

互联网+远程一体化智慧数字教材

TONGJIXUE JIAOCHENG

# 统计学教程

主 编 金勇进



互联网+教育云服务

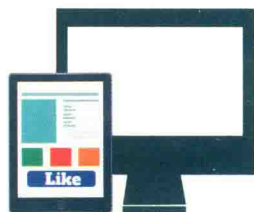
富媒体教材：权威性、实用性、个性化和开放性  
微视频随时随地助力学习



支持多终端学习



丰富的数字课程资源



强大的题库和作业、考试系统

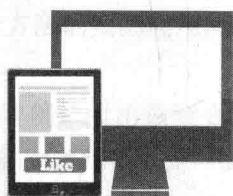


互联网+远程一体化智慧数字教材

TONGJIXUE JIAOCHENG

# 统计学教程

主 编 金勇进



中国人民大学出版社

·北京·

## 图书在版编目 (CIP) 数据

统计学教程: 数字教材版/金勇进主编. —北京: 中国人民大学出版社, 2018. 6  
互联网+远程一体化智慧数字教材  
ISBN 978-7-300-25852-2

I. ①统… II. ①金… III. ①统计学-远程教育-教材 IV. ①C8

中国版本图书馆 CIP 数据核字 (2018) 第 113664 号

互联网+远程一体化智慧数字教材  
统计学教程 (数字教材版)  
主 编 金勇进  
Tongjixue Jiaocheng

---

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	<a href="http://www.crup.com.cn">http://www.crup.com.cn</a>		
	<a href="http://www.ttrnet.com">http://www.ttrnet.com</a> (人大教研网)		
经 销	新华书店		
印 刷	北京昌联印刷有限公司		
规 格	185 mm×260 mm 16 开本	版 次	2018 年 6 月第 1 版
印 张	12.75	印 次	2018 年 6 月第 1 次印刷
字 数	280 000	定 价	42.00 元

---

版权所有 侵权必究 印装差错 负责调换

# PREFACE 前言

统计学是关于数据的科学，在大数据浪潮滚滚而来的背景下，欲对数据进行更好的挖掘与使用，基本的统计学知识是必不可少的。本书的宗旨是介绍基本的统计方法，同时解读背后其蕴含的统计思想。本书的最大特点是通俗、易懂，没有追求统计方法中的公式推导，而是借助大量生动活泼的例子阐述统计方法的实际应用。与其他同类教科书相比，本书的主要特点表现在：

1. 内容全面，重点突出。本书覆盖了一般统计学教科书中的核心内容，但对一些内容有所侧重。例如，在第二章侧重讨论了误差问题，我们认为统计学的核心问题是对误差的描述和分析；在第三章介绍统计图表的内容后，专门讨论了如何制作一个好的统计图，因为在实践中经常出现统计图误用的情形；第五章和第六章遥相呼应，把重点集中于一个总体和两个总体的参数估计与检验；第七章讨论的列联表中的相关测量是许多教科书中还没有涉及的内容，但在数据分析中十分必要；第十章介绍综合评价指数，因为对数据的综合评价是现实中经常遇到的问题。

2. 省略公式推导，强调实际应用。“统计学教程”主要讲授统计方法，方法中的公式推导是统计理论的重要内容。本书没有追求公式推导，而是强调统计方法的实际应用。所以本书中的例题较多，试图通过例题告诉读者统计方法的具体应用。

3. 本书内容通俗易懂、生动活泼。每章前都精心挑选一个生动案例，引出该章介绍的内容；作者尽量采用生动的语言，设计“学习导航”“小词典”“你知道吗”“试一试”“想一想”等栏目，介绍重要概念，拓展相应知识，引发读者思考，培养动手能力。各章中也有数量不等的“人物小传”，介绍著名统计学家生平和主要学术贡献，帮助读者在学习统计方法的同时，对统计学发展史有所了解。每章的最后有“本章小结”，强调该章核心知识点。

本书特别适合于自学“统计学”课程的读者，也可以作为各类院校“统计学”课程的教材。在本书的写作过程中，中国人民大学统计学院的刘展博士和于力超博士帮助作者整理材料，经作者最后修改形成本书。在此，向刘展博士和于力超博士表示深深的感谢！本书的出版得到了中国人民大学出版社的大力支持，在此一并致以谢意！

# CONTENTS 目录

<b>第一章</b>	<b>引 论</b> .....	1
	第一节 统计数据与统计学 .....	2
	第二节 一些基本概念 .....	5
	第三节 统计与统计软件 .....	10
<b>第二章</b>	<b>数据的收集</b> .....	13
	第一节 数据的来源 .....	14
	第二节 数据误差 .....	20
<b>第三章</b>	<b>数据的描述——数据的直观显示</b> .....	26
	第一节 用统计表描述数据 .....	27
	第二节 用统计图描述数据 .....	30
<b>第四章</b>	<b>数据的描述——重要的统计量</b> .....	42
	第一节 集中趋势的描述 .....	43
	第二节 离散趋势的描述 .....	51
	第三节 数据的标准化处理 .....	58
<b>第五章</b>	<b>参数估计</b> .....	62
	第一节 抽样分布 .....	63
	第二节 点估计和区间估计 .....	65
	第三节 一个总体情形下参数的区间估计 .....	66
	第四节 两个总体情形下参数的区间估计 .....	69
	第五节 关于样本量 .....	73

## 第六章

<b>假设检验</b> .....	79
第一节 假设检验的基本问题 .....	80
第二节 一个总体参数的假设检验 .....	87
第三节 两个总体参数的假设检验 .....	93

## 第七章

<b>列联分析</b> .....	103
第一节 定类数据与列联表 .....	104
第二节 拟合优度检验 .....	108
第三节 独立性检验 .....	113
第四节 列联分析中应注意的问题 .....	116

## 第八章

<b>相关与回归</b> .....	121
第一节 相关分析 .....	122
第二节 回归分析 .....	127

## 第九章

<b>时间序列分析</b> .....	139
第一节 时间序列的描述 .....	140
第二节 时间序列的分解法 .....	148
第三节 时间序列的平滑法 .....	160

## 第十章

<b>指 数</b> .....	168
第一节 基本问题 .....	169
第二节 总指数编制方法 .....	172
第三节 指数体系 .....	179
第四节 几种典型的指数 .....	183
第五节 综合评价指数 .....	190

<b>参考文献</b> .....	197
-------------------	-----

# 第一章

## 引 论

---

### 统计无处不在

我们每天都可以在报纸上、网络上、电视机中看到、听到各种数据，这些数据是统计研究的结果。这些数据意味着什么？数据是如何收集的？如何进行数据分析以获得直观的结论，从而反映我们感兴趣的变量的现状和发展态势？请看下面一些统计数据 and 统计研究结果：

(1) 2009年7月北京期房商品住宅的销量为12 840套，剔除其中的经济适用房和限价房外的期房商品住宅签约套数为10 862套，日均销售约350套，较6月的日均销售385套下跌9.1%。

(2) 央行在公开市场发行了50亿元一年期央行票据，发行利率水平较前次发行直降近95个基点，为2.2459%，达到了2006年5月的水平。

(3) 统计分析结果显示，第二个出生的子女没有第一个聪明，第三个出生的子女没有第二个聪明，以此类推。

(4) 2000年，洛杉矶湖人队14名球员的年薪平均数约为410万美元，中位数约为260万美元。

(5) 全球平均气温近些年总体具有变高的趋势。

这些都是统计学研究的问题和应用统计方法得出的结论。其他类似的股票行情、物价指数、房价波动、汇率变化，乃至交通事故件数、死亡人数等各种数据围绕着我们，我们生活在数据之中，想避开都难，可以说，有数据的地方就有统计，统计无处不在。

## 【学习导航】

- 统计数据的不同类型：按照数据的不同计量尺度、不同收集方法和是否与时间相关划分。
- 统计学的应用领域。
- 统计学的一些基本概念。
- 常用的统计软件。

## 第一节

# 统计数据与统计学

### 一、统计数据的类型

数据是统计的起点，要进行统计分析，首先要获得相关数据。在学习统计学前，有必要了解统计数据的类型。

#### （一）按照所采用计量尺度的不同分类

统计数据按照所采用计量尺度的不同可划分为三种类型，即数值型数据、分类型数据和顺序型数据。

##### 1. 数值型数据 (metric data)

数值型数据是指直接使用自然数或度量衡单位进行计量的用数字尺度测量的观察值。

例如，GDP 等的宏观经济运行数据、每天进出海关的旅游人数、某地流动人口的数量等。数值型数据的表现就是具体的数值，统计处理中的大多数都是数值型数据。

##### 2. 分类型数据 (categorical data)

分类型数据是指反映事物类别的数据。分类属性具有有限个（可能很多）不同值，值之间无序。

例如，人口按性别分为男、女两类，受教育程度也可以按不同类别来区分。这种数据通常用频数（率）表示分类的结果，例如，某地区男性常住人口占 52%、女性占 48%。

##### 3. 顺序型数据 (rank data)

只能归于某一有序类别的非数字型数据，称为顺序型数据。

例如，满意度调查中的选项有“非常满意”“比较满意”“比较不满意”“非常不满意”等；再如，将学生的考试成绩分为优、良、中、及格、不及格等。

在这三类数据中，数值型数据由于说明了事物的数量特征，因此可归为定量数据 (quantitative data)，分类型数据和顺序型数据由于定义了事物所属的类别，说明了事物的



品质特征，因而可统称为定性数据（qualitative data）。区分数据的类型非常重要，这是因为不同类型的数据在一些情况下，需要用不同的统计方法进行处理。

统计数据按所采用计量尺度的不同分类如图 1-1 所示。

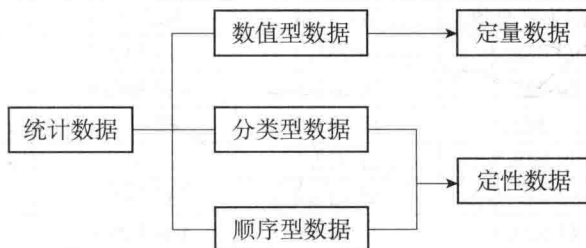


图 1-1 统计数据按所采用计量尺度的不同分类

## （二）按照收集方法的不同分类

按照收集方法的不同，可将统计数据分为观测数据和实验数据两类。

### 1. 观测数据（observational data）

观测数据是在没有对事物进行人为控制的条件下，通过调查或观测而收集到的数据。

观测数据主要集中在社会经济领域。例如，对商品零售价格变动水平进行测量可以得到商品零售价格指数，对股票价格变动水平进行测量可以得到股票价格指数等。

### 2. 实验数据（experimental data）

实验数据指的是通过在实验中控制实验对象而收集到的数据。

实验数据主要集中在自然科学领域。比如某种新型电池的使用寿命、一种新型降压药疗效的实验数据等。

## （三）按照是否与时间相联系分类

按照是否与时间相联系，还可以将统计数据分为截面数据和时间序列数据。

### 1. 截面数据（cross-section data）

在相同或近似相同的时间点上收集的数据，称为截面数据。

截面数据描述了现象在某一时刻的变化情况，它通常是在不同的空间上获得的数据，表 1-1 显示的数据就是截面数据。

表 1-1 2007 年我国部分地区国内生产总值及其构成数据 单位：亿元

地区	国内生产总值	第一产业	第二产业	第三产业
北京	9 353.32	101.26	2 509.40	6 742.66
天津	5 050.40	110.19	2 892.53	2 047.68
河北	13 709.50	1 804.72	7 241.80	4 662.98
山西	5 733.35	269.68	3 438.58	2 025.09
内蒙古	6 091.12	762.10	3 154.56	2 174.46

### 2. 时间序列数据（time series data）

在不同时间上收集到的数据，称为时间序列数据。

时间序列数据常用于描述现象随时间变化的情况。表 1-2 显示的数据就是时间序列数据。

表 1-2 2001—2007 年我国国内生产总值及其构成数据 单位: 亿元

年份	国内生产总值	第一产业	第二产业	第三产业
2001	109 655.2	15 781.3	49 512.3	44 361.6
2002	120 332.7	16 537.0	53 896.8	49 898.9
2003	135 822.7	17 381.7	62 436.3	56 004.7
2004	159 878.3	21 412.7	73 904.3	64 561.3
2005	183 217.5	22 420.0	87 364.6	73 432.9
2006	211 923.4	24 040.0	103 162.0	84 721.4
2007	249 529.8	28 095.0	121 381.3	100 053.5

## 二、统计学

统计学是一门收集、整理和分析数据的科学,有数据的地方,就有统计学的应用。收集数据并研究如何得到数据,与之对应的是统计学中的抽样调查和试验设计等理论;整理数据指的是将数据用图或表的形式展现出来,与之对应的是描述统计的方法;分析数据指的是选择适当的统计方法研究数据,并从数据中提取有用信息进而得出结论,更多地对应于推断统计的理论与方法,包含参数估计、假设检验、相关分析、回归分析、时间序列分析等诸多内容。

统计学的应用领域非常广泛,是一门适用于几乎所有学科领域的通用数据分析方法。无论是学术研究还是政府管理,无论是公司或企业的生产经营管理还是人们的日常生活,都离不开统计方法的应用。

表 1-3 列出了统计的一些应用领域,从中可以看到,统计无处不在,统计学是非常有实用价值的。

表 1-3 统计的应用领域

Actuarial work (精算)	Hydrology (水文学)
Agriculture (农业)	Industry (工业)
Animal science (动物学)	Linguistics (语言学)
Anthropology (人类学)	Literature (文学)
Archaeology (考古学)	Manpower planning (劳动力计划)
Auditing (审计学)	Management science (管理科学)
Crystallography (晶体学)	Marketing (市场营销学)
Demography (人口统计学)	Medical diagnosis (医学诊断)
Dentistry (牙医学)	Meteorology (气象学)
Ecology (生态学)	Military science (军事科学)
Econometrics (经济计量学)	Nuclear material safeguards (核材料安全管理)
Education (教育学)	Ophthalmology (眼科学)
Election forecasting and projection (选举预测和策划)	Pharmaceutics (制药学)
Engineering (工程学)	Physics (物理学)

Epidemiology (流行病学)	Political science (政治学)
Finance (金融)	Psychology (心理学)
Fisheries research (水产渔业研究)	Psychophysics (心理物理学)
Gambling (博彩业)	Quality control (质量控制)
Genetics (遗传学)	Religious studies (宗教研究)
Geography (地理学)	Sociology (社会学)
Geology (地质学)	Survey sampling (抽样调查)
Historical research (历史研究)	Taxonomy (分类学)
Human genetics (人类遗传学)	Weather modification (气象改善)

资料来源：贾俊平，何晓群，金勇进．统计学．北京：中国人民大学出版社，2007：4.

当今世界正进入大数据的时代，全球知名咨询公司麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域。”随着信息技术，特别是数据库技术的发展，社会各行业和领域收集、存储数据的能力有了很大的提高，同时积累了大量的数据，这为统计自身理论的进一步发展和在社会各领域的更广泛应用提供了机会。在知识经济和信息时代，统计学具有十分广阔的前景。



### 你知道吗?

在统计学产生和发展的过程中，出现了一批著名的统计学家，他们为统计学的发展做出了卓越的贡献，下面列举了一些历史上著名的国内外统计学家，在本书中会给出部分统计学家的小传。

- (1) 雅各布·伯努利 (Jacob Bernoulli, 1654—1705)。
- (2) 亚伯拉罕·棣莫佛 (Abraham De Moivre, 1667—1754)。
- (3) 托马斯·贝叶斯 (Thomas Bayes, 1701—1761)。
- (4) 卡尔·弗里德里希·高斯 (Carl Friedrich Gauss, 1777—1855)。
- (5) 弗洛伦斯·南丁格尔 (Florence Nightingale, 1820—1910)。
- (6) 卡尔·皮尔逊 (Karl Pearson, 1857—1936)。
- (7) 罗纳德·艾尔默·费歇尔 (Ronald Aylmer Fisher, 1890—1962)。
- (8) 许宝騄 (1910—1970)。

## 第二节

### 一些基本概念

#### 一、随机性和规律性

随机现象是相对于确定性现象而言的。在一定条件下，必然发生某一结果的现象称

为确定性现象。例如，在标准大气压下，纯水在温度降低到零摄氏度时会凝固等。随机现象则是指在基本条件不变的情况下，每一次试验或观察前，不能肯定会出现哪种结果，呈现出偶然性。许多社会现象是随机的，带有不确定性。例如，买一注“双色球”福利彩票，能否中奖是不确定的；在犁好的土地中播下 1 000 颗种子，有多少会发芽也是不确定的。也有许多社会现象是有规律的，比如，人口密度高的地方犯罪率通常也比较高。

随机试验的每一可能结果称为一个基本事件，一个或一组基本事件统称随机事件，或者简称事件。典型的随机试验有掷骰子、扔硬币、抽扑克牌以及轮盘游戏等。

现实中，随机性与规律性并非完全对立，社会现象通常是随机性和规律性的有机结合体，随机之中带有规律性。抛一枚均匀的硬币，可能出现正面，也可能出现反面，其结果是随机的。但是，随机之中又有规律，如果重复抛该硬币若干次，则硬币正面出现的频率会随着重复次数的增加逐渐趋于 0.5 这个数值。图 1-2 记录了 500 次掷硬币试验中正面出现频率的波动情况。在重复次数较少时，该频率波动剧烈；随着重复次数的增加，该频率波动的幅度逐渐减小，并向 0.5 靠拢。这种现象可以用统计学中著名的大数定律解释。

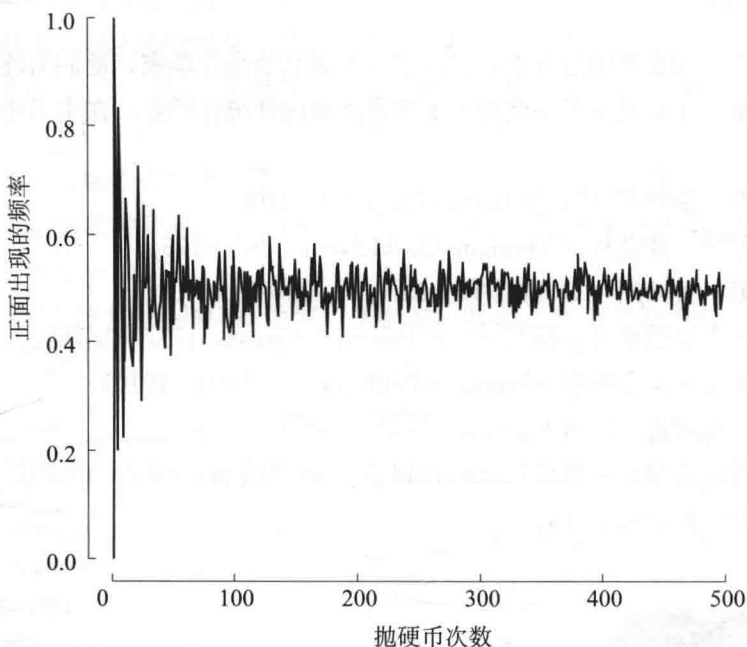


图 1-2 抛一枚均匀硬币，正面出现频率的变化趋势

类似的例子还有很多：你很难预测股票在某一天是上涨还是下跌，但是，作为国民经济的晴雨表，在经济向好的那些年份，股票价格指数总的趋势是上涨的；人的身高很难预先确定，受生活习惯、基因、体育锻炼等因素的影响，存在较大的个体差异，带有一定的随机性，但是，从总体来说，我国公民的平均身高是非常稳定的，并且随着生活水平的提高在逐步增加。这都说明了随机之中有规律，这种规律称为统计规律。

对统计数据进行分析，就是利用数据产生的随机性和统计规律进行推断和决策。例如，要比较甲和乙两所高中的英语教学水平，分别在两所学校各随机抽取 30 名学生进行试卷测试。一般情况下，两个学校学生的成绩会有差异，这种差异有可能来自抽取到水平不同的学生，也可能来自两所高中英语教学水平的不同。在很多情况下，我们会面临不同背景的观测数据，需要研究这两组不同背景下的数据是否来自同一随机现象，即需要研究两组数据之间的差异是否大到超过随机性本身所能解释的地步，这种研究的理论就是概率论 (probability theory)。

### 小词典

在随机事件的大量重复出现中，往往呈现几乎必然的规律，这个规律就是大数定律。通俗地说，这个定律就是，在试验不变的条件下，重复试验多次，随机事件的频率近似于它的概率。比如，我们向上抛一枚硬币，硬币落下后哪一面朝上本来是偶然的，但当我们上抛硬币的次数足够多后，如上万次甚至几百万次，我们就会发现，硬币每一面向上的次数约占总次数的二分之一。这就说明了偶然之中包含着某种必然。

## 二、概率

概率是对机会的描述，度量某件事情发生的可能性，其取值在 0 和 1 之间（也可能是 0 和 1）。概率为 0，对应那些绝对不可能发生的事情，比如在标准大气压下，水加热到  $60^{\circ}\text{C}$  就沸腾，这是不可能的；概率为 1，对应那些一定会发生的事情，比如在真空状态下，自由落体在经过  $t$  秒钟后，落下的距离  $s$  必定是  $gt^2/2$ 。现实中，概率为 0 或 1 的事件都比较少，绝大部分事件发生的概率都介于 0 和 1 之间，为随机事件。这样的例子有：

- (1) 随意抛掷一颗骰子，出现的点数为 6。
- (2) A、B 两队进行一场足球比赛，比赛结果为平局。
- (3) 在某交易日，上证指数以红盘报收。
- (4) 某对夫妇生下的是一名男孩。
- (5) 某天出现雷雨天气。

通过对这些随机事件以概率的形式进行表述，可以为进一步的统计推断奠定基础。

### 小词典

概率论是研究随机现象数量规律的数学分支。事件的概率是衡量该事件发生的可能性的量度。虽然在一次随机试验中某个事件的发生是带有偶然性的，但那些可在相同条件下大量重复的随机试验往往呈现出明显的数量规律。例如，多次测量一个物体的长度，其测量结果的平均值随着测量次数的增加，逐渐稳定于一个常数，并且诸测量值大都落在此常

数的附近,其分布状况呈现中间多、两头少的对称性。大数定律及中心极限定理就是描述和论证这些规律的。

现代概率论的主要分支有极限理论、随机过程、随机分析、应用概率论、金融数学等。



### 你知道吗?

设从均值为 $\mu$ 、方差为 $\sigma^2$  (有限)的任意一个总体中抽取样本量为 $n$ 的样本,当 $n$ 充分大时,样本均值的抽样分布近似服从均值为 $\mu$ 、方差为 $\sigma^2/n$ 的正态分布。

## 三、参数和统计量

统计学经常涉及参数和统计量的概念。为研究某一问题,需要对研究对象进行界定。在统计学中,将所要研究的全部个体(数据的集合)称为总体,其特征的一些概括性数字度量称为参数。例如,某一地区的平均受教育年限,一批袋装牛奶的合格品率等。

作为总体特征概括性数字度量,参数的种类可以有多种,但研究者通常关心的主要有以下几种:总体平均数(记为 $\mu$ )、总体方差(记为 $\sigma^2$ )、总体中具有某特征的个体所占的比例(记为 $\pi$ )等。

现实中,研究对象的范围虽然在大多数情况下容易界定,但是要把这些研究对象相关特征的数据全部收集到通常面临着很大的困难,涉及时间、人力、物力、财力等诸多因素。例如,想要知道全国的总人口数,若逐个进行统计,实施普查,花费的时间和费用是惊人的,现阶段只能是每十年进行一次人口普查。我们更多地采用抽样调查的方法获得感兴趣的总体参数。有时候,对应的研究对象可能是无限总体,即该总体所包括的元素是无限、不可数的,此时收集总体的全部数据变得根本不可行。例如,在科学试验中,可将每一个试验数据看作总体的一个元素,而试验可以无限地进行下去。为此,很自然的一个想法是我们不必去收集总体的全部数据,而是从总体中随机抽取一小部分元素的集合作为样本,根据样本提供的信息来推断总体的特征。再如,我国每隔五年进行一次百分之一的人口调查,从出厂的某批次灯泡中随机抽出几个检测其寿命等。

与参数相对应,我们称用来描述样本特征的概括性数字度量为统计量。它根据样本数据计算得出,是样本的函数。常用的统计量和参数类似,主要有样本平均数(记为 $\bar{x}$ )、样本方差(记为 $s^2$ )、样本比例(记为 $p$ )等。

统计学中的绝大多数问题都是研究如何根据统计量去推断参数。例如,如何用样本平均数( $\bar{x}$ )去估计总体平均数( $\mu$ ),如何用样本方差( $s^2$ )去估计总体方差( $\sigma^2$ ),如何用样本比例( $p$ )去估计总体比例( $\pi$ )等。图1-3形象地展示了这一过程。

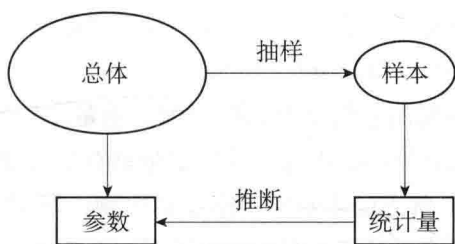


图 1-3 绝大多数统计问题的本质

## 小词典

抽样调查是一种非全面调查，它是从总体（全部调查研究对象）中抽选一部分单位进行调查，并据以对总体的一些参数（如总体均值、总体方差等）作出估计和推断的一种调查方法。显然，抽样调查虽然是非全面调查，但它的目的在于取得反映总体情况的信息资料，因而也可起到全面调查的作用。由于全面调查有测量误差等非抽样误差的存在，故抽样调查有时比全面调查获得的参数估计结果更加真实可信。根据抽选样本的方法的不同，抽样调查可以分为概率抽样和非概率抽样两类。概率抽样是从调查研究的总体中，根据概率论与数理统计的原理，采用随机原则来抽选样本，并从数量上对总体的某些特征作出估计推断，对推断出的可能出现的误差可以从概率意义上加以控制。非概率抽样是指从总体中非随机地选择特定的要素（单位）。

## 四、随机变量

### （一）随机变量与变量的类型

随机变量是相对于常量（也叫常数）而言的，是说明随机现象某种特征的概念。例如，某个小学中每天上学的人数是不同的，可能今天多几个，明天少几个，因为有些人生病，有些人逃学，还有一些人因为其他原因没有来，上课学生人数就是一个随机变量，但学校登记注册的学生人数则是常量，是一个固定的已知数目。

#### 1. 随机变量 (random variable)

随机变量是随机现象（在一定条件下，并不总是出现相同结果的现象称为随机现象）各种结果的实值函数（一切可能的样本点）。如未来某一天的气温、下一年度美职篮（NBA）总冠军等，都是随机变量的实例。

事实上，所有随机取值的数据都归属于某个随机变量，是随机变量的某一次具体实现。数据可划分为定性数据和定量数据两类，类似地，随机变量的类型也区分为定性变量和定量变量。

#### 2. 变量类型

##### （1）定性变量 (categorical variable)。

观测的个体只能归属于几种互不相容类别中的一种，一般是用非数字来表达其类别，

这样的观测数据称为定性变量。

(2) 定量变量 (quantitative variable)。

定量变量也就是通常所说的连续变量,如长度、重量、产量、人口、速度和温度等,它们是通过测量或计数、统计所得到的量。这些变量具有数值特征,称为定量变量。

性别、受教育程度等,都是定性变量;商品销售额、职工工资等,则都属于定量变量。统计学所面对的是随机变量和各变量之间的关系。

## (二) 变量之间的关系

事物是普遍联系的,作为随机现象某种特征的表达,变量与变量之间同样存在着千丝万缕的联系。与区分数据的类型类似,这里同样需要区分不同变量类型之间的关系,因为研究不同变量类型之间的关系,对应着不同种类的统计模型。

按照变量所属类型的不同组合,可以将变量之间的关系区分为定性变量之间的关系、定量变量之间的关系和定性变量与定量变量之间的关系。

性别与文化程度是否有关、不同国家的人民对陌生人的态度倾向是否存在差异、居民家庭订阅报纸和开通宽带上网之间是否有联系等,这些都属于定性变量之间的关系。研究定性变量之间关系的统计模型与方法主要有列联分析、对数线性模型等。

广告投入是否会影响销售额、城镇居民人均收入和人均支出相互影响有多大、复习时间和考试成绩之间是否存在必然联系等,这些都是定量变量之间的关系。研究定量变量之间关系的统计模型与方法主要有线性回归、非线性回归等。

手机品牌和手机销售量之间的关系、是否违约和信用卡用户年龄以及月收入之间的关系、上市公司所属行业与基本财务指标之间的关系等,都可归为定性变量与定量变量之间的关系。研究此类关系的统计模型与方法主要有方差分析、Logistic 回归、判别分析等。

上述统计模型与方法,有些是较为基础的内容,会在本书的相关章节予以介绍,比如列联分析、方差分析和回归分析;有些则属于更深层次的内容,需要参阅相关专业统计书籍。

## 第三节

### 统计与统计软件

随着科技的飞速发展和计算机的普及,原本枯燥、庞大的统计计算等工作,如今都可以通过统计软件完成,这为统计应用的普及提供了条件。

统计软件的种类有很多,这里介绍常见的几种:

(1) Excel。虽然严格来说,Excel并不是一款统计软件,但它自带了一些统计计算功能。Excel中设计了种类十分齐全的统计函数,并且通过加载宏安装数据分析的功能,能



够进行一些诸如方差分析、线性回归等简单的统计分析。

(2) SPSS。SPSS的全称是 statistical product and service solutions, 即统计产品与服务解决方案, 是目前非常受欢迎的一款统计软件。它囊括了各种成熟的统计方法与模型, 并提供了各种数据准备与整理技术。

(3) SAS。SAS的全称为 statistics analysis system。在数据处理和统计分析领域, SAS是一款权威统计软件。SAS以编程为主, 在编程操作时需要用户对所使用的统计方法有较清楚的了解, 非统计专业人员掌握起来较为困难。

(4) S-plus。S-plus是由美国 MathSoft 公司开发的一种基于 S 语言的统计软件, 是世界上公认的三大统计软件之一, 主要用于数据挖掘、统计分析和统计作图等。其最大的特点在于它可以交互地从各方面发现数据中的信息, 并可以很容易地实现一个新的统计方法, 兼容性好。

(5) R。R是一款国际自由统计软件, 由一群致力于推广统计应用的志愿者组织管理。它完全免费, 其统计功能的实现源自不断加入的由各个研究方向的统计学家编写的软件包, 是目前更新速度最快的软件。R同样需要编程, 但与 SPSS 和 SAS 中的编程语言相比, R语言是彻底面向对象的统计编程语言, 十分简洁和高效。R的官方网址是 <http://www.r-project.org>, 在这个网站上可以下载到各种程序包及相关资料。

(6) Eviews。在时间序列数据的分析和处理上, Eviews 是一款非常专业的软件, 擅长于多种常用的计量经济模型。它通过建立序列间的统计关系式, 实现预测和模拟等功能。该软件在科学数据分析与评价、金融分析、经济预测、销售预测和成本分析等领域应用广泛。

还有许多其他的统计软件, 这里不一一罗列。值得注意的是, 统计软件的使用必须建立在熟悉相关统计理论与方法的基础上, 否则容易误用。而学习统计软件的最好方式是在使用中学习, 并多看软件的帮助和说明。

## 人物小传

雅各布·伯努利 (Jacob Bernoulli, 1654—1705), 伯努利家族代表人物之一, 瑞士数学家。他被公认为是概率论的先驱之一。他是最早使用“积分”这个术语的人, 也是较早使用极坐标系的数学家之一。他较早阐明了随着试验次数的增加, 频率稳定在概率附近。概率论中的伯努利试验与大数定律也是他提出来的。雅各布·伯努利一生最有创造力的著作就是 1713 年出版的《猜度术》, 他在这部著作里提出了概率论中著名的“伯努利定理”, 这是大数定律的最早形式。这本书的出版是组合数学及概率论史上的一件大事。

