Statistics

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

统计学

贾俊平　编著

中国人民大学出版社

Statistics

Statistics and Actuarial Science

中国人民大学统计与精算系列教材

统计学

贾俊平　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计学/贾俊平编著. —北京：中国人民大学出版社，2015.6
中国人民大学统计与精算系列教材
ISBN 978-7-300-21401-6

Ⅰ.①统… Ⅱ.①贾… Ⅲ.①统计学-高等学校-教材 Ⅳ.①C8

中国版本图书馆 CIP 数据核字（2015）第 113959 号

中国人民大学统计与精算系列教材
统计学
贾俊平 编著
Tongjixue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511770（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京密兴印刷有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2015 年 7 月第 1 版
印　　张	16.5 插页 1	**印　　次**	2015 年 7 月第 1 次印刷
字　　数	412 000	**定　　价**	35.00 元

版权所有　侵权必究　　印装差错　负责调换

前　言

本书概要

本书是为大学本科生编写的一本统计学基础教材，可供一个学期（约 3 学分）的教学使用。全书内容共 12 章，第 2 章和第 3 章介绍数据的描述性分析方法，包括图表的使用和常用统计量的计算与分析方法。第 4～6 章介绍统计推断的基本原理和方法，包括作为推断理论基础的概率分布及参数估计和假设检验。第 7～11 章介绍实际中常用的一些统计方法，包括类别变量分析、方差分析、回归分析和时间序列预测等。第 12 章介绍几种常用的非参数检验方法。

本书特色

本书体现了作者的一贯写作风格，表述简明，强调应用。具体有以下几个特点：

➢ **强调统计软件的使用**。本书所有例题的计算和分析完全使用统计软件来实现。书中使用了 SPSS（19.0 中文版）和 R（3.1.2 版本）两种软件。其中以 SPSS 为主，SPSS 能够实现的分析均给出 SPSS 实现结果，对于少数 SPSS 不能实现的部分内容给出了 R 的输出结果，并在每章的最后给出了 SPSS 的详细操作步骤和 R 的实现程序。同时，在本书最后的附录里给出了两种软件的使用简介。

➢ **注重统计思想和方法应用**。本书完全避免统计方法的数学推导，并把繁杂的计算交给统计软件来完成，让读者拿出更多精力去理解统计方法的思想和原理。读完本书后你就会发现统计学并不那么难学，而且比你想象的有趣、有用。

➢ **体现统计方法之间的逻辑**。在每章开头均给出了本章的学习目标，可快速浏览本章的学习要点。在第 1 章最后以图解方式给出了本书的逻辑框架，其他章的最后均给出了本章的内容框架，以便于读者把握各章节内容的内在联系。

➢ **方便教学和学习**。每章都配有详细的 PPT，并配备书中例题和练习题的电子版数据，方便教师教学和学生学习。

读者对象

本书适用的读者包括：高等院校统计学专业的本科生；经济管理类专业本科生；其他文科专业及部分理、工、农、林、医、药专业的本科生；具备少量数学知识的其他读者。

致谢

本书是中国人民大学 985 经费资助项目，特此感谢。

贾俊平

中国人民大学统计学院

目　录

试读结束：需要全本请在线购买：　www.ertongbook.com

第1章　数据与统计学

学习目标

学完本章内容应该达到以下目标：

1. 理解统计学的含义。
2. 理解描述统计和推断统计的含义。
3. 了解统计学的应用。
4. 掌握变量和数据的分类。
5. 了解数据的来源。
6. 理解几种不同概率抽样方法的含义。
7. 能使用软件从总体中抽取简单随机样本。

在日常生活中，经常会接触到统计数据或一些统计研究结果。比如，在电视、报纸、网络等各种媒体中就会经常看见一些报道使用统计数据、图表等。作为一门科学的统计学研究什么呢？怎样获得所需要的统计数据呢？这就是本章将要介绍的内容。

1.1　统计学及其应用

每个人都离不开统计，了解一些统计学知识对每个人来说都是必要的。比如，在外出旅游时，你需要关心一段时间内的详细天气预报；在投资股票时，你需要了解股票市场价格的信息，了解某只特定股票的有关财务信息；在观看足球比赛时，除了关心进球数的多少，你还要知道各支球队的多项技术统计，等等。要看懂数据似乎并不困难，但要正确理解统计数据或统计的某些结论，就需要具备一些统计学知识了。

1.1.1　什么是统计学

在日常工作或管理中，总会面对各种各样的数据。如果不去分析，那它也仅仅是一堆数据，没有太多价值。如何分析这些数据，用什么方法分析数据，并从分析中得出某些结论以帮助我们做出决策，这正是统计学要解决的问题。简言之，**统计学**（statistics）是收集、处理、分析、解释数据并从数据中得出结论的原则和方法。统计学所提供的是一系列有关数据收集、数据处理和分析的方法。

数据收集就是取得所需要的数据。数据的收集方法可分为两大类：一是观察方法；二是实验方法。观察方法是通过调查或观测而获得数据；实验方法是在控制试验对象的条件下通过实验而获得数据。

数据处理是对所获得的数据进行加工和处理，包括数据的计算机录入、筛选、分类和汇总等，以符合进一步分析的需要。

数据分析是利用统计方法对数据进行分析。数据分析所用的方法大体上可分为描述统计和推断统计两大类。**描述统计**（descriptive statistics）主要是利用图表对数据进行展示，计算一些简单的统计量（诸如比例、比率、平均数、标准差等）进行分析。**推断统计**（inferential statistics）主要研究如何根据样本信息来推断总体的特征，内容包括参数估计和假设检验两大

类。参数估计是利用样本信息推断所关心的总体特征，假设检验则是利用样本信息判断对总体的某个假设是否成立。比如，从一批灯泡中随机抽取少数几个灯泡作为样本，测出它们的使用寿命，然后根据样本灯泡的平均使用寿命估计这批灯泡的平均使用寿命，或者是检验这批灯泡的使用寿命是否等于某个假定值，这就是推断统计要解决的问题。

数据解释是对分析的结果进行说明，包括结果的含义、从分析中得出的结论等。

统计学是一门关于数据的科学，它研究的是来自各领域的数据，提供的是一套通用于所有学科领域的获取数据、分析数据并从数据中得出结论的原则和方法。统计方法是通用于所有学科领域的，而不是为某个特定的问题领域构造的。当然，统计方法和技术并不是一成不变的，使用者在给定的情况下必须根据所掌握的专门知识选择使用这些方法，而且，如果需要，还要进行必要的修正。

正如有的学者所指出的那样："统计学基本上是寄生的。靠研究其他领域内的工作而生存。这不是对统计学的轻视，这是因为对很多寄主来说，如果没有寄生虫就会死。对有的动物来说，如果没有寄生虫就不能消化它们的食物。因此，人类奋斗的很多领域，如果没有统计学，虽然不会死亡，但一定会变得很弱。"① 这看上去似乎将统计边缘化了，但实际上正说明了统计在各学科领域的独特地位和作用，也表明了统计作为一门独立的学科所具有的特点。

1.1.2 统计学的应用

说出哪些领域应用统计，这很困难，因为几乎所有的领域都用统计；说出哪些领域不用统计，同样也很困难，因为几乎找不到一个不用统计的领域。可以说，统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。只要有数据的地方，就会用到统计方法。

1. 统计学的应用领域

统计学被广泛应用到各个学科领域，对各学科的发展做出了重要贡献。这里，我们不想列举统计学的应用领域，只想通过几个简单的例子说明统计学的应用。

【例 1—1】 用统计识别作者。1787—1788 年，三位作者亚历山大·汉密尔顿（Alexander Hamilton）、约翰·杰伊（John Jay）和詹姆斯·麦迪逊（James Madison）为了说服纽约人认可宪法，匿名发表了著名的 85 篇论文。这些论文的作者大多已经得到了识别，但是，其中 12 篇论文的作者身份引起了争议。通过对这些论文中不同单词的频数进行统计分析，得出的结论是詹姆斯·麦迪逊最有可能是这 12 篇论文的作者。现在，对于这些存在争议的论文，认为詹姆斯·麦迪逊是原创作者的说法占主导地位，而且几乎可以肯定这种说法是正确的。

【例 1—2】 用简单的描述统计量得到一个重要发现。费希尔（R. A. Fisher）在 1952 年的一篇文章中举了一个例子，说明如何由基本的描述统计量的知识引出一个重要的发现。20 世纪早期，哥本哈根卡尔堡实验室的施密特（J. Schmidt）发现不同地区所捕获的同种鱼类的脊椎骨和鳃线的数量有很大不同，甚至在同一海湾内不同地点所捕获的同种鱼类，也有这样的倾向。然而，鳗鱼的脊椎骨的数量变化不大。施密特从欧洲各地、冰岛、亚速尔群岛以及尼罗河等几乎分离的海域里所捕获的鳗鱼样本中，计算发现了几乎一样的均值和标准偏差值。由此，施密特推断所有各个不同海域内的鳗鱼是由海洋中某公共场所繁殖的。后来名为"戴纳"（Dana）的科学考察船在一次远征中发现了这个场所。

【例 1—3】 挑战者号航天飞机失事预测。1986 年 1 月 28 日清晨，载有 7 名宇航员的挑战者号进入发射状态。发射几分钟后，航天飞机发生爆炸，机上的宇航员全部遇难。在此次失事前，

① C. R. 劳：《统计与真理——怎样运用偶然性》，北京，科学出版社，2004。

试读结束：需要全本请在线购买： www.ertongbook.com

该航天飞机 24 次发射成功。将航天飞机送入太空的两个固体燃料推进器由 6 只 O 型项圈密封，在几次飞行中，曾发生过 O 型项圈被腐蚀或气体泄漏事故。这类事故与气温是否有关系呢？天气预报预报本次发射时的气温为零下 0.56℃。下面的表 1—1 是 23 次飞行中 O 型项圈发生腐蚀或泄漏事故损坏的个数（因变量 y）及发射时火箭连接处的温度（自变量 x）数据。

表 1—1　　挑战者号航天飞机 23 次飞行中损坏的 O 型项圈个数和发射时的温度

飞行次数	O 型项圈的损坏个数	温度（℃）	飞行次数	O 型项圈的损坏个数	温度（℃）
1	2	11.7	13	1	21.1
2	1	13.9	14	1	21.1
3	1	14.4	15	0	22.2
4	1	17.2	16	0	22.8
5	0	18.9	17	0	23.9
6	0	19.4	18	2	23.9
7	0	19.4	19	0	24.4
8	0	19.4	20	0	25.6
9	0	20.0	21	0	26.1
10	0	20.6	22	0	27.2
11	0	21.1	23	0	24.4
12	0	21.1			

根据表 1—1 的数据进行线性回归，得到的回归方程为 $\hat{y}=2.177\,1-0.085\,6x$。由此得到当温度为 -0.56℃时，O 型项圈发生事故的预计个数为 2.225 个。结果显示连接处的温度与 O 型项圈事故之间有一定的相关性。如果当时管理者看到了回归的预测结果，推迟发射也许会成为最佳选择。

前两个是统计得以应用并取得成效的例子，后一个是统计结果未被采纳而酿成惨剧的例子。不管怎样，它们都表明统计在许多领域都有广泛应用。

2. 统计的误用与滥用

大约在一个世纪以前，政治家本杰明·迪斯雷利（Benjamin Disraeli）曾有一个著名的论断："有三类谎言：谎言、糟透的谎言和统计。"统计常常被人们有意或无意地滥用，比如错误的统计定义、错误的图表展示、一个不合理的样本、数据的篡改或造假等。这些误用有些是常识性的，有些是技术性的，有些则是故意的。作为从数据中寻找事实的统计，却被有些人变成了歪曲事实的工具。你也许常常看到这样的产品质检报告：某产品的抽样合格率是 80%。乍看上去没什么问题，但如果事实上只抽查了 5 件产品，有 4 件合格。这样的合格率能说明什么问题呢？在马路上随便采访几个人，他们的看法能代表大多数人的观点吗？"调查结果表明……"调查了多少个人？是随机调查的吗？样本是怎样选取的？这看上去是在用事实说话，实际上是统计陷阱。

在管理领域，统计也往往被作为两个极端使用：一个极端是复杂问题简单化，一些不懂或不太懂统计的人认为统计没什么用，他们因为不懂统计而瞧不起统计，他们不用或几乎不用统计方法分析数据，即使做些统计分析，往往也是表面上的。走入这一极端的人，他们决策的依据就是自己的大脑：一些杂乱无章的信息组合出的某种直觉。如果决策是正确的，更增加了他们的自信，更加感到不用统计也挺好；如果决策出了毛病，便会找出一大堆推脱的理由：市场难测，环境突变，竞争激烈，需求疲软，价格下跌，管理不善，成本上升，出口

下降……另一个极端是把简单问题复杂化，特别是在管理领域，一些管理者把本来可以用简单方法解决的问题故意复杂化，他们不用简单的分析方法，而用复杂的分析方法；他们为证明管理的科学性，建立一个别人看不懂的模型，编一大堆程序，输出一大堆数字和符号；他们得出用统计语言陈述的结论，提出一些似是而非的建议……这样的分析往往脱离了管理问题，对实际决策也未必有用。在统计应用中，这两个极端都是不可取的。管理决策中不用统计几乎不可想象，把简单问题复杂化对管理决策也未必有用。从统计的实际应用来看，简单的方法不一定没用，复杂的方法也不一定有用。统计应该被恰当地应用到它能起作用的地方。不能把统计神秘化，更不能歪曲统计，把统计作为掩盖实事的陷阱。

曲解统计是一种常见现象。在有些人看来，使用统计就是寻找支持：他们的心目中可能有了某种“结论”性的东西，或者说他们希望看到一种符合他们需要的某种结论，而去找些数据来支持他们的结论。如果数据分析的结果与他们预期的结论一致，他们就会宣称自己是用科学方法得到的结论；如果与预期的不一致，他们要么篡改数据，要么对统计弃而不用。这恰恰歪曲了数据分析的本质。数据分析的真正目的是从数据中找出结论，从数据中寻找启发，而不是寻找支持。真正的数据分析事先是没有结论的，通过对数据的分析才得出结论。

1.2 数据及其来源

统计分析离不开数据，没有数据，统计方法就成了无米之炊。数据是什么？怎样获得所需的数据？这就是本节将要介绍的内容。

1.2.1 变量与数据

观察一个企业的销售额，你会发现这个月和上个月有所不同；观察股票市场上涨股票的家数，今天与昨天数量不一样；观察一个班学生的生活费支出，一个人和另一个人不一样；投掷一枚骰子观察其出现的点数，这次投掷的结果和下一次也不一样。这里的“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等就是变量。简言之，**变量**（variable）是描述观察对象某种特征的概念，其特点是从一次观察到下一次观察可能会出现不同结果。变量的观测结果就是**数据**（data）。

根据观测结果的特征，变量可以分为类别变量和数值变量两种。

类别变量（categorical variable）是取值为事物属性或类别以及区间值的变量，也称**分类变量**（classified variable）或**定性变量**（qualitative variable）。比如，观察人的性别、公司所属的行业、用户对商品的评价时，得到的结果就不是数字，而是事物的属性。例如，观测性别的结果是“男”或“女”，公司所属的行业为“建造业”、“零售业”、“旅游业”等，用户对商品的评价为“很好”、“好”、“一般”、“差”、“很差”。人的性别、公司所属的行业、用户对商品的评价等作为变量取的值不是数值，而是事物的属性或事物的类别。此外，学生月生活费支出的档次可能分为1 000元以下、1 000～1 500元、1 500～2 000元、2 000元以上4档，作为变量的“月生活费支出档次”的这4档取值也不是普通的数值，而是数值区间，因而也称为**区间值类别变量**。人的性别、公司所属的行业、用户对商品的评价、学生月生活费支出的档次等都是类别变量。

类别变量根据取值是否有序通常分为两种：**名义**（nominal）值类别变量和**顺序**（ordinal）值类别变量。名义值类别变量也称无序类别变量，其取值是不可以排序的。例如“公司所属的行业”这一变量取值为“建造业”、“零售业”、“旅游业”等，这些取值之间不存在顺序关系。又比如“商品的产地”这一变量的取值为甲、乙、丙、丁，这些取值之间也不存

在顺序关系。顺序值类别变量也称有序类别变量，其取值间可以排序。例如“对商品的评价”这一变量的取值为很好、好、一般、差、很差，这 5 个值之间是有序的。取区间值的变量当然是有序的类别变量。当类别变量只取两个值时也称为**二值**（binary）类别变量，例如“性别”这一变量取值为男和女。二值变量可以看成名义变量，也可以看成有序变量。

类别变量的观测结果称为**类别数据**（categorical data）。类别数据也称为分类数据或定性数据。与类别变量相对应，类别数据相应分为名义值类别数据和顺序值变量数据两种。其中只取两个值的类别数据也称为二值类别数据。

数值变量（metric variable）是取值为数字的变量，也称为**定量变量**（quantitative variable）。例如“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等这些变量的取值可以用数字来表示，都属于数值变量。数值变量的观察结果称为**数值型数据**（metric data）或定量数据。

数值变量根据其取值的不同，可以分为**离散变量**（discrete variable）和**连续变量**（continuous variable）。离散变量是只能取有限个值的变量，而且其取值可以一一列举，如“企业数”、“产品数量”等就是离散变量。连续变量是可以在一个或多个区间中取任何值的变量，它的取值是连续不断的，不能一一列举，如“年龄”、“温度”、“零件尺寸的误差”等都是连续变量。当离散变量的取值很多时，也可以将离散变量当作连续变量来处理。

图 1—1 给出了变量的基本分类。

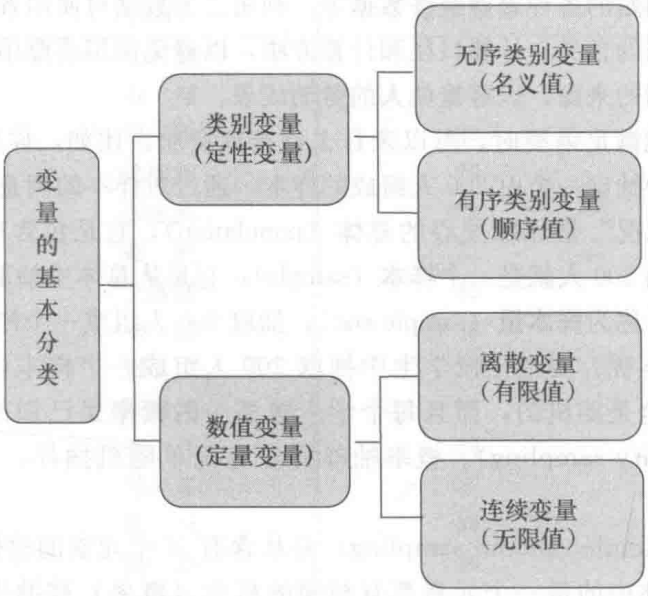

图 1—1　变量的基本分类

上面介绍的是变量的基本分类。当然，也可以从其他角度进行分类，比如随机变量、经验变量和理论变量等。随机变量用数值来描述特定实验一切可能出现的结果，它的取值事先不能确定，具有随机性。经验变量所描述的是周围环境中可以观察到的事物。理论变量则是由统计学家用数学方法所构造出来的一些变量，比如后面的有些章节中将要用到的 z 统计量、t 统计量、χ^2 统计量、F 统计量等都是理论变量。

数据也可以从其他角度进行分类。比如，按照数据的收集方法可分为**观测数据**（observational data）和**实验数据**（experimental data）。观测数据是通过调查或观测而收集到的数据，这类数据是在没有对事物进行人为控制的条件下得到的，有关社会经济现象的数据几乎都是

观测数据。实验数据则是在实验中控制实验对象而收集到的数据，比如，对一种新药疗效的实验数据，对一种新的农作物品种的实验数据。自然科学领域的大多数数据都为实验数据。按照被描述的现象与时间的关系，可以将数据分为**截面数据**（cross-sectional data）和**时间序列数据**（time series data）。截面数据是在相同或近似相同的时间点上收集的数据，这类数据通常是在不同的空间上获得的，用于描述现象在某一时刻的变化情况，比如，2014 年我国各地区的国内生产总值（GDP）数据就是截面数据。时间序列数据是在不同时间上收集到的数据，这类数据是按着时间顺序收集到的，用于描述现象随时间而变化的情况。比如 2000—2014 年我国的国内生产总值数据就是时间序列数据。

区分数据的类型是必要的，因为对不同类型的数据，需要采用不同的统计方法来处理和分析。比如，对类别数据通常进行比例和比率分析、列联表分析和 χ^2 检验等；对数值型数据可以用更多的方法进行分析，如计算各种统计量、进行参数估计和检验等。

1.2.2 数据的来源

从哪里取得所需的数据呢？对大多数使用者来说，亲自去做调查或实验往往是不可能的。所使用的数据大多是别人通过调查或科学实验得到的数据，对使用者来说就是二手数据。

二手数据主要是公开出版或公开报道的数据，这类数据主要来自各研究机构、国家和地方的统计部门、其他管理部门、专业的调查机构以及各种报纸、杂志、图书、广播、电视传媒。现在，随着计算机网络技术的发展，也可以在网络上获取所需的各种数据。比如，各种金融产品的交易数据、官方统计网站的各种宏观经济数据等。利用二手数据对使用者来说既经济又方便，但使用时应注意统计数据的含义、计算口径和计算方法，以避免误用或滥用。同时，在引用二手数据时，一定要注明数据的来源，以尊重他人的劳动成果。

当已有的数据不能满足需要时，可以亲自去调查或实验。比如，你想了解全校学生的生活费支出状况，可以从中抽出一个由 200 人组成的样本，通过对样本的调查获得数据。这里“全校所有学生生活费支出状况”是你所关心的**总体**（population），它是包含所研究的全部个体（数据）的集合。所抽取的 200 人就是一个**样本**（sample），它是从总体中抽取的一部分元素的集合。构成样本的元素的数目称为**样本量**（sample size），抽取 200 人组成一个样本，样本量就是 200。

怎样获得一个样本呢？要在全校学生中抽取 200 人组成一个样本，如果全校学生中每一个学生被抽中与否完全是随机的，而且每个学生被抽中的概率是已知的，这样的抽样方法称为**概率抽样**（probability sampling）。概率抽样方法有简单随机抽样、分层抽样、系统抽样、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，抽取 n 个元素组成一个样本，使得总体中的每一个元素都有相同的机会（概率）被抽中。采用简单随机抽样时，如果抽取一个个体记录下数据后，再把这个个体放回到原来的总体中参加下一次抽选，这样的抽样方法叫做**有放回抽样**（sampling with replacement）；如果抽中的个体不再放回，再从所剩下的个体中抽取第二个元素，直到抽取 n 个个体为止，这样的抽样方法叫做**无放回抽样**（sampling without replacement）。当总体数量很大时，无放回抽样可以视为有放回抽样。由简单随机抽样得到的样本称为**简单随机样本**（simple random sample）。多数统计推断都是以简单随机样本为基础的。

分层抽样（stratified sampling）也称分类抽样，它是在抽样之前先将总体的元素划分为若干层（类），然后从各个层中抽取一定数量的元素组成一个样本。比如，要研究学生的生活费支出，可先将学生按性别分成两类，然后从每一类中抽取一定数量的学生组成一个样本。假定总体

共有 N 个元素，要抽取 n 个元素作为样本，如果总体的 N 个元素被分成 N_1，N_2，…，N_k 几个类别，按照 N_1/N，N_2/N，…，N_k/N 的比例抽取样本，则称为等比例分层抽样。如果人为指定要在某一类中抽取指定数量的样本元素，则称为不等比例抽样。分层抽样的优点是可以使样本分布在各个层内，从而使样本在总体中的分布比较均匀，可以降低抽样误差。

系统抽样（systematic sampling）也称等距抽样，它是先将总体各元素按某种顺序排列，并按某种规则确定一个随机起点，然后，每隔一定的间隔抽取一个元素，直至抽取 n 个元素组成一个样本。比如，要从全校学生中抽取一个样本，可以找到全校学生的花名册，按花名册中的学生顺序，用随机数找到一个随机起点，然后依次抽取，就得到一个样本。

整群抽样（cluster sampling）是先将总体划分成若干群，然后以群作为抽样单元从中抽取部分群组成一个样本，再对抽中的每个群中包含的所有元素进行观察。比如，可以把每一个学生宿舍看作一个群，在全校学生宿舍中抽取一定数量的宿舍，然后对抽中的宿舍中的每一个学生都进行调查。整群抽样的误差相对要大一些。

下面通过一个例子说明从总体中抽取随机样本的过程。

【例 1—4】 表 1—2 是一个班级 50 个学生的名单，采用简单随机抽样抽出 10 个学生组成一个随机样本。

表 1—2　　**某班级 50 个学生的名单**

学生编号	姓名	学生编号	姓名
1	张　松	26	姜　洋
2	王　翔	27	隗　佳
3	田　雨	28	于　静
4	徐丽娜	29	李　华
5	张志杰	30	高　云
6	赵　颖	31	金梦迪
7	王智强	32	徐海涛
8	宋　媛	33	张　洋
9	袁　方	34	李冬茗
10	张建国	35	李宗洋
11	李　佳	36	刘皓天
12	马凤良	37	刘文涛
13	陈　风	38	卢　阳
14	杨　波	39	马　强
15	孙学伟	40	孟子铎
16	林　丽	41	潘　凯
17	谭英键	42	邱　爽
18	欧阳飞	43	邵海阳
19	吴　迪	44	王浩波
20	周　祥	45	孙梦婷
21	刘晓军	46	唐　健
22	李国胜	47	尹　韩
23	蒋亚迪	48	王　迪
24	崔　勇	49	王　倩
25	黄向春	50	王思思

解：由 SPSS 在 50 人中抽取大约 20%样本的结果如表 1—3 所示。

表 1—3 在 50 人中抽取大约 20%样本的结果（部分显示）

	学生编号	姓名	filter_$
1	1	张 松	1
2	2	王 翔	0
3	3	田 雨	1
4	4	徐丽娜	1
5	5	张志杰	0
6	6	赵 颖	1
7	7	王智强	0
8	8	宋 媛	0
9	9	袁 方	0
10	10	张建国	1
11	11	李 佳	0
12	12	马凤良	0
13	13	陈 风	0
14	14	杨 波	1
15	15	孙学伟	0
16	16	林 丽	0
17	17	谭英键	0
18	18	欧阳飞	0
19	19	吴 迪	0
20	20	周 祥	0

表 1—3 中标号栏中画有斜杠“/”标记的表示未被选中的个案，同时系统会自动产生一个名为“filter_$”的筛选指示变量，被选中的记录取值为 1，未被选中的记录取值为 0。

表 1—4 给出了指定抽取大约 20%的一个随机样本的结果[①]（实际上抽取了 12 个人）。

表 1—4 指定抽取大约 20%的一个随机样本的结果（共 12 人）

姓名

		频率	百分比	有效百分比	累积百分比
有效	高 云	1	8.3	8.3	8.3
	黄向春	1	8.3	8.3	16.7
	李宗洋	1	8.3	8.3	25.0
	刘皓天	1	8.3	8.3	33.3
	孙梦婷	1	8.3	8.3	41.7
	田 雨	1	8.3	8.3	50.0
	徐海涛	1	8.3	8.3	58.3
	徐丽娜	1	8.3	8.3	66.7
	杨 波	1	8.3	8.3	75.0
	张 松	1	8.3	8.3	83.3
	张建国	1	8.3	8.3	91.7
	赵 颖	1	8.3	8.3	100.0
	合计	12	100.0	100.0	

表 1—5 给出了指定抽取 10 个人的一个随机样本的结果。

① 由于样本是随机采取的，每次抽样都可能会产生一个不同的样本。

试读结束：需要全本请在线购买： www.ertongbook.com