 X 的均值，$\hat{\sigma}_{e_0}$ 就越小，预测的置信区间就越小；反之，X_0 越远离样本区间内解释变量 X 的均值，预测的置信区间就越大。

例 8.10. 接例8.9，求当解释变量 wt 的取值分别为 3 和 4 时，油耗的个体预测值的 95% 的置信区间。

解：利用 R 软件计算置信区间的编程如下

```
mod <- lm(mpg~wt,data=mtcars)
b0 <- coef(mod)[1]
b2 <- coef(mod)[2]
n <- nrow(mtcars)
wt<- c(3,4)
```

```
hat.mpg <- b0+b2*wt
hat.sigu <- sigma(mod)
mean.wt <- mean(mtcars$wt)
dis.sq <-(wt-mean.wt)^2
SST.wt <- (n-1)*sd(mtcars$wt)^2
sig.e <- hat.sigu*sqrt(1+1/n+dis.sq/SST.wt)
alpha <- 0.05
crt.tv <- qt(1-alpha/2,df=n-2)
l.mpg <- hat.mpg-crt.tv*sig.e
u.mpg <- hat.mpg+crt.tv*sig.e
cofd.3 <- c(hat.mpg[1],l.mpg[1],u.mpg[1])
cofd.4 <- c(hat.mpg[2],l.mpg[2],u.mpg[2])

result <- rbind(cofd.3,cofd.4)
colnames(result) <- c("fit.mpg","l.mpg","u.mpg")
result
##         fit.mpg l.mpg u.mpg
## cofd.3    21.3 14.93  27.6
## cofd.4    15.9  9.53  22.3
```

在本例中，也可以用 *predict()* 命令来直接得到个别值的预测区间，但在利用该命令前，需要将 wt 转换为 **data.frame** 形式。

```
wt <- data.frame(wt)
predict(mod,wt,interval = 'prediction')
##    fit   lwr  upr
## 1 21.3 14.93 27.6
## 2 15.9  9.53 22.3
```

2. 因变量的均值的区间估计

由前面的推导可知，$\hat{Y}_0$ 是 $E(Y_0)$ 的一个无偏估计量。令

$$\delta_0 = E(Y_0) - \hat{Y}_0$$

可以得到，随机变量 δ_0 是一个随机变量。其均值和方差分别为：

$$\begin{aligned} \delta_0 &= 0 \\ Var(\delta_0) &= \sigma_u^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \end{aligned} \tag{8.6.6}$$

利用 $\hat{\sigma}_u^2$ 去替代总体方差，可以得到其方差的估计量为：

$$\hat{\sigma^2}(\delta_0) = \hat{\sigma_u^2} \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \tag{8.6.7}$$

因此，在给定显著性水平 α 后，可得 $E(Y_0)$ 的 $1-\alpha$ 的置信区间为：

$$[\hat{Y}_0 - t_{1-\alpha/2}\hat{\sigma}_{\delta_0},\ \hat{Y}_0 + t_{1-\alpha/2}\hat{\sigma}_{\delta_0}] \tag{8.6.8}$$

例 8.11. 接例 (8.10)，求相应的预测均值的区间。

```
sig.del<- hat.sigu*sqrt(1/n+dis.sq/SST.wt)
l.mpg <- hat.mpg-crt.tv*sig.del
u.mpg <- hat.mpg+crt.tv*sig.del
cofde.3 <- c(hat.mpg[1],l.mpg[1],u.mpg[1])
cofde.4 <- c(hat.mpg[2],l.mpg[2],u.mpg[2])

result <- rbind(cofd.3,cofd.4)
```

```
colnames(result) <- c("fit.mpg","l.mpg","u.mpg")
result
##         fit.mpg l.mpg u.mpg
## cofd.3    21.3 14.93  27.6
## cofd.4    15.9  9.53  22.3
```

同样，利用 *predict()* 可以得到同样的结果

```
wt <- data.frame(wt)
predict(mod,wt,interval = 'confidence')
```

```
##    fit  lwr  upr
## 1 21.3 20.1 22.4
## 2 15.9 14.5 17.3
```

附录 A　概率论初步

A.1　事件、概率和随机变量

试验、事件和样本空间

1. 试验

定义 A.1. 试验是对一个或多个试验对象进行一次观测。

下面就是一些常见的试验：

- 掷一枚均匀的硬币，观察其出现正面或反面的情况；
- 掷一枚骰子，观察其出现的点数；
- 从一批产品中随机抽取一件，观察其实合格品还是次品；

试验的特点：

- 试验可以在相同条件下重复进行；
- 每次试验的可能结果不止一个，但在试验开始前就可以确定试验所有可能结果；
- 在试验结束之前，无法确定该次试验的确切结果。

我们将满足上面三个条件的试验称为随机试验。

2. 事件

定义 A.2. 事件指的是试验的结果。

事件有时又称为随机事件。

A.2 离散型随机变量

A.3 连续型随机变量

呐，到这里朕的书差不多写完了，但还有几句话要交待，所以开个附录，再啰嗦几句，各位客官稍安勿躁、扶稳坐好。

附录 B　抽样和抽样分布