



Statistics

21世纪统计学系列教材

Statistics with Excel

统计学 —— 基于Excel

贾俊平 编著

 中国人民大学出版社

Statistics 统计学系列教材

Statistics with Excel

统计学 —— 基于Excel

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：基于 Excel/贾俊平编著. —北京：中国人民大学出版社，2017. 4
21 世纪统计学系列教材
ISBN 978-7-300-24276-7

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 (2017) 第 057804 号

21 世纪统计学系列教材
统计学——基于 Excel
贾俊平 编著
Tongjixue: Jiyu Excel

出版发行	中国人民大学出版社	邮政编码	100080
社 址	北京中关村大街 31 号		
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	北京鑫丰华彩印有限公司		
规 格	185 mm×260 mm 16 开本	版 次	2017 年 4 月第 1 版
印 张	11.25 插页 1	印 次	2017 年 4 月第 1 次印刷
字 数	232 000	定 价	29.00 元

版权所有 侵权必究 印装差错 负责调换

前言

● 本书概要

本书是一本统计学基础教材，全书共9章，第1章主要介绍变量、数据及其分类以及数据的来源。第2章介绍数据的预处理和频数分布表的制作。第3章和第4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5~第7章介绍统计推断的基本原理和方法，包括统计量的概率分布及参数估计和假设检验。第8章介绍实际中广泛应用的相关与回归分析方法。第9章介绍时间序列分析和预测。

● 本书特色

◇全部使用 Excel 实现计算与分析。本书所有例题的计算和分析全部使用 Excel (2013 版) 实现。每种方法均以文本框的形式给出了详细操作步骤。考虑到读者对象的特点，Excel 能够实现的一些方法并没有写进书中，如方差分析、双样本正态检验和 t 检验、双样本方差比 F 检验等。此外，在本书末的附录里给出了书中用到的 Excel 函数的列表，以及用 Excel 生成常用概率分布表的方法。

◇注重统计方法应用。每章开头均以问题与思考的形式引出本章的内容。对方法的介绍侧重于原理和思想，完全避免数学推导，并把繁杂的计算交给 Excel 来完成。读完本书后你会发现统计学并不那么难学，而且比你想象的有趣和有用。

◇方便教学和学习。每章都配有详细的 ppt，并配有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。这些资源放在中国人民大学出版社工商分社网站 (www.rdjg.com.cn) 上，读者可自行下载。

● 读者对象

本书适用的读者包括：普通高等院校非统计学专业的本科生；中等职业教育或继续教育的本科生和大专生；实际工作者及各领域的管理人员；对统计学知识感兴趣的其他读者。

贾俊平

于中国人民大学统计学院

目 录

第 1 章 数据及其来源	1
1.1 统计学与数据	1
1.1.1 什么是统计学	2
1.1.2 变量、数据及其分类	2
1.2 数据的来源	4
1.2.1 数据的间接来源	4
1.2.2 数据的直接来源	4
1.3 用 Excel 产生随机数	8
1.3.1 Excel【数据分析】工具的安装	8
1.3.2 用 Excel 产生随机数	9
习题	12
第 2 章 数据处理与频数分布	14
2.1 数据的预处理	14
2.1.1 数据审核与验证	14
2.1.2 数据排序与筛选	16
2.2 类别数据的频数分布	21
2.2.1 简单频数分布表	21
2.2.2 二维列联表	23
2.2.3 类别数据的简单分析	23
2.3 数值数据的类别化	24
2.3.1 数据分组	25
2.3.2 用 Excel 制作频数分布表	26
习题	28



第 3 章 数据的可视化	30
3.1 类别数据的可视化	30
3.1.1 条形图	31
3.1.2 饼图和环形图	34
3.2 数值数据的可视化	35
3.2.1 直方图	36
3.2.2 散点图和气泡图	38
3.2.3 雷达图和轮廓图	41
3.3 合理使用图表	44
习题	44
第 4 章 数据的描述统计量	47
4.1 描述水平的统计量	47
4.1.1 平均数	48
4.1.2 分位数	49
4.1.3 众数	53
4.2 描述差异的统计量	54
4.2.1 全距和四分位距	55
4.2.2 方差和标准差	55
4.2.3 离散系数	57
4.2.4 标准分数	58
4.3 描述分布形状的统计量	60
4.3.1 偏度系数	60
4.3.2 峰度系数	61
4.4 Excel【数据分析】工具的应用	62
习题	63
第 5 章 统计量及其概率分布	66
5.1 概率与随机变量	67
5.1.1 什么是概率	67
5.1.2 随机变量及其概括性度量	68
5.2 随机变量的概率分布	69
5.2.1 二项分布	70
5.2.2 正态分布	71
5.2.3 χ^2 分布、 t 分布和 F 分布	76
5.3 样本统计量的概率分布	80
5.3.1 统计量及其分布	80

5.3.2 样本均值的分布	81
5.3.3 其他统计量的分布	83
5.3.4 统计量的标准误	84
习题	85
第6章 参数估计	86
6.1 参数估计的原理	87
6.1.1 点估计与区间估计	87
6.1.2 评价估计量的标准	90
6.2 总体均值的区间估计	92
6.2.1 大样本的估计	92
6.2.2 小样本的估计	93
6.3 总体比例的区间估计	95
6.4 总体方差的区间估计	95
6.5 样本量的确定	97
6.5.1 估计总体均值时样本量的确定	97
6.5.2 估计总体比例时样本量的确定	98
习题	98
第7章 假设检验	100
7.1 假设检验的步骤	100
7.1.1 提出假设	101
7.1.2 确定显著性水平	102
7.1.3 做出决策	103
7.1.4 表述结果	106
7.2 总体均值的检验	107
7.2.1 大样本的检验	107
7.2.2 小样本的检验	110
7.3 总体比例的检验	112
7.4 总体方差的检验	113
习题	114
第8章 相关与回归分析	116
8.1 变量间关系的度量	116
8.1.1 变量间的关系	117
8.1.2 相关关系的描述	117
8.1.3 相关关系的度量	120



8.2 回归模型及其参数估计	122
8.2.1 一元线性回归模型与回归方程	122
8.2.2 参数的最小平方估计	123
8.3 模型评估和检验	127
8.3.1 模型评估	127
8.3.2 显著性检验	129
8.4 利用回归方程进行预测	131
8.4.1 平均值的置信区间	131
8.4.2 个别值的预测区间	131
8.5 残差分析	133
8.5.1 残差与标准化残差	134
8.5.2 残差图及其解读	135
习题	136
第9章 时间序列分析和预测	139
9.1 增长率分析	140
9.1.1 增长率与平均增长率	140
9.1.2 年化增长率	142
9.2 时间序列的成分和预测方法	143
9.2.1 时间序列的成分	143
9.2.2 预测方法的选择与评估	145
9.3 平滑法预测	147
9.3.1 移动平均预测	147
9.3.2 简单指数平滑预测	147
9.4 趋势预测	151
9.4.1 线性趋势预测	151
9.4.2 非线性趋势预测	153
9.5 分解法预测	158
习题	163
附录1 Excel 中的统计函数	165
附录2 用 Excel 生成概率分布表	168
参考文献	173

C 第 1 章

Chapter 1

数据及其来源

思考一下

- ▶ 在你的印象中，统计学是什么？
- ▶ 你日常生活和工作中接触过哪些数据？
- ▶ 调查一批人的性别、职业、月收入等，这里涉及哪些变量？
- ▶ 如果让你在全校大学生中做一次调查，你会调查所有的学生还是抽取一部分学生做调查？假定让你从全校 10 000 名学生中随机抽取 200 人，你会怎么做？
- ▶ 你使用过哪些软件，Excel，SPSS，R 还是其他？
- ▶ 你认为不使用软件能做统计分析吗？



在日常工作和生活中，经常会接触各类数据，比如，空气质量（PM 2.5）数据、全国的 GDP（国内生产总值）数据、CPI（居民消费价格指数）数据、股票交易数据、某品牌手机的市场占有率，等等。这些数据如果不加以分析，将仅仅是数据，给你提供的信息十分有限。那么，如何分析这些数据？用什么方法分析？这就是统计学要解决的问题。

1.1 统计学与数据

如果有了数据不去分析，那么数据就没有太多价值。这就涉及两个问题：一是到哪里去找所需要的数据；二是用什么方法分析这些数据。本节首先介绍统计学的含

义，然后介绍数据及其分类。

1.1.1 什么是统计学

统计学 (statistics) 是一门分析数据的科学，它提供的是收集数据、处理数据和分析数据的一套方法和技术，通过对数据的分析得出结论。

收集数据就是获得所需要的数据。处理数据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** (descriptive statistics) 和**推断统计** (inferential statistics) 两大类。描述统计主要是利用图表形式对数据进行汇总和展示，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。推断统计主要是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，从一批电池中随机抽取少数几块电池作为样本，测出它们的使用寿命，然后根据样本电池的平均使用寿命估计这批电池的平均使用寿命，或者检验这批电池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。

概括地讲，统计学是一门关于数据的科学，它研究的是来自各领域的的数据，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。这些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而构造的。统计方法不是一成不变的，使用者在特定的情况下必须根据所掌握的专门知识作出选择，而且如果需要，还要进行必要的修正。

在人类社会进入大数据时代的今天，数据分析越来越引起人们的重视，统计学在各领域也发挥越来越大的作用。可以说，统计学提供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。

1.1.2 变量、数据及其分类

观察一个企业的销售额，这个月与上个月可能不同；观察一只股票的价格，今天与昨天不一样；观察每个人的月收入，一个人和另一个人不一样；投掷一枚色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”“股票价格”“月收入”“投掷一枚色子出现的点数”等就是**变量** (variable)。简言之，变量是描述所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的结果。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大类。

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**定性变量** (qualitative variable)。比如，观察人的性别、上市公司所属的行业、用

用户对商品满意度的评价，得到的结果就不是数字，而是事物的属性。比如，观测性别的结果是“男”或“女”，上市公司所属的行业为“金融业”“地产业”“旅游业”等；用户对商品满意度的评价为“很满意”“满意”“一般”“不满意”“很不满意”。人的性别、上市公司所属的行业、用户对商品满意度的评价等，其结果就不是数值，而是事物的属性或类别。此外，将从业人员的月收入分为5 000元以下、5 000~10 000元、10 000~15 000元、15 000~20 000元、20 000元以上5档，这里的“月收入档次”的取值也不是普通的数值，而是数值区间，这实际上是将数值转化成了类别。人的性别、上市公司所属的行业、用户对商品满意度的评价、月收入的档次都是类别变量。类别变量的观测结果称为**类别数据**（categorical data）或定性数据。

类别变量根据取值是否有序通常分为无序类别变量和有序类别变量两种。无序类别变量也称**名义**（nominal）值类别变量，其取值不可以排序。例如“上市公司所属的行业”这一变量取值为“金融业”“地产业”“旅游业”等，这些取值之间不存在顺序关系。再比如“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。有序类别变量也称**顺序**（ordinal）值类别变量，其取值间可以排序。例如“对商品满意度的评价”这一变量的取值为“很满意”“满意”“一般”“不满意”“很不满意”，这5个值之间是有序的。

数值变量（metric variable）是取值为数字的变量，也称**定量变量**（quantitative variable）。例如“企业销售额”“股票价格”“月收入”“投掷一枚色子出现的点数”等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“上市公司家数”“一个社区的居民户数”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”“股票价格”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显示了变量的基本分类。

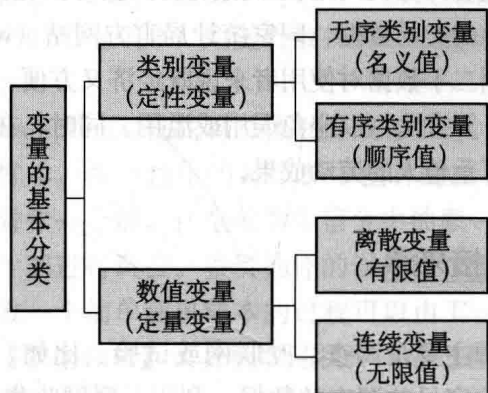
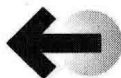


图 1—1 变量的基本分类



由于数据是变量的观测结果，因此数据的基本分类与图 1—1 所示的变量分类相同。此外，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）和试验数据（experimental data）。观测数据是通过调查或观测收集到的数据，这类数据是在没有对事物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。试验数据则是在试验中控制试验对象收集到的数据，比如，对一种新药疗效的试验数据，对一种新的农作物品种的试验数据。自然科学领域的大多数数据都是试验数据。按照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获得的，用于描述现象在某一时刻的变化情况，比如，2016 年我国各地区的 GDP 数据就是截面数据。时间序列数据是在不同时间收集到的数据，这类数据是按时间顺序收集的，用于描述现象随时间变化而变化的状况，比如 2000—2016 年我国的 GDP 数据就是时间序列数据。

1.2 数据的来源

从使用者的角度看，数据主要来源于两种渠道：一是直接的调查和试验，属于直接来源；二是别人调查或试验的数据，属于间接来源。

1.2.1 数据的间接来源

对大多数使用者来说，亲自去做调查或试验往往不现实。他人调查或试验的数据，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，广泛分布在报刊、图书、广播、电视传媒中。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、国家统计局官方网站（www.stats.gov.cn）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。

1.2.2 数据的直接来源

数据的直接来源主要是调查、互联网或试验。比如，统计部门调查取得的数据；其他部门或机构为特定目的调查的数据；利用互联网收集的各类产品交易、生产和经营活动等产生的大数据。试验是取得自然科学数据的直接来源。

已有的数据不能满足需要时，可以亲自去调查或试验。比如，你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，称为**有放回抽样**（sampling with replacement）；如果抽中的个体不再放回，再从剩下的个体中抽取第二个元素，直到抽取 n 个个体为止，这样的抽样方法称为**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）。简单随机抽样是其他抽样方法的基础，多数统计推断也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，可先将学生按地区进行分类，然后从各地区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层，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每一个学生都进行调查。整群抽样的误差相对要大一些。

在实际应用中，抽取一个简单随机样本的过程可以由 Excel 来完成。下面通过一个例子说明用 Excel 的【数据分析】工具抽取随机样本的过程。



例 1—1

表 1—1 是 60 个房地产类上市公司的股票代码和股票名称，随机抽取 6 个上市公司组成一个样本。

表 1—1 60 个房地产类上市公司的股票代码和股票名称

序号	股票代码	股票名称	序号	股票代码	股票名称	序号	股票代码	股票名称
1	000558	莱茵置业	21	600791	京能置业	41	600067	冠城大通
2	600082	海泰发展	22	600895	张江高科	42	000931	中关村
3	600193	创兴置业	23	000042	深长城	43	000046	泛海建设
4	600322	天房发展	24	000514	渝开发	44	600638	新黄浦
5	600665	天地源	25	000005	世纪星源	45	000402	金融街
6	000006	深振业 A	26	000671	阳光城	46	600533	栖霞建设
7	600657	信达地产	27	000029	深深房 A	47	600240	华业地产
8	600745	中茵股份	28	600606	金丰投资	48	600159	大龙地产
9	000534	万泽股份	29	600185	格力地产	49	600663	陆家嘴
10	002305	南国置业	30	000511	银基发展	50	000011	深物业 A
11	600684	珠江实业	31	000517	荣安地产	51	000573	粤宏远
12	000711	天伦置业	32	600622	嘉宝集团	52	000002	万科
13	002285	世联地产	33	000718	苏宁环球	53	600734	实达集团
14	002133	广宇集团	34	600246	万通地产	54	600048	保利地产
15	000056	深国商	35	600773	西藏城投	55	000024	招商地产
16	000838	国兴地产	36	000150	宜华地产	56	600823	世茂股份
17	600743	华远地产	37	000616	亿城股份	57	600648	外高桥
18	600052	浙江广厦	38	000502	绿景地产	58	600383	金地集团
19	000036	华联控股	39	000031	中粮地产	59	600266	北京城建
20	600639	浦东金桥	40	000009	中国宝安	60	600675	中华企业

解：首先将 60 个公司的股票代码和股票名称录入 Excel 工作表中的一列，并对每只股票进行编号，如 1, 2, …, 60，然后用 Excel 【分析工具】中的【抽样】命令抽取随机样本（【数据分析】工具的安装见文本框 1—2）。操作步骤如文本框 1—1 所示。

文本框 1—1

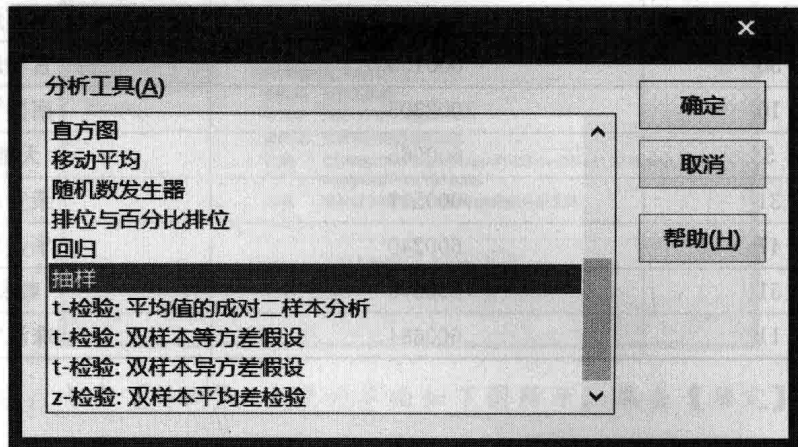
用 Excel 抽取随机样本



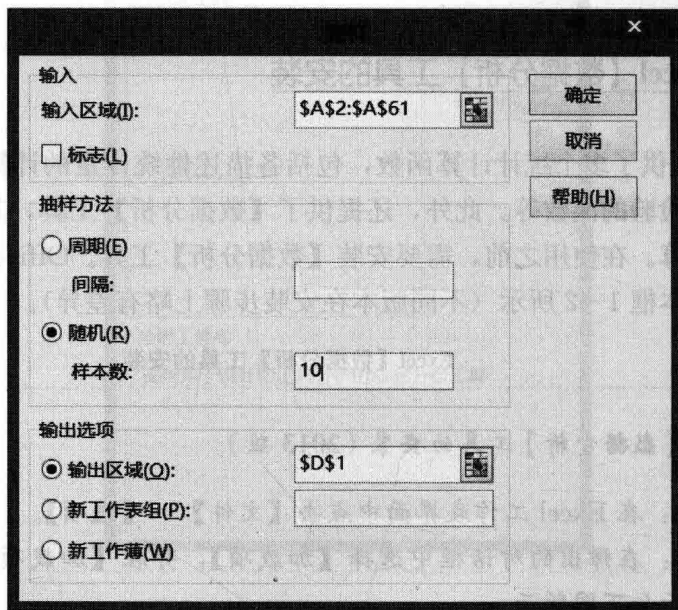
使用【数据分析】工具抽取样本

第1步：在工作表中点击【数据】→【数据分析】。

第2步：在弹出的对话框中选择【抽样】，界面如下图所示。



第3步：单击【确定】。在出现的对话框【输入区域】中输入代码区域（数值型数据直接输入数据区域）；在【抽样方法】中单击【随机】，在【样本数】中输入需要抽样的样本量；在【输出区域】中选择抽样结果放置的区域。出现的界面如下图所示。



单击【确定】，即得到一个随机样本。

按上述步骤得到的随机样本如表 1—2 所示。

表 1—2 用【数据分析】工具抽取的例 1—1 一个随机样本

样本序号	股票代码	股票名称
4	600322	天房发展
59	600266	北京城建
55	000024	招商地产
36	000150	宜华地产
10	002305	南国置业
5	600665	天地源
31	000517	荣安地产
47	600240	华业地产
51	000573	粤宏远
11	600684	珠江实业

1.3 用 Excel 产生随机数

有时需要生成各种分布的随机数做模拟分析。用 Excel 提供的统计函数或【数据分析】工具中的【随机数发生器】可以产生一些常用分布的随机数。

1.3.1 Excel【数据分析】工具的安装

Excel 提供了多个统计计算函数，包括各描述性统计量的计算函数、概率分布函数、估计和检验的函数等。此外，还提供了【数据分析】工具，其中包含多种基本统计方法的计算。在使用之前，需要安装【数据分析】工具。Office 2013 版本的具体安装步骤如文本框 1—2 所示（不同版本在安装步骤上略有差异）。

文本框 1—2 Excel【数据分析】工具的安装



Excel【数据分析】工具的安装（2013 版）

第 1 步：在 Excel 工作表界面中点击【文件】→【选项】。

第 2 步：在弹出的对话框中选择【加载项】，并在【加载项】下选择【分析工具库】，界面如下图所示。