HZ BOOKS
华章教育

“十二五”普通高等教育本科国家级规划教材

统计学

第4版

Statistics

李金昌 苏为华 编著

机械工业出版社
China Machine Press

能够从总体观、数量观和差异观出发，做到运筹帷幄、胸中有数、准确定位；在决策分析时，能够科学收集和运用数据，从容应对事物的不确定性，把风险控制在最低程度。然而这一切并不是人类天生就具有的，需要不断地学习和实践才能得到。对于学生来说，首先需要的就是系统的学习。正因为如此，国家教育部早已明确规定，“统计学”是高等学校财经类专业必须开设的核心课程之一。

要学习，就需要好的教材。为了满足高校“统计学”课程教学的需要，国内已有各种版本、各种名称的“统计学”教材上百种之多，我们也曾编写出版过数种。随着科技的发展和知识更新的加快，“统计学”教材也需要不断补充、更新和完善，出于此目的，我们重新组织编写了“统计学”教材，并作为浙江省高校重点建设教材予以出版。本教材系统地介绍了统计学的基本理论和方法。全书共分11章：总论，统计数据的收集、整理与显示，变量分布特征的描述，抽样估计，假设检验，方差分析，相关回归分析，时间数列分析，统计指数分析，统计综合评价和非参数统计方法。通过学习，学生将能较好地掌握基本统计思想和各种定量分析方法，提高分析问题的能力。本教材的特色是：内容全面、完整、有新意；体系自成，逻辑严密；深入浅出，通俗易懂，注重思想，注重应用；每章后附有小结、练习与思考；在附录中还介绍了Excel在统计学中的应用。每章都配有著名统计学家的名言和人物介绍，增加了教材的知识性、趣味性和可读性。本教材适合高等院校财经类本科各专业学生使用。

本教材是浙江省高校人文社科重点研究基地、浙江省重点学科、浙江省重点专业和浙江省精品课程——“统计学”课程建设的成果之一，由浙江工商大学李金昌教授和苏为华教授共同编写而成。编写的具体分工为：李金昌编写第1～5章和第9章，苏为华编写第6～8章、第10章和第11章。附录A“Excel在统计学中的应用”由浦国华副教授编写。在教材的编写过程中，我们参考和吸收了一些同类教材的成果，在此一并表示感谢！当然，文责自负，错误之处，敬请批评指正！

李金昌

2014年春于杭州

统计学。历史上许多有关领域的著名专家，往往也是著名的统计学家。

统计学的最主要奠基人费希尔（R. A. Fisher）曾经说过："给20世纪带来了人类进步独特方面的是统计学。"我们也相信，统计学将为人类社会的进步做出更大的贡献。

1.2 统计数据类型与研究方法

1.2.1 统计数据类型

既然统计学是关于统计数据的科学，那么统计数据有哪些类型呢？大致上可以从以下几个角度进行分类。

1.2.1.1 统计数据按照所采用的计量尺度不同，可以分为定性数据与定量数据两类

定性数据是指只能用文字或数字代码来表现事物的品质特征或属性特征的数据，具体又分为定类数据与定序数据两种。定类数据是对事物进行分类的结果，表现为类别，由定类尺度计量而成。例如，人口按照性别分为男与女两种类别，人的消费按照支出去向分为衣、食、住、行、烧、用、医、文、娱、健等类别，都属于定类数据。为了便于统计处理（计算机录入等计数处理），常用数字代码来表示各个类别，例如分别用1、0表示男性与女性，分别用1、2、3、4、5、6、7、8、9、10等表示衣、食、住、行、烧、用、医、文、娱、健等。需要注意的是，这时的数字没有任何程度上的差别或大小多少之分，只是符号而已。定序数据是对事物按照一定的排序进行分类的结果，表现为有顺序的类别，由定序尺度计量而成。例如，学生的考试成绩表示为优、良、中、及格、不及格，课题成果的鉴定等级表示为A、B、C，消费者对某产品的满意程度表示为很满意、满意、一般、不满意、很不满意，等等，都属于定序数据。同样，定序数据也可以用数字代码来表示，例如学生的考试成绩可以分别用5、4、3、2、1来表示优、良、中、及格、不及格。这时，数字代码能体现一种顺序或程度的不同，但还不能体现事物之间或不同结果之间（例如及格与不及格之间，很满意与满意之间）的具体数量差别。定序数据所包含的信息量大于定类数据。

定量数据是指用数值来表现事物数量特征的数据，具体又分为定距数据与定比数据两种。定距数据是一种不仅能反映事物所属的类别和顺序，还能反映事物类别或顺序之间数量差距的数据，由定距尺度计量而成。例如，两位学生的考试成绩分别为85分和55分，不仅说明前者良好，后者不及格，前者高于后者，而且还说明前者比后者高30分。再如，某日甲、乙、丙三地的最高气温分别为30℃、20℃和10℃，说明该日甲与乙之间最高温的温差等于乙与丙之间的温差，都是10℃。但要注意的是，定距数据一般只适合于进行加减计算而不适合于乘除运算，例如气温30℃与10℃相比，并不能说明前者的暖和程度是后者的3倍，因为气温可以是0℃或0℃以下，而0℃或0℃以下并不代表没有温度。这种情况称为不

1.2.2 统计数据研究过程

统计数据研究过程，也就是统计研究过程，包括以下四个基本环节：统计设计、数据收集、数据整理以及数据分析与解释。

1.2.2.1 统计设计

统计设计就是制定统计数据研究方案的环节，是关于以后各环节的总体安排。统计设计要在有关学科的理论指导下，根据研究问题的性质、目的和任务，科学地确定统计研究的总体对象，明确所要收集数据的种类，确定相应的统计指标及其体系并给出统一的定义和标准，确定统计数据收集、整理、推断和分析的基本方法，规定研究工作的进度安排和质量要求，拟定研究工作的资源配置和组织实施方式等。统计设计对于统计数据研究质量至关重要，要求设计者不仅要掌握系统的统计学理论和方法，而且要具有所研究领域的有关知识和理论素养。

1.2.2.2 数据收集

数据收集就是按照统计设计的要求，有针对地获取所需的统计数据的环节，也称为统计调查环节。也就是说，要通过统计观测或实验的方式、方法去收集各种各类计算统计指标所需的原始数据，以及其他已经存在的各种相关数据。数据收集是否准确、及时、完整，直接影响到统计分析的质量。

1.2.2.3 数据整理

数据整理就是对通过统计观测或实验所获得的原始数据，进行必要的系统化处理，使之条理化、综合化，成为能反映总体特征的统计数据的环节，也称为统计整理环节。数据整理也包括对已有数据的再加工和深加工。数据整理的手段有统计分组、汇总和计算等，整理结果表现为统计图、统计表或统计指标。

1.2.2.4 数据分析与解释

数据分析是在数据整理的基础上，围绕统计设计所确定的研究任务，运用各种统计方法对数据进行各种统计分析，得出某些有用的定量结论的环节，也称为统计分析环节。数据分析实质上就是对数据的深加工整理，是整个统计研究的核心，也是统计研究的最终目的。在这个环节，既要用到描述统计方法，又要用到推断统计方法。数据解释则是对整理和分析的数据或有关数量结果进行说明，即说明为什么会得出这些数据，这些数据的含义分别是什么，从中能得出哪些具有规律性的结论，需要进一步探讨哪些问题，等等。数据解释是对数据分析的深化。

1.2.3 统计数据研究方法

统计数据研究的基本方法有大量观察法、统计分组法、综合指标法、统计推断法和统

使用，又能在许多场合更深入地揭示出事物的本质。例如，若考查某一批产品的质量，则研究的是具体总体，结果是表明这一批产品的质量高低；而若考查某种工艺条件下的该种产品的质量，则研究的是抽象总体，结果不仅能表明产品本身的质量，而且更重要的是还可以说明这种工艺条件的性能及先进性。

(3) 总体按照其个体能否计数可以分为可计数总体和不可计数总体两类。可计数总体是指能对其所包括的个体进行计数且计数结果能加总的总体，例如人口总体，每个个体是可计数的，而计数的结果即人数是可相加的；工业企业总体、某批产品总体等也都是如此。可计数总体的特征是，它所包含的个体具有相同的计量单位，可以计算总体单位总数。不可计数总体是指对其所包括的个体不可计数或计数结果不能加总的总体，例如零售商品总体，虽然每件商品都具有商品的共性，但由于各自的使用价值形态和计量单位不同，所以在商品的件数上是不能直接相加的。然而，零售商品总体的物价水平和销售数量的总变动情况却是统计研究的内容之一。不可计数总体的特征是，它所包含的个体通常不具有相同的计量单位，不能计算总体单位总数。

(4) 总体按照其个体是否人为划定可以分为自然总体和人为总体两类。自然总体是由自然确定的个体所组成的，即个体是明确的、易定的，例如人口总体中的个人、企业总体中的企业、家庭总体中的家庭等都是自然个体。人为总体是由人为确定的个体所组成的，其个体往往不明显或难以确定，例如在考查某种小麦的出粉率时，总体是全部该种小麦，但个体显然不能是每一粒小麦，那么该以1千克小麦还是以100千克小麦或1吨小麦作为一个个体，并没有明确的规定。再如，要研究林区的木材储藏量，也不能以每一棵树作为个体，但应该以多大面积的区域作为一个个体也没有明确的规定。对于个体不明显的现象，要根据研究对象的具体情况和研究目的的不同恰当地加以确定。

在实践中，还经常需要对被研究的总体进行分组或分类研究，尤其是要对总体中的某特定组或特定类进行分析研究。这时，总体中的一个组或类，就被称为一个研究域或一个子总体。例如在研究消费者的购买力时，对某特定类型的消费者群（例如老年消费者群、儿童消费者群、学生消费者群、妇女消费者群等）进行特别的研究，就是对消费者总体中的一个子总体进行研究，这在市场营销学中被称为市场细分研究。子总体具有和原总体同样的性质。

3. 总体与个体的关系

总体与个体的关系不是一成不变的，其可变性体现为三方面。一是总体容量随着个体数的增减可变大或变小。二是随着研究目的不同，总体中的个体可发生变化。例如要研究某市的居民身体素质，则总体是该市所有人口，若要研究该市的居民家庭生活水平和消费结构，则总体是该市所有居民家庭。三是随着研究范围的变化，总体与个体的角色可以变换。例如，在研究某地区某校学生的学习状况时，总体由该校所有学生构成，即学校是总体。而若要研究该地区所有学生的构成状况，则总体由该地区的所有学校构成，学校则成为个体。这说明，个体与总体要根据研究目的和对象范围而定。

个。再如，“您夏天喜欢喝什么饮料？①开水，②矿泉水，③纯净水，④可乐，⑤雪碧，⑥芬达，⑦果汁，⑧其他____”，备选答案有八个，由被调查者从中选择一个或多个。多项式设计的回答和统计处理都比较容易，但要列出所有可能的备选答案往往有一定困难（不能太多），故常用“其他____”来处理。

(3) 顺位式。顺位式要求被调查者对问题的备选答案，按照重要性程度或喜爱程度定出先后顺序，做出比较性的回答。例如，“请您对下列不正之风按您痛恨的程度以1，2，3…的顺序加以排列：□用公款大吃大喝，□用公款送礼，□拉关系走后门，□用公款旅游，□用公款购买小车和手机等，□任人唯亲，□领导干部官僚主义、脱离群众，□滥发文凭，□拉帮结派，□其他_____”等。这种设计便于被调者去衡量比较，能比多项式了解更多的信息，适用于要求区分答案的缓急轻重或先后顺序的问题。但它难以体现答案之间的差异大小，并且当备选答案较多时，各答案在问卷中的位置也会对被调查者产生一定影响。

(4) 程度评价式。这是一种观念计量的方法，所得结果即为定类数据。一般地，对问题列出几个不同程度的答案，并对每一个答案事先按顺序给分，相邻答案的分差相等，由被调查者从中选择一个答案来表达他对事物的感受程度。例如：

您对您目前从事的职业有多满意？

	很满意	满意	一般	不满意	很不满意
	2	1	0	−1	−2
或	5	4	3	2	1

这种设计能从计分的角度进行统计处理，有利于综合了解被调查者的总体态度和程度。但计分本身是非客观的，只是一种人为规定。有时，也可以把答案按程度分为3档、7档或9档，档数越多，了解的信息就越细，但相邻答案之间的区别就越微小。

(5) 比较式。比较式指把若干可比较的事物整理成两两对比的形式，由应答者进行比较。这种方式比将许多事物放在一起，让应答者做比较要简便容易一些，并可获得针对性明显的具体结果。例如：

请您比较下列每一对不同的广告，哪一种更吸引人？

① □甲广告和□乙广告　② □丙广告和□丁广告

③ □甲广告和□丁广告　④ □乙广告和□丙广告

⑤ □甲广告和□丙广告　⑥ □乙广告和□丁广告

此外，问题答案还有过滤式、倾向偏差式、竞争选好式、回想式等形式。

2. 问题答案的设计原则

(1) 所列答案应包括所有可能的回答。只有将全部可能的答案列出，才能使每个应答者都有答案可选，不至于无合适答案而放弃回答。为防止答案遗漏，可用“其他_____”来弥补。

(2) 不同答案之间不能相互包含。一个问题所列出的各个答案必须互不相容，互不重叠，否则应答者可能做出有重复内容的双重选择，影响调查效果。例如，“您喜欢阅读哪类图书？①文学艺术类，②自然科学类，③社会科学类，④经济管理类，⑤会计类，⑥统计类”这一设计中，有关答案之间就相互包容了，因为会计类属于经济管理类或社会科学类，因

序排列在相应的表格内，就形成统计表。广义的统计表还包括统计调查表和统计分析表。统计表具有简明扼要、一目了然的特点，可以清楚地显示统计数据，直观地反映统计分布特征和各部分之间的关系，便于进行对比、计算并开展统计分析，便于保存统计数据。

2.3.1.2　统计表的结构

统计表的结构可以从表式和内容两个方面来认识。

从表式上看，统计表是由纵横交错的线条所构成的一种表格，包括总标题、横行标题、纵栏标题和指标数值四个部分。总标题是统计表的名称，概括地说明统计表的内容（包括时间和空间），写在统计表的正上面。横行标题是表示横行内容的名称，是所要说明的对象，可以是总体、个体，也可以是组，或者是时间，一般写在表的左方。纵栏标题是纵栏的名称，即用以说明横行标题的指标名称，一般写在表的右上方。指标数值列在横行标题与纵栏标题的交叉处，是用以表明横行标题数量特征的具体数值，列在表的右下方，是统计表的核心部分。具体表式如表2-7所示。

表2-7　2010年我国三次产业增加值及增长情况表　← 总标题

产　业	增加值（亿元）	占国内生产总值的比重（%）	比2004年增长（%）
第一产业	40 497	10.20	4.30
第二产业	186 481	46.80	12.20
第三产业	171 005	43.00	9.50
合计	397 983	100.00	10.30

（横行标题：产业；纵栏标题：增加值（亿元）、占国内生产总值的比重（%）、比2004年增长（%）；指标数值：各数值；主词：横行标题；宾词：纵栏标题和指标数值）

资料来源：国家统计局，中华人民共和国2010年国民经济和社会发展统计公报，www.stats.gov.cn。

从内容上看，统计表由两部分组成：主词和宾词。主词是统计表所要说明的总体、个体或者组的名称，一般列于表的左方，即横行位置。宾词是用以说明总体及其组成部分数量特征的各种统计指标，一般列于表的右方，即纵栏标题和指标数值的位置。有时，主词与宾词的位置可以互换。

此外，有些统计表还有补充资料、资料来源、注释、填表单位、填表人和填表日期等内容。

2.3.1.3　统计表的种类

统计表按照主词是否分组，以及分组标志多少，可以分为未分组表、简单分组表和复合分组表三种。

未分组表是指主词未进行任何分组的统计表，即主词只按一定顺序罗列总体中每个个体的名称，或者将主词按时间顺序简单排列。

简单分组表是指主词按一个标志分组的统计表，表2-4、表2-5和表2-7都是简单分组表。

复合分组表是指主词按两个或两个以上标志分组的统计表，可以通过多个标志的结合

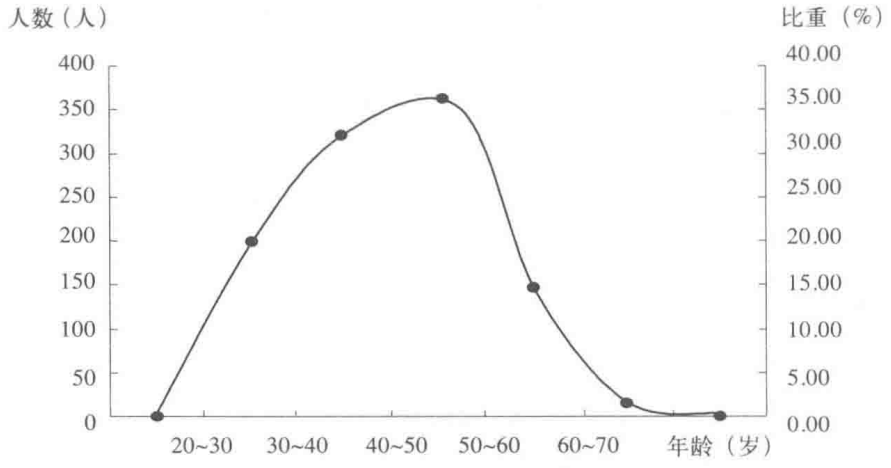
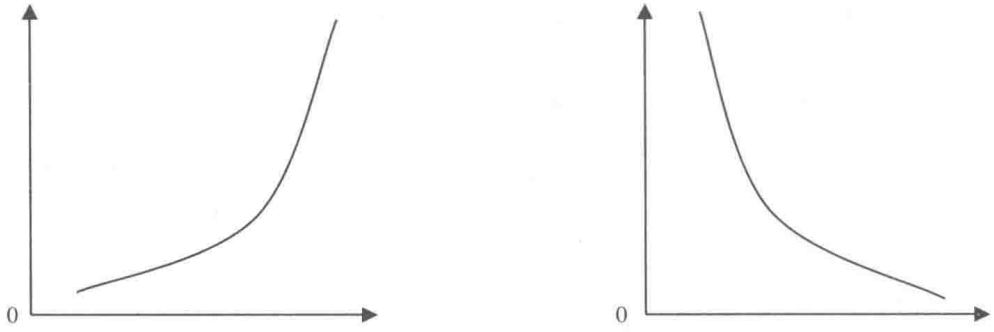


图2-4 某高校教师年龄分布曲线图

变量分布曲线图种类很多，常见的有J形分布、U形分布和钟形分布三种。

J形分布有两种类型：一种是变量分布的频数或频率随变量值的增大而变大，称为正J形分布，例如商品供应量随着价格的上升而增加。另一种是变量分布的频数或频率随变量值的增大而变小，称为反J形分布，例如商品需求量随着价格的上升而下降。J形分布曲线如图2-5a、图2-5b所示。



a) 正J形分布　　b) 反J形分布

图2-5 J形分布

U形分布是一种“两头大，中间小”的分布，即靠近中间变量值的分布频数小、频率低，靠近两端变量值的分布频数大、频率高，曲线形式犹如英文字母“U”。例如人口死亡率的年龄分布就是幼儿和老年人死亡率高，青少年和中年的死亡率低，如图2-6所示。

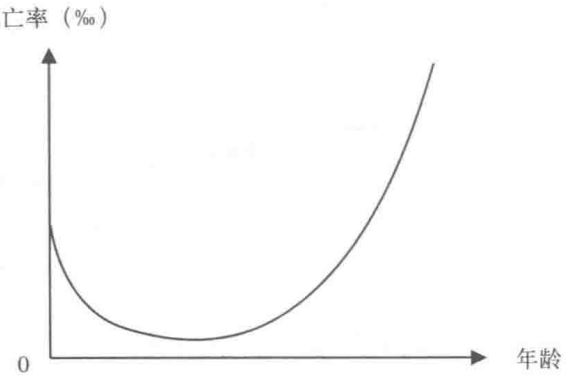


图2-6 U形分布

钟形分布与U形分布正好相反，是一种“中间大，两头小”的分布，即靠近中间变

A. 普查　　B. 重点调查

C. 抽样调查　　D. 科学推算

2. 调查小学男生的身高，则身高是（　）。

A. 观测标志　　B. 观测单位　　C. 调查对象　　D. 变量值

3. 抽样调查中不可避免的误差是（　）。

A. 系统性误差　　B. 偶然性误差　　C. 观测性误差　　D. 登记性误差

4. 在组距式数列中，对组限值的处理原则是（　）。

A. 上组限不在内、下组限在内　　B. 下组限不在内、上组限在内

C. 上下组限均不在内　　D. 上下组限均在内

5. 最常见的变量分布类型是（　）。

A. 正J形分布　　B. U形分布

C. 钟形分布　　D. 反J形分布

三、简答题

1. 如何设计统计数据收集方案？试举例说明。

2. 概率抽样与非概率抽样有什么本质区别？试举例说明。

3. 分层抽样与整群抽样有什么区别？试举例说明。

4. 什么是重点调查？有什么特点？

5. 什么是统计数据收集的实验方式？应遵循哪些原则？

6. 常用的实验设计有哪些？试分别举例说明。

7. 在统计数据收集过程中，可能存在哪些误差？试分别举例说明。

8. 什么是问卷？该如何设计问卷的问题和答案？

9. 统计数据整理有哪些基本步骤？

10. 如何理解统计分组的含义与性质？

11. 试举例说明J形分布、U形分布和钟形分布。

四、计算题

根据书中例2-6关于55名工人日加工零件数资料，要求：

(1) 编制频数分布数列和频率分布数列；

(2) 编制向上、向下累计频数分布数列和累计频率分布数列；

(3) 绘制直方图、折线图、曲线图、箱形图和累计分布曲线图（可利用Excel）；

(4) 说明工人日加工零件数的分布特征。

五、实践题

请同学们组成5人小组，自行确定调查主题，设计问卷，并进行实际调查（有效问卷50份以上），利用Excel进行问卷数据处理，编制必要的统计表并绘制必要的统计图，写出简单的调查报告。

第3章

变量分布特征的描述

“统计学具有处理复杂问题的非凡能力，当科学的探索者在前进的过程中荆棘载途时，唯有统计学可以帮助他们打开一条通道。”

“很难理解为什么统计学家通常限制自己的调查于平均数，而不着迷于更广泛的考虑。对于变化的魅力，他们的灵魂看来如同平坦的英格兰国家之一的当地人的一样迟钝，那些当地人关于瑞士的回顾是，如果可以将它的山脉扔进它的湖泊，那么两种讨厌的东西将立即去除。”

——弗朗西斯·高尔顿

本章介绍如何对变量分布的特征进行描述，内容包括集中趋势与平均指标、离中趋势与离散指标、分布形状与形状指标三大方面。本章内容对于以后各章的学习非常重要，具体要求：①理解变量分布三大特征，即集中趋势、离中趋势和分布形状的含义；②理解平均指标、离散指标和形状指标的意义与作用；③熟练掌握各种平均数的计算方法并加以正确的应用，科学理解加权平均数中权数的意义，正确认识算术平均数与调和平均数之间的应用关系，以及算术平均数、中位数和众数三者之间的数量关系；④熟练掌握各种离散指标的计算方法并加以正确的应用，尤其是要深刻理解方差、标准差和离散系数的内涵；⑤熟练掌握偏度系数和峰度系数的计算方法并加以正确的应用，尤其是要了解动差的含义。

3.1 集中趋势的描述

变量分布特征可以从以下三个方面加以描述：一是变量分布的集中趋势，反映变量分布中各变量值向中心值靠拢或聚集的程度；二是变量分布的离中趋势，反映变量分布中各

布的特征，我们不仅要观察其集中趋势和离中趋势，也要观察其形状。

变量分布的形状要用形状指标来反映。形状指标就是反映变量分布具体形状，即左右是否对称、偏斜程度与陡峭程度如何的指标。具体来说，变量分布的形状一般从对称性和陡峭性两方面来反映，因此形状指标也有两个方面：一是反映变量分布偏斜程度的指标，称为偏度系数；二是反映变量分布陡峭程度的指标，称为峰度系数。

偏度系数可以告诉我们变量分布是左偏还是右偏，即受低端变量值的影响大还是受高端变量值的影响大。而峰度系数则可以告诉我们分布是尖陡还是扁平，即频数（频率）分布绝大部分集中于众数附近还是各变值的频数（频率）相差不大（如果各变量值的频数或频率相等，则分布呈一条直线，无峰顶可言）。由此可见，形状指标与平均指标、离散指标一样，都是变量分布特征的重要体现。

3.3.2　偏度系数

偏度的概念首先由统计学家皮尔逊（Pearson）于1895年提出，是对变量分布对称性的测度，是指变量分布偏斜的方向及其程度。在本章3.1节论述算术平均数、中位数和众数三者的关系时，曾经涉及这个问题。图3-3表示变量分布对称无偏，图3-4表示变量分布向右偏斜（即右偏或正偏），图3-5表示变量分布向左偏斜（即左偏或负偏）。

偏度的测定是通过计算偏度系数来实现的，偏度系数通常用S_k来表示。偏度系数的计算主要有三种方法。

3.3.2.1　利用算术平均数与众数或中位数的离差求偏度系数

前面已提到，如果算术平均数、众数与中位数三者相等，则变量分布无偏；如果三者不相等，则变量分布有偏，而且三者之间的差距越大变量分布的偏度也越大。因此，我们可以利用算术平均数与众数或中位数的离差求偏度系数并标记为 $S_k^{(1)}$，计算公式为

$$S_k^{(1)} = \frac{\bar{x} - m_o}{s} \tag{3-34}$$

将 $\bar{x} - m_o$除以标准差s，一是为了消除不同计量单位的影响，二是为了把不可直接比较的绝对数转化为可相互比较的相对数。

一般情况下，偏度系数 $S_k^{(1)}$的变动范围为（−3，3）。当 $\bar{x} > m_o$时，$S_k^{(1)}$为正值，变量分布属于正偏；当 $\bar{x} < m_o$时，$S_k^{(1)}$为负值，变量分布属于负偏；当 $\bar{x} = m_o$时，$S_k^{(1)}$为零，变量分布属于无偏（即对称分布）。$S_k^{(1)}$的绝对值越接近于3，表明变量分布的偏斜程度越严重；$S_k^{(1)}$的绝对值越接近于零，表明变量分布的偏斜程度越轻微。

3.3.2.2　利用四分位数求偏度系数

根据四分位数的特点可知，如果变量分布对称、无偏斜，那么第一个四分位数Q_L与第三个四分位数Q_U是关于中位数对称分布的，即$Q_U - m_e = m_e - Q_L$，因此我们可以通过$Q_U - m_e = m_e - Q_L$这个等式是否成立来判断变量分布是否对称，并且可以根据第一个、第三个四

4. 在实际应用中，调和平均数与算术平均数有什么联系？
5. 从数学上看，算术平均数、几何平均数和调和平均数三者有什么关系？
6. 什么是中位数？有什么特点？试举例说明其应用。
7. 什么是众数？有什么特点？试举例说明其应用。
8. 算术平均数、中位数和众数三者的数量关系说明什么样的变量分布特征？
9. 什么是离散指标？有什么作用？常用的离散指标有哪些？
10. 什么是方差和标准差？有哪些性质？
11. 如何反映变量分布的形状？

四、计算题

1. 某司机开车从A地到B地的时速是100公里，从B地返回A地的时速是120公里，问平均时速是多少？
2. 菜场上某鱼摊大鲫鱼每条约重0.4千克，售价为每千克20元，小鲫鱼每条约重0.25千克，售价为每千克12元。某顾客向摊主提出大、小鲫鱼各买一条，一起称重，价格为每千克16元。摊主应允，问这次买卖谁占了便宜？为什么？
3. 某公司下属27家企业的资金利润率分组数据和各组年利润额数据如下表所示：

按资金利润率分组（%）	企业数	年利润额（万元）
8以下	2	300
8～12	6	1 000
12～16	12	2 600
16～20	5	1 200
20以上	2	400
合计	27	5 500

请计算：

(1) 平均每个企业的利润额。

(2) 全公司的平均资金利润率（分别用绝对数权数和相对数权数）。

4. 某年某企业3个车间的产品生产情况如下表所示：

车间	合格率（%）	合格品产量（辆）	年生产工时数（小时）
A	98	19 600	6 800
B	95	18 620	7 200
C	99	18 434	8 000
合计		56 654	22 000

问：

(1) 若3个车间依次完成整个产品某一工序的加工装配任务，全厂总的合格率、平均合格率和平均废品率分别是多少？

(2) 若3个车间分别独自完成整个产品的生产加工过程，则全厂总的合格率、平均合格率

业对他并无吸引力。22岁那年他获得一笔可观的遗产，决定弃医。1850～1852年，他与友人远赴非洲进行科学考察，1853年被选为英国皇家地理学会会员，1856年又被选为英国皇家学会会员。高尔顿研究涉猎范围包括地理、天文、气象、物理、机械、人类学、民族学、社会学、统计学、教育学、医学、生理学、心理学、遗传学、优生学、指纹学、照相术、登山术、音乐、美术、宗教等，是一位百科全书式的学者。主要著作有《气象测量》《遗传的天才》《自然的遗传》《指纹》等15部，撰写各种学术论文220篇。高尔顿主张“无论何时，能算就算”，对统计学的最大贡献是相关性概念的提出和回归分析方法的建立。高尔顿的生物统计学思想经过他的学生皮尔逊、韦尔登的参与和发挥，在英国形成了一个颇有影响的生物统计学派。1901年，高尔顿、皮尔逊、韦尔登创办《生物统计》杂志，成为生物统计学派的一面旗帜。1909年，弗朗西斯·高尔顿被英国王室授予勋爵称号。

4.1.1.1 总体分布及其特征

总体分布就是总体中所有个体关于某个变量（标志）的取值所形成的分布。假设X为总体随机变量，那么总体分布就是指X的分布。很显然，同一变量不同的总体或同一总体不同的变量，其分布是不同的。第3章已经谈到，变量分布的形态很多，例如J型分布、U型分布和钟形分布等，不同的分布会有不同的特征，认识总体分布特征是统计研究的任务之一。

反映总体分布特征的指标叫做总体参数，一般用θ来表示。在抽样实践中，常用的总体参数有两个：一是总体均值（包括是非变量的均值）；二是总体方差或标准差（包括是非变量的方差或标准差）。

假设有限总体的容量为N，第i个个体的变量值为X_i（$i = 1, 2, 3, \ldots, N$），均值为$\overline{X}$，方差为S^2，那么就有

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{4-1}$$

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N-1} \tag{4-2}$$

特殊地对于是非变量，如果两类变量值个数分别为N_1和N_0（$N_1 + N_0 = N$），N_1个变量值为1，N_0个变量值为0，并且令$P = \frac{N_1}{N}$，$Q = \frac{N_0}{N}$，那么如果以$\overline{X}_P$表示总体均值，以S_P^2表示总体方差，就有

$$\overline{X}_P = P \tag{4-3}$$

$$S_P^2 = \frac{N}{N-1} PQ \tag{4-4}$$

显然，$P + Q = 1$。这时，$\overline{X}_P$也称为总体比例或总体成数。

从理论上看，总体参数θ的值是唯一确定的，是根据总体中所有个体的变量值计算而得的。然而，我们不可能经常对总体进行全面观测调查，以获取所有个体的变量值数据，所以总体参数θ的值通常都是未知的，正因为如此才需要通过样本观测结果来加以估计。

4.1.1.2 样本分布及其特征

样本分布就是样本中所有个体关于某个变量（标志）的取值所形成的分布。假设x为总体随机变量X在样本中的体现，那么样本分布就是指x的分布，或者说是关于n个观测值的分布。同样，同一变量不同的样本或同一样本不同的变量，其分布是不同的。由于样本来自于总体，包含了一部分关于总体的信息，所以样本分布是一种经验分布。当样本容量n很大，或是当n逐渐增大时，样本分布会接近总体分布。如果样本容量很小，那么样本分布就有可能与总体分布相差很大，抽样估计的结果就会很差。所以，如何抽样、应该有多大的样本容量才能使样本分布充分接近总体分布，这是抽样中很重要的问题。

反映样本分布特征的指标叫样本统计量，通常用T来表示。与总体参数相对应，常见

3. 抽样分布特征

任一抽样分布都有自己的特征，这个特征就是样本统计量的数学期望和方差。其中，样本统计量的数学期望就是所有样本统计值的平均数，样本统计量的方差就是所有样本统计值关于数学期望的方差。当估计量就是样本统计量时，数学期望与方差分别表示为 $E(\hat{\theta}) = \sum \hat{\theta}_i \pi_i$ 和 $V(\hat{\theta}) = \sum \left[\hat{\theta}_i - E(\hat{\theta})\right]^2 \pi_i$ 。

在简单随机抽样下，样本均值的数学期望为总体均值即 $E(\bar{x}) = \sum \bar{x}_i \pi_i = \bar{X}$，样本均值的方差为 $V(\bar{x}) = \sum (\bar{x}_i - \bar{X})^2 \pi_i$ 。在例4-1中，不论是重复抽样还是不重复抽样，样本均值都等于6，但重复抽样与不重复抽样的方差则有不同的结果，重复抽样下的方差为 $V(\bar{x}) = \frac{8}{3}$，不重复抽样下的方差为 $V(\bar{x}) = \frac{4}{3}$。同理，在简单随机抽样下，样本成数的数学期望为总体成数，即 $E(p) = \sum p_i \pi_i = P$，样本成数的方差为 $V(p) = \sum (p_i - P)^2 \pi_i$ 。在例4-2中，不论是重复抽样还是不重复抽样，样本成数的均值都是0.6，但重复抽样与不重复抽样的方差也有不同的结果，重复抽样下的方差为 $V(p) = 0.06$，不重复抽样下的方差为 $V(p) = 0.04$。

根据第3章关于离散指标的含义可知，在均值相同的情况下，方差不同就代表分布的离散程度不同，即方差越小（大）抽样分布的离散程度越弱（强）或抽样分布的集中趋势越强（弱）。由于在各种抽样方法和抽样组织形式下，样本统计量的数学期望等于总体参数（例如样本均值的数学期望等于总体均值、样本成数的数学期望等于总体成数）这个性质基本都能得到满足（无偏性），因而抽样分布的特征主要是通过抽样分布的方差来体现的。很显然，由于抽样估计是以所抽取样本所提供的特征（样本统计值）为依据，因而抽样分布越集中、样本统计量分布的方差越小，则所抽取样本的统计值就越可能接近总体参数，抽样估计的误差就越小，抽样估计的结果就越精确。因此，如何在遵循随机原则和节省费用的前提下，设计出抽样分布方差最小的抽样方案，始终是我们追求的目标。当然，样本统计量无偏并不等于抽样分布无偏，抽样分布的偏差性需要用偏度系数，例如样本均值分布的偏差要用 $\sum_{i=1}^{m} \left[\bar{x}_i - E(\bar{x})\right]^3 \Big/ \left[\sqrt{V(\bar{x})}\right]^3$ 来反映。

需要说明的是，我们每次抽样一般只能抽取一个样本，所得到的样本统计值只是m个可能值中的一个，不可能按上述形式列出样本均值或样本成数的实际抽样分布，因此也不可能按前述的公式来计算抽样分布的期望和方差。但是，我们对样本统计量抽样分布的理解，能帮助我们掌握样本统计量分布的规律和样本统计量与总体参数之间的内在联系，从而使我们由样本去估计总体有据可循。

4.1.2 常用的抽样分布定理

4.1.2.1 样本均值的抽样分布定理

1. 正态分布的再生定理

如果某样本的n个个体完全随机地来自数学期望为 $\bar{X}$、方差为S^2的正态总体，则不论样

本容量n多大，样本均值$\bar{x}$服从数学期望为$\overline{X}$、方差为$V(\bar{x})=\frac{S^2}{n}$（重复抽样时）或$V(\bar{x})=\frac{(N-n)S^2}{N_n}$（有限总体且不重复抽样时）的正态分布。标准统计量$z=\frac{\bar{x}-\overline{X}}{\sqrt{V(\bar{x})}}$则服从数学期望为0、方差为1的标准正态分布。这就是正态分布的再生定理。

2. 中心极限定理

对于任一具有平均数$\overline{X}$和方差S^2的有限总体，当样本容量n足够大时（例如$n>30$或$n>50$），样本均值$\bar{x}$的分布也趋于服从正态分布，其数学期望和方差与再生定理的相同。这就是中心极限定理。

3. t分布定理

当正态总体的方差未知且n较小，或任一方差为S^2的总体但n较小，则样本均值$\bar{x}$的分布服从自由度为$n-1$的t分布。t分布曲线与正态分布相近，其中数学期望相同。

4.1.2.2 样本成数的抽样分布定理

1. 二项分布定理

从一个数学期望为P、方差为$\frac{N}{N-1}PQ$的是非变量（0-1分布）总体中随机重复地抽取容量为n的样本，那么样本中含有n_1个某类变量值的概率为

$$\pi(n_1)=C_n^{n_1}P^{n_1}Q^{n-n_1} \tag{4-9}$$

其中$n_1=0,1,2,3,\cdots,n$；$\sum\pi(n_1)=1$。

对于特定的n和P，可以求出$n_1=0$至$n_1=n$的所有概率，也就是可以求出$p=0$至$p=1$的所有概率，从而形成一个分布，这个分布就是二项分布。当$P=0.5$时，二项分布是对称的；当$P\neq0.5$时，二项分布是不对称的。

2. 超几何分布定理

从一个数学期望为P、方差为$\frac{N}{N-1}PQ$的是非变量（0-1分布）总体中随机不重复地抽取容量为n的样本，那么当$N_1\geqslant n$同时$N_0\geqslant n$时，样本中含有n_1个某类变量值的概率为

$$\pi(n_1,n_0|N_1,N_0)=\frac{C_{N_1}^{n_1}C_{N_0}^{n_0}}{C_N^n} \tag{4-10}$$

其中$n_1=0,1,2,3,\cdots,n$；$\pi(n_1,n_0|N_1,N_0)=1$。

对于给定的n和P，可以求出$n_1=0$至$n_1=n$的所有概率，也就是可以求出$p=0$至$p=1$的所有概率，从而形成一个分布，这个分布就是超几何分布。当N无限增大时，超几何分布趋向于二项分布。

3. 中心极限定理

从任一数学期望为P、方差为$\frac{N}{N-1}PQ$的是非变量（0-1分布）总体中随机抽取容量足够

抽样极限误差实际上就是对估计量可允许取的最高值或最低值进行了限制，因为每一次抽样都有一定的精度要求。如果抽样极限误差过大，即所允许的估计值过高或过低，那么抽样估计的结果就可能毫无意义。例如，某些社会经济指标平均每年能递增5%就算不错了，如果抽样极限误差比5%都要大，则抽样估计的价值就难以体现了。

那么抽样极限误差Δ该如何确定呢？它取决于两个因素。一是抽样标准误，即抽样分布本身具有多大的标准差。如果说抽样标准误是一把衡量抽样误差大小的尺子，那么抽样极限误差就是以该尺子来衡量的一个长度。在其他条件既定时，抽样标准误越大（小），抽样极限误差也就越大（小）。二是抽样估计概率保证程度，也称为置信水平，通常表示为$1-\alpha$，其中α就是显著性水平。以样本估计总体，除有精度要求外，还有可靠度要求，即以多大的概率来保证估计是准确的。根据抽样分布曲线可知（见图4-2），抽样分布曲线与估计量坐标轴之间的极限面积为1，或者说抽样分布曲线涵盖所有可能估计值的概率为100%。在抽样分布标准差（抽样标准误）既定时，所要求的概率保证程度越高（低），被曲线覆盖的可能最高估计值或最低估计值就越远离抽样分布的中心位置（估计量的期望值），抽样极限误差也就越大（小）。

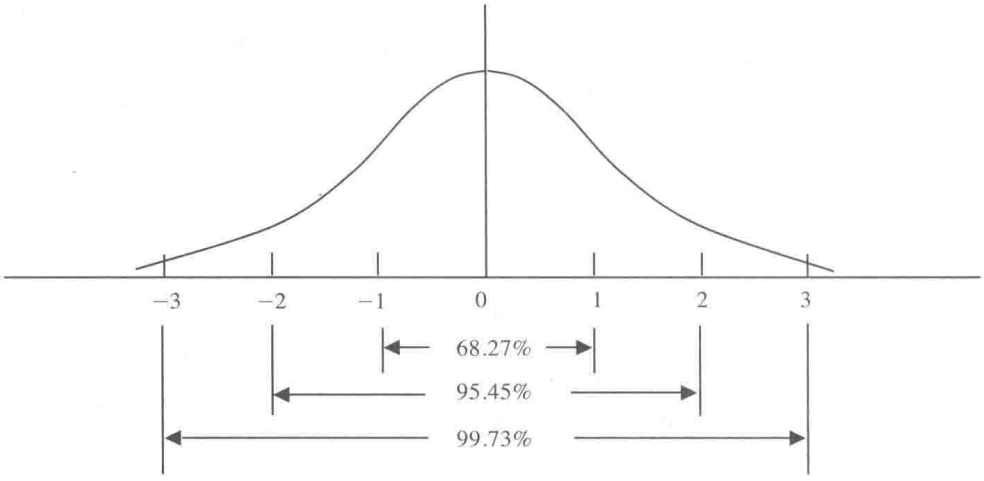


图4-2 标准正态分布临界值与置信水平

为了把抽样极限误差、抽样标准误和抽样概率保证程度三者关系更清楚地表达出来，我们把抽样极限误差与抽样标准误之比的系数称为抽样概率度。在正态分布下，抽样概率度用$z_{\alpha/2}$来表示，即

$$\Delta = z_{\alpha/2} SE(\hat{\theta}) \tag{4-11}$$

或

$$z_{\alpha/2} = \frac{\Delta}{SE(\hat{\theta})} \tag{4-12}$$

不难发现，Δ分别与$z_{\alpha/2}$、$SE(\hat{\theta})$成正比，而$z_{\alpha/2}$与$SE(\hat{\theta})$成反比。因此，在一定的概率保证下，要想提高抽样估计精度，就必须缩小抽样极限误差，就必须通过抽样设计来降低抽样标准误。