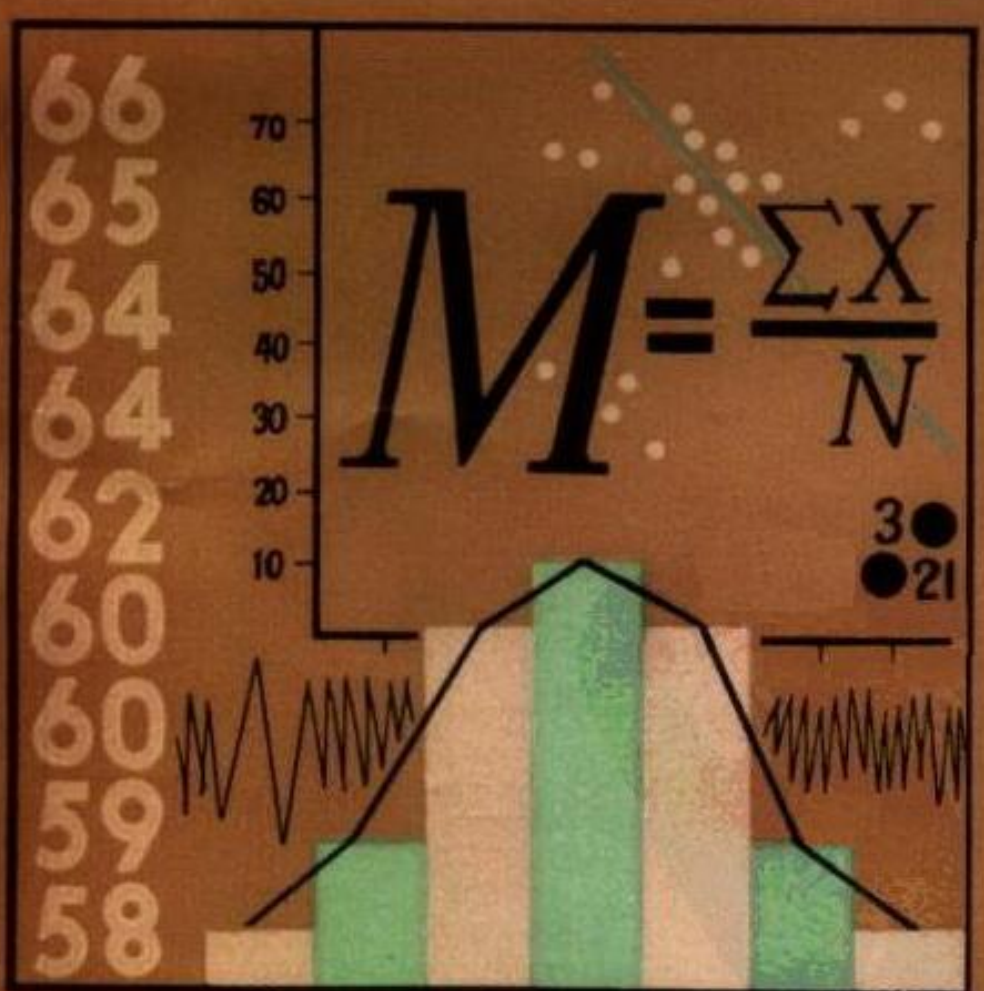
《自修数学》小丛书

统计世界

〔英〕D. A. 约翰逊
W. H. 格伦 著



科学出版社

《自修数学》小丛书

统 计 世 界

〔英〕D. A. 约翰逊
W. H. 格 伦，著

周焕山 惲简馨 译

科 学 出 版 社

1981

出 版 说 明

英国出版的《自修数学》小丛书（Exploring Mathematics on Your Own）是给具有中等文化程度的读者编写的一套近代数学通俗读物．该丛书自1964年初版后，于1974年、1976年多次再版印刷．为开阔读者眼界、增长数学知识，我们将选其中的一部分翻译出版，其目次如下：

大家学数学
测量世界
数型
毕达哥拉斯定理
统计世界
集合、命题与运算
数学逻辑与推理
曲线
拓扑学——橡皮膜上的几何学
概率与机率
向量基本概念
有限数学系统
无限数
矩阵

正好象
当我们学会驾驶汽车，
就能愉悦轻快地
驶向我们要去的地方；
而不必一步步走过去，
既费力又浪费时光。
你同样会看到
我们越是用数学武装，
生活就会越发舒畅。
利用数学这个工具，
事情做得顺顺当当。
如果光凭赤手空拳，
实现这些事本无指望。
数学就是这样一个法宝，
它能帮助我们的手足和大脑，
从而使我们人类，
发展得更加美好。

——利利安·阿·利伯

引自《米特的教育》

写 在 前 面

这本小册子向你介绍一个引人入胜的称为“统计学”的数学分支．希望你能从这本讲统计的书里，发现许多新奇的值得深究的问题，找到许多富有趣味的有用的知识．读任何一本数学书，只要你肯钻研，都可能象阅读一本惊险小说或侦探一个神秘洞穴那样趣味无穷．希望你在自修本书时，能饱尝发现许多新概念的乐趣．

你在阅读本书时，需要采用比阅读故事之类书籍更为认真的读书方法．从一开始就要慢慢地读．要养成面前放张纸、手中握支笔的习惯．一有疑问就要动手算一算，弄通了再读下去．如果你起初对一段话或一个句子不太明瞭，不要着急，而要有耐心．如果你能勤练习、勤画图、完成书中的作业，你将发现本书所介绍的统计知识是容易掌握的．

目　　录

一、数据与研究数据的科学

1.什么叫统计

"猫儿活到老，难改好奇心."这是一句古老的英国谚语.象猫一样,我们对现实世界都怀有好奇心.不同的是猫的好奇心并不能使猫聪明起来,也不能使世界发生什么改变,而人的好奇心却促使人们提出各种各样的问题，并进一步去思考问题和解决问题，去改变世界的面貌．比如每个人都可能提出这样的问题：明天会发生什么事？旋风是怎样形成的？怎样才能使宇宙飞船登上火星？全世界究竟蕴藏着多少石油?石油用完之后怎样解决能源问题？其它星球上有人类吗？原子能的广泛应用将会怎样改变世界的面貌？

诸如此类的问题不胜枚举，其中许多问题现在还无法回答．对于大多数问题来说,要解答它,必须通过调查研究,搜集与它有关的事实依据．我们搜集到的事实依据往往是数量依据,比如要知道一个短跑运动员的水平,搜集到的事实依据是他跑完一定距离的秒数．我们把搜集、记录下来的数量依据叫做**数据**,把研究数据的搜集、整理与分析方法的学问叫做**统计学**．但在口语中，常把统计学简称为统计，有时也把数据本身称为统计或统计资料，有时也把搜集有关某一问题的

数据的过程称为统计．因此，“统计”一词的具体含义要结合具体的语言环境来确定．许多问题都可用统计来解决．可用统计来解决的问题，其答案往往是以非确定性现象和不完全的情报作为依据的．所以，这样的答案只是大概正确的，有时把这样的答案叫做预测．以统计工作为职业的人叫做统计人员，统计人员必须知道较多的数学知识，并应用数学知识来提高结论的准确程度．

本书将向你介绍一些最基本的统计方法，你可以用这些方法去搜集数据，并利用数据去发现未知的规律和结论．

2. 统计学在现代和未来的应用

我们今天生活的世界，是一个迅速地变化着的世界．如果你看看报纸，听听广播，就会发现在报告新闻时十分频繁地使用着数字，只要是关于“如何多”、“如何快”、“何时”、“何地”、“怎样广泛”之类的问题，通常都要用数字去回答．所以，如果你学一点统计学，将能更好地理解世界上发生的事情．让我们来看几个现今如何应用统计学的事例．

如果你喜爱板球[1]之类的体育运动，你大概会关心体育比赛的统计数字．怎样了解一个板球运动员的击球水平呢？也许你看到，迈克尔击球水平是 45 分，而比尔是 35 分．这意味着迈克尔每场比赛的得分在 45 分左右，而比尔每场比赛的得

1) 板球是双方各十一人玩的英国特有的球类运动．——译者

分在 35 分左右．但这是否表明在任何一场比赛中，迈克尔的得分一定高于比尔呢？事实证明未必如此．也许在某一天的比赛中，比尔取得比迈克尔更高的得分．而在另外一天的比赛中，迈克尔的得分可能高于 45 分，比尔的得分可能低于 35 分．由此可见，在板球比赛这类非确定性现象中，对于运动员的水平所作的评价只是大概正确的．这是体育运动中应用统计的一个例子．

几乎不论在什么地方，你都可以找到应用统计的实例．例如天气变化是人们惯常的话题，而说起天气，通常是指温度、风、雨，或者日照强度等．这些情报在世界各地的气象站里都有记录．气象预报员是在研究过大片地区上搜集的气象资料之后才进行预报的．他在预报时明白，天气变化是一种非确定性现象，因而他的预报只是大概正确的，得冒失败的风险．这是应用数据进行预报的一个实例．

每个人都关心经济生活．在资本主义国家里，关于商业行情、物价、工资和失业人数等情报，遍见于报纸杂志．根据这些情报，人们决定何时买一辆新汽车，何时建造一所新房子，何时出卖牲畜，何时购买公债．工商界人士特别仔细地研究这些情报，以决定储存多少股票，何时建造工厂，何时出卖产品．政府部门也仔细研究这些情报，以便了解明年可望收入多少税款，有多少必要的开支，必须修建多少英里公路．教育部门也必须研究数据，以确定明年有多少学生和教师，何时何地建造新的学校，需要多少投资，等等．这些例子说明，许多事情都必须应用统计来作出决定．

人们以种种方式应用统计来改变我们的世界。医生研究实验数据，以研制新药品并试验它的效果。各种科学研究部门，竞相建造专门的实验室，以获得关于建筑材料、医疗药品、汽车引擎、火箭燃料和高速电子计算机等新产品的情报。而电子计算机的广泛使用又反过来为统计学的应用开拓了新的广阔的前景。由此我们看到，统计学在发现和试验新产品、新方法、新思想方面的重要作用和广泛应用。

虽然统计学是一门现代科学，但它的起源可追溯到很早以前。例如在《圣经》的《民数记》中就记载着许多数据。在十九世纪的英国，统计学常用于研究社会问题，因而当时被人们叫做"行政科学"。现在统计学几乎广泛应用于一切领域。人们以各种方式应用它，既用它来分析已经发生和正在发生的事情，也用它来预测未来可能发生的事情。

练习1　数据的应用

1．找几则应用数据描述下列情况的新闻或广告：

a．体育运动；　b．天气情况；　c．农作物收成；　d．商业行情；

e．政治；　f．科学；　g．教育；　h．贸易。

2．你单位有哪些问题需搜集数据后才能作出决定？

3. 用数据求答案

一所中学将迁移到离开市中心 3 英里的新校舍去。在迁入新校舍之前，就该校 300 名学生如何到校问题进行了一次

调查，结果得到下列数据：

步行	30 人
骑自行车	60 人
乘小汽车	10 人
坐公共汽车	200 人

根据这些数据，学校需要建造自行车停车棚，并承包七辆专用公共汽车。

这是学校当局与校舍设计者在制订计划时所需情报的一例。此外，他们还必须了解这一类问题：建造自行车停车棚需要多大面积？有多少学生回家吃午饭？跑回家要多少时间？关于这些问题都需要了解准确的数据。

上例中的一组数据，是就有关问题作某些重要决定的依据。在这个例子中，要得到包括学校中全部学生的数据并不困难，由于搜集起来的数据比较简单，即使不经整理使用起来也不困难。但是如果你要在一个有 2000 名学生的学校中进行同样的调查，恐怕就没有那么容易，而必须找一个搜集、分析和罗列数据的简捷方法。如果你在一个拥有许多学校的城市里进行调查，由于学生人数众多，这时恐怕你就只能从全城所有学校中抽出一部分学生进行调查。在统计学中，把所考虑的对象（物体、符号、分数、产品、人、生物等）的全体叫做总体，把每一个对象叫做个体，而把从总体中抽取的部分个体叫做样本。一个样本所含个体的多寡，称为这个样本的大小。在上面这个例子中，全城所有学校的全部学生组成总体，每个学生是个体，其中一部分学生就组成一个样本。从全城所有学

校中抽出一部分学生进行调查，用统计学上的术语讲，就是抽取样本进行调查，简称抽样调查。

如果你准备用抽样调查的方法去调查学生上学的交通工具，那么根据你所调查的学校的位置，该学校的总人数，调查那天的天气，以及在调查中学生回答问题的方式，可能得到不同的结果。这就是说，由于所取样本不同，结果也就可能不同，因此，这里就有一个选取适当的样本的问题。

你由此可以看到，从样本中得到的情报可能与实际情况有出入，因而是不完全可靠的。所以，根据抽样调查得到的情报做决定时，你必须考虑到是在何时、何地、用何种方式抽取样本的，样本有多大。即使考虑到所有这些因素，也无法消除和实际情况的差距。因而根据抽样调查作出的每个决定，都带有一定的不可靠性。虽然通常认为数学结论是非常可靠的，但以后我们将看到，数学将向我们提供根据不可靠数据做出决定的方法。

通过考察本节中的这道例题，让我们看看解一个统计问题需要哪几个步骤。

首先，我们确定所要解决的问题——本例中问题是"学生每天是如何到校(或回家)的？"其次，我们决定怎样搜集关于这问题的情报。是否需要抽取样本？如果需要，样本应有多大，应怎样抽取？然后，我们搜集必要的数据，记录下来。为了醒目起见，可把数据列成表，画出图形，再分析数据并解释数据的意义。最后，我们根据样本结果对总体作出结论。当这样做时，我们必须估计到这种结论的不可靠性以及可能冒

得到错误结论的风险.

由此可见，用统计解决问题将涉及许多方面. 在某些情况下,可能用准确的数据去解决问题. 在另外的情况下,必须抽取样本,因而结果具有不可靠性. 无论在哪种情况下,都得想方设法去整理、分析和表示数据，并解释数据的意义. 因此,为了学会并应用统计学,我们必须善于同数表、平均值、图形、样本等东西打交道.

练习 2　搜集数据

为了解决下列问题应搜集什么数据？哪些问题必须应用样本？并就怎样选取样本说出自己的想法.

1. 一辆长途公共汽车使用 10 公斤汽油能走多少公里？
2. 附近新办一所学校，应为学生饭厅订购多少张餐桌？
3. 为什么有些学生上课迟到？
4. 初中一年级学生每星期约看电视几个小时？

4.“数字不会说谎，但说谎者利用数字”

正如人们以种种方法应用统计学去探求新事物一样，也有人以种种手法滥用统计学去达到不正当的目的. 有时候实际上是在利用统计数据说谎. 有些广告用统计数字把一个产品说得神乎其神. 例如,一则牙膏广告说：“八个牙科医生中就有七个使用这种牙膏.”但是它永远不会告诉你,这八个牙科医生是怎样挑选出来的，他们到底是哪些人. 一种专卖药

品的广告说，实验证明，这种药在数分钟内可杀死数百万细菌。但这种药究竟杀哪一种细菌？说不定它杀死的倒不是病菌，而是无害的甚至是有益的细菌。我们还可以问一下这种药是在什么条件下杀死这些细菌的。即使日光也可在数分钟内杀死数百万细菌。因此，日光也许是比这种药更有效的杀菌剂呢！

假如有则广告说：10 个农民中有 7 个评论说，某种奶饼比另一种奶饼加倍有味。这时我们就要问：你是怎样衡量滋味好坏的？按照什么标准断定某种奶饼加倍有味？你的这 10 个农民的样本，是用什么方法在什么时间什么地点挑选出来的？

有人说，统计表明现代的青少年比过去的青少年惹出更多的麻烦，这些人经常引用关于犯罪和车祸的报道做证据。就算现代青少年确实表现不好吧！但是要断定他们比你的父辈在年青时更坏，就必须比较有关这两代人的数据。现代之所以有更多的青少年犯罪，也许仅仅因为现代青少年的人数比过去更多了。

如果你懂得一点统计学，你就不会被一些虚假的声明或狡猾的广告所愚弄。你就会懂得，用一个经过选择的小样本的优点来赞美一种产品，并不能证明这种产品真正优越。你就会知道，所谓“普通”公民的意见并非总是“正确的”意见。你就会知道，只有在对比过关于“两代人”的可以比较的数据之后，才能作出有关两代人的结论。无论什么时候，只要你看到引用统计数字的声明，你必须问这样几个问题：是谁搜集

这些数据的，数据的来源，搜集这些数据用的什么方法，并思考一下这些数据的真实意义。

练习3 分析数据

在问题1至问题5中，根据已知数据作出的结论错在哪里？

1．你在一年中上学多少天？

一年共有天数：	365
你每天睡觉至少8小时，即 $\frac{1}{3}$ 年：	−122
剩下：	243
你有52个星期六[1]和星期天：	−104
剩下：	139
你有7个星期的暑假：	−49
剩下：	90
你有圣诞节假，复活节假和寒假：	−50
剩下：	40
你每天吃饭至少花 $2\frac{1}{2}$ 小时，全年共计：	−40
所以你上学的天数是：	0

2．1960年死于飞机失事的人数多于1928年．所以，1960年乘飞机要比1928年更危险。

3．吉尔西种乳牛的产量比其它品种的乳牛多26%，所以，吉尔西种乳牛是最好的乳牛。

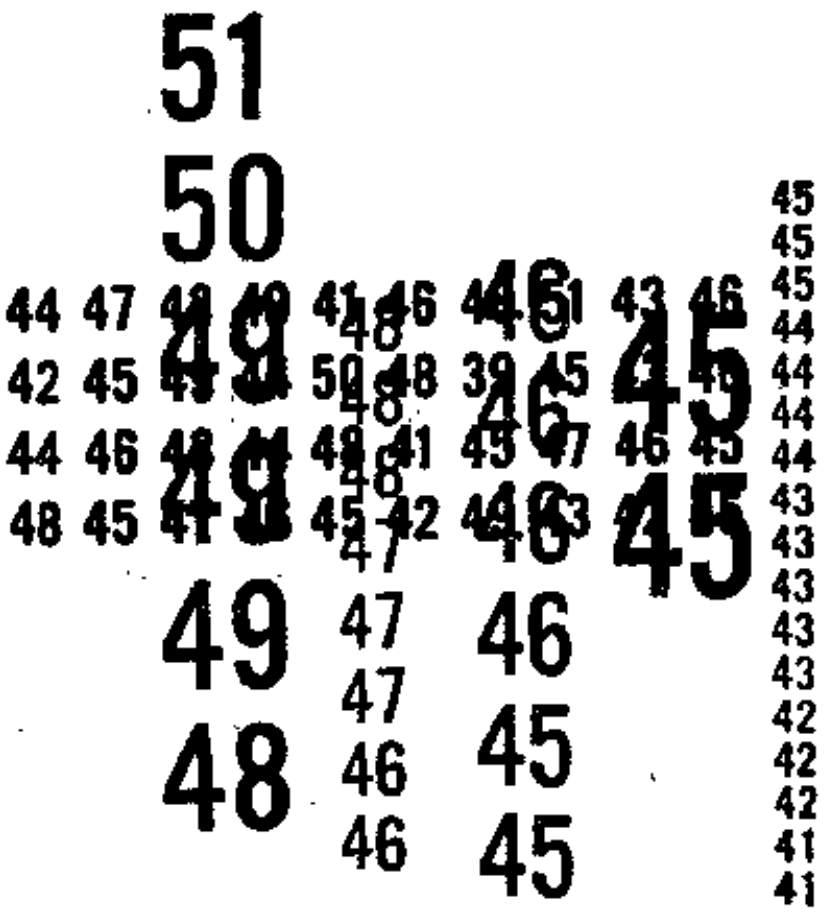
4．法国的车祸比德国的少．所以，在法国驾驶汽车要比在德国安全。

5．任何人服用某种药后，七天内伤风得以痊愈．所以，某种药是治疗伤风的特效药。

1）英国一些学校星期六不上课．——译者

6. 找出几则引用数据说明问题的新闻、论文或社论. 分析一下文中引用的数据是否清楚准确.

7. 找出几则涉及实验数据或统计数字的广告. 分析它的结论是否正确,或者是否有疑点.



二、数据的表示方法

1. 数 据 的 整 理

我们已经看到，在许多情况下可应用数据去解决问题，但同时也看到，数据也容易被曲解或滥用．因此，需要有某种方法和原则来帮助我们整理数据，并且作出解释．

假设你在一次数学测验中得到 45 分．这是高分还是低分呢？如果你不知道参加这次测验的其他同学的分数，你就难以回答这一问题．一个分数（或量值）本身不能说明什么，

我们要知道的是它与其它数值相比较是怎样的.

假设在这次数学测验中的全部分数如下:

44, 47, 43, 49, 41, 46, 48, 51, 43, 46,
42, 45, 49, 44, 50, 48, 39, 45, 43, 46,
44, 46, 43, 44, 48, 41, 45, 47, 46, 45,
48, 45, 47, 46, 45, 42, 49, 43, 42, 45.

即使有象这 40 个分数那样的一大批数字,当它们按上面这种方式排列时也不能说明多少问题. 仔细观察这 40 个分数,你可能找到最高分数和最低分数;但要更好地理解这组数据的意义, 你必须用某种方法重新整理. 方法之一是按照它们的大小顺序, 从最高到最低重新写出这 40 个分数, 并标出每一分数出现的次数. 表 1 中的斜线记号表明怎样标出每一分数出现的次数. 每一分数出现的次数叫做这分数的 频 数.

表 1 数学测验成绩

分数	次数记录	频数
51	/	1
50	/	1
49	///	3
48	////	4
47	///	3
46	~~////~~ /	6
45	~~////~~ //	7
44	////	4
43	~~////~~	5
42	///	3
41	//	2
40		0
39	/	1

这样一种数据一览表叫频数分布表。

上页中的分数总共有 40 个。数据的总个数通常用符号 N来表示。看了这张表，你对于你的 45 分在班级上的名次就能有一个比较清楚的概念了。

有时数据是相差很大的一些数值，这时必须把相近的数值并入一组。表 2 中所列数据，是一堂体育课上举重测验的成绩。

表 2　举重成绩（单位：磅）

129.6	60.3	80.2	93.0	100.0	136.4	92.1	122.6
91.8	102.0	129.4	99.1	103.6	78.0	91.0	139.1
92.0	87.6	72.9	153.4	75.5	114.0	82.3	98.0
93.0	61.9	86.0	81.1	96.0	75.6	84.2	108.3
48.0	107.7	72.3	81.4	100.2	28.5	103.6	95.2
83.3	62.1	102.1	96.0	77.7	96.6	57.6	118.2
56.6	38.1	63.3	64.4	81.0	66.6	88.1	
58.0	110.1	43.3	115.0	65.1	116.2	59.4	

因为这些数值分布很散，须要列很多项，因此将它们分组为好。组的范围称为组区间。组的大小称为组距。在本例中，组距看来以十磅为宜。表 3 就是根据表 2 中的成绩取组距为 10 个单位（磅）列出的频数分布表。

成绩按如下衡量规则列入表内。例如 129.6 磅这一重量，对于 129 磅和 130 磅来说，它更接近于 130 磅，因此它应列入 130--139 的区间内。同样，129.4 磅应列入 120—129 的区间

表 3　举重测验成绩（单位：磅）

重量区间	组　限	区间中点	频　数
150—159	149.5—159.5	154.5	1
140—149	139.5—149.5	144.5	0
130—139	129.5—139.5	134.5	3
120—129	119.5—129.5	124.5	2
110—119	109.5—119.5	114.5	5
100—109	99.5—109.5	104.5	8
90—99	89.5—99.5	94.5	12
80—89	79.5—89.5	84.5	10
70—79	69.5—79.5	74.5	6
60—69	59.5—69.5	64.5	7
50—59	49.5—59.5	54.5	4
40—49	39.5—49.5	44.5	2
30—39	29.5—39.5	34.5	1
20—29	19.5—29.5	24.5	1

内．区间 120— 29 的实际界限是 119.50—129.49．这通常写成 119.5—129.5．我们应将 129.5 这样的界限值列人较高的一组，即列人 129.5—139.5 这一组内．

组区间的中点常常用来作为该组的代表值．因为 150—159 的区间包括 149.5—159.5 的数值，因此这个区间的中点应是 149.5 和 159.5 的中间值，即 154.5．

数据一旦按区间列表，有些资料就反映不出来．例如在表 3 的区间 120—129 中有两个数值，我们只知道这两个数值

.在 119.5—129.49 之间，但却无法从表中知道这两个数值的准确值.

练习 4　制作频数分布表

1. 下面的数据是一次数学测验的分数，按照从高到低的顺序把它们重新排列. 中间的一个分数是什么？从上向下数第三个分数是什么？小于 32 的分数所占的百分数是多少？

19, 21, 36, 25, 32, 23, 28, 20, 34, 28, 31, 33.

2. 把下列数据列成频数分布表，但不要分组.

92, 96, 87, 93, 92, 90, 97, 89, 86, 90, 91, 95,
88, 90, 87, 91, 93, 90, 93, 92, 95, 91, 88, 90.

3. 下面的频数分布表给出六年级学生身长（单位：英寸）.

身长区间	次数记录
73 — 74	/
71 — 72	//
69 — 70	𝍸 /
67 — 68	𝍸
65 — 66	𝍸 ///
63 — 64	𝍸 𝍸 //
61 — 62	////
59 — 60	//
57 — 58	
55 — 56	/

a. 这些六年级学生中，哪个区间的身长最为常见？

b. 这张表可能记录到的最高身长是多少？

c. 量值 62.6 英寸应划入哪一组？

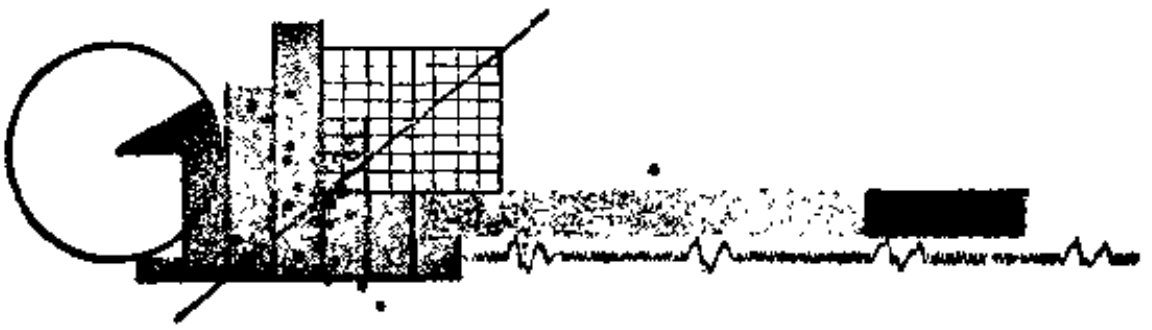
d. 身长低于 64.5 英寸的学生有多少个？

4. 抄下这张表并填充空白.

	记录区间	实际数值界限	组距	组中点
a.	73—77	72.5—77.5	5	75.0
b.	70—74			

c. 70—　　10

d.　　3　　75.0



2. 数据的图示

我们大多数人都觉得，图形所反映的资料要比文字材料所反映的更易于理解．一幅照片或一幅图形可能提供大量的资料．数学工作者用图形描绘数量关系正是基于这一事实．因为图形能形象地表示数据，并能清楚迅速地概括出数量关系，所以数据的图示在统计学中用处很大．

能够用来描绘数据的图形有好几种．包括圆形图、矩形分布图、柱形图、点频数图、直方图、折线图、频数多边形等．在用图形表示数据时，首要任务是选择适当类型的图形以表示手边的数据．然后必须给图形写上标题，并标出符号，以便识读并解释其意义．

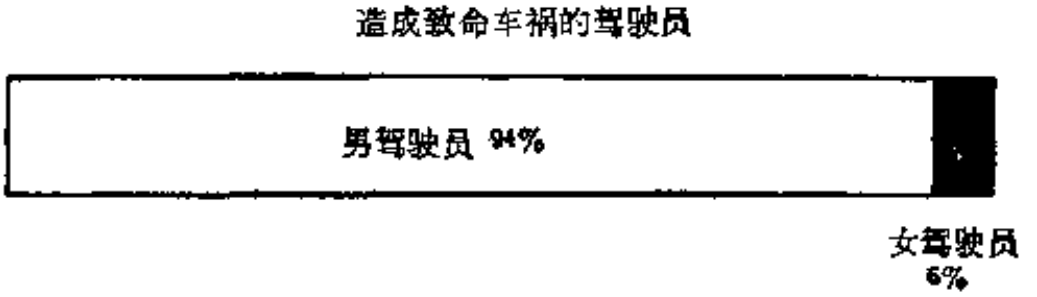
考察表 4 的数据．第三栏表明，在造成致命车祸的驾驶员中，男性占 94%，女性占 6%．这说明在造成致命车祸的每 100 个驾驶员中，94 个是男人，这表示出男驾驶员对驾驶员总数的比率．这种类型的比率可以用矩形分布图或圆形图

来描绘，如图 1 所示，其中每个区域的面积画成和已知数据成正比例．例如表示女驾驶员的部分都是总面积的 6%．

表 4　驾驶员事故统计表

	造成致命车祸的驾驶员人数	百 分 率
男　　性	36,700	94
女　　性	2,300	6
总　　数	39,000	100

矩形分布图



圆形图

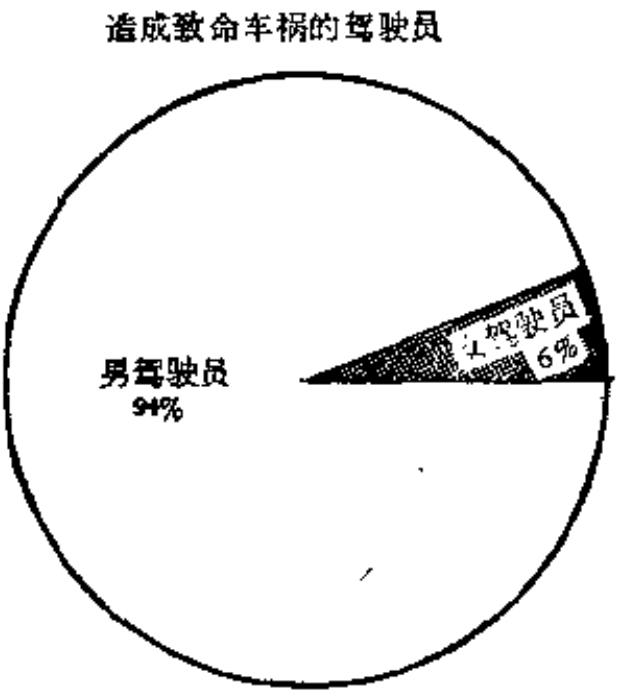


图　1

在圆形图中，要把这个比率转化为度数，才能对圆作适当的划分．例如，360° 的 6% 应为：

$$0.06 \times 360^\circ = 21.6^\circ \approx 22^\circ.$$

这样，圆中 22° 的弧上的扇形，就是用以表示造成致命车祸的女驾驶员的部分．

我们必须仔细地体会这些图形的意义．例如，上面的图形并不表明女驾驶员一定比男驾驶员好．如要得出那样的结论，我们还需要更多的资料．例如男、女驾驶员各自行驶的里程，驾驶的时间和地区，以及男驾驶员和女驾驶员的总数．

表 5 列出了四年级四个班的学生在一次数学测验中的平

表 5

班　级	平均成绩
1	25
2	20
3	12
4	9

均成绩．这类数据可以用柱形图表示出来，如图 2．

在柱形图中，每个柱子的高度应与相应的量值、分数或百分率的大小成比例．各个柱子的宽度应一致，每两个柱子之间的间隔应与柱子的宽度相等．水平标尺和竖直标尺（有时也叫横轴和纵轴）应当标注清楚、完整，水平标尺上标注的数字通常写在各个柱子底部的正中位置上．

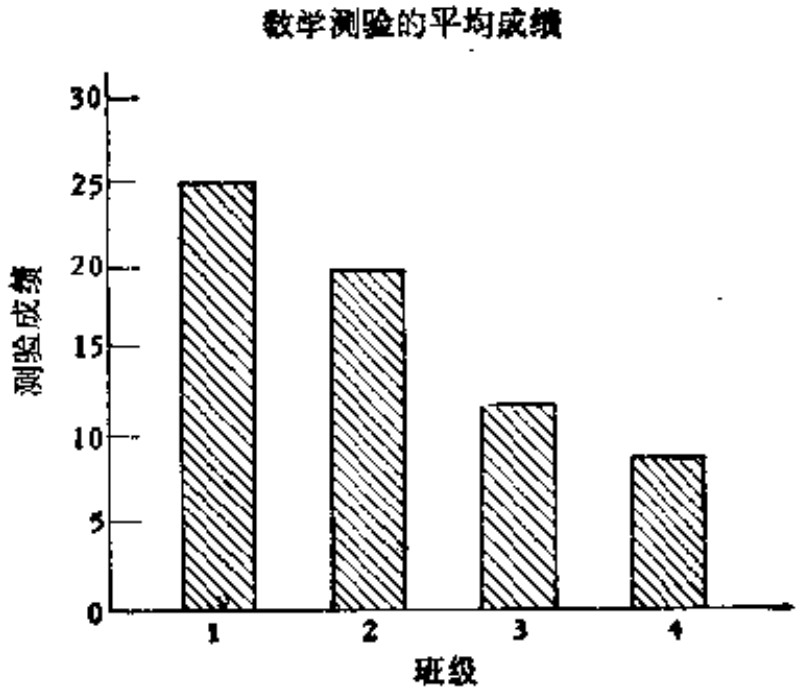


图 2

数据不分组的频数分布，可用点频数图表示. 例如表 6 列出了某班 21 个学生的数学等第的频数分布.

表 6

数学等第	频　数
A	3
B	5
C	9
D	2
E	2

表 6 中的数据可用图 3 来表示，在图 3 中，每个等第出现的次数用点来表示——1 点表示频数为 1，2 点表示频数为 2，等等.

英国的拉格比橄榄球协会的 38 个橄榄球队，在 8 场连续

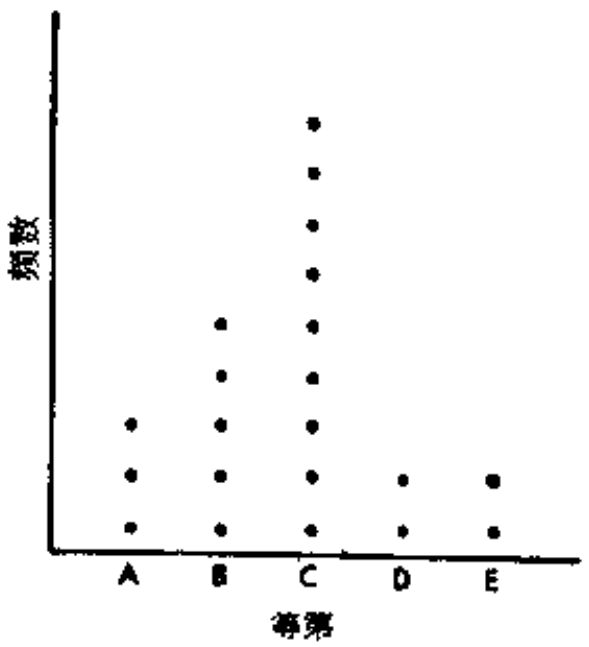


图 3

比赛中所得总分列成频数分布表，如表 7. 根据这一频数分

表 7

得分区间	组　　限	频数(球队数)
22—23	21.5—23.5	2
20—21	19.5—21.5	5
18—19	17.5—19.5	8
16—17	15.5—17.5	12
14—15	13.5—15.5	6
12—13	11.5—13.5	4
10—11	9.5—11.5	1

布画成直方图，如图 4. 直方图有些象柱形图，只是相邻两个柱子之间没有间隙. 直方图经常被用来作为频数分布的图

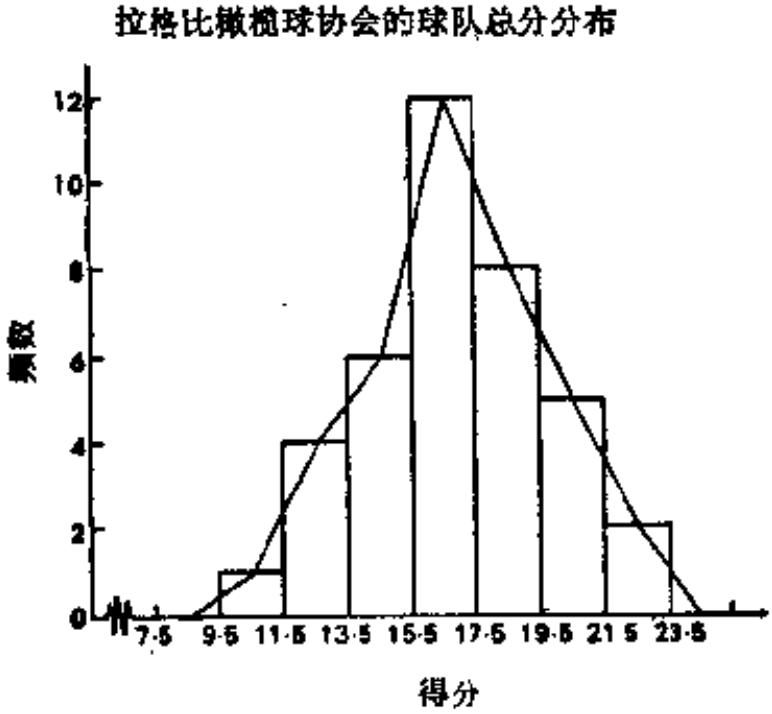


图 4

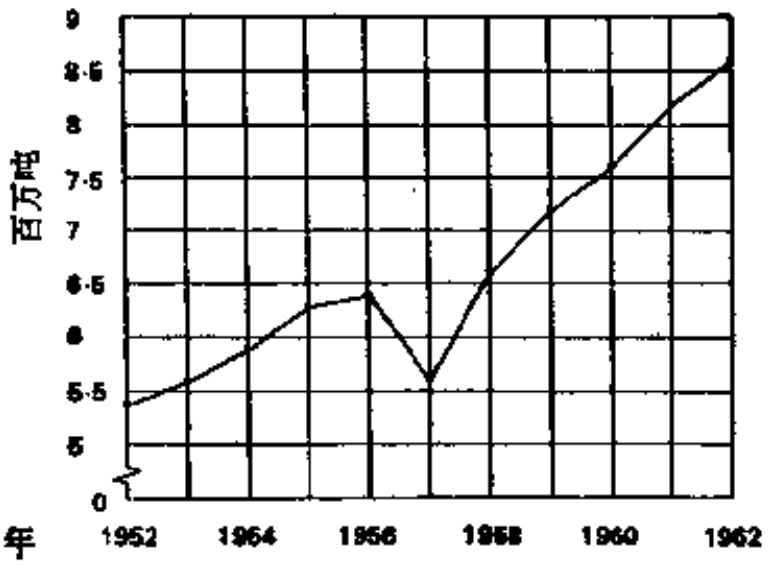
示．在直方图中，标度均匀的横轴上标出组限，纵轴上标出频数．例如，在图 4 中，横轴上标出组限 7.5，9.5，11.5，等等，而纵轴上标出频数 2，4，6，等等．这样，图 4 所示的直方图清楚地表明，14—15 这组的频数为 6，16—17 这组的频数为 12，等等．图 4 中横轴上的断裂号表示在 0 到 7.5 之间的距离是缩短了的．

频数多边形经常用来作为频数分布的图示．它的横轴和纵轴的标度与直方图相同．实际上只要连结直方图中每相邻两个矩形的顶部中点就可得到频数多边形，如图 4 中细实线所示．

许多种统计上用的折线图是由频数多边形演化而成的，练习 5 第 1 题和第 2 题的图形，就是其它一些类型的折线图的两个例子．

练习5 数据的图示

1．英国从1952年到1962年间的石油消耗量(单位：百万吨)，可用下图表示。



英国石油消耗图示

a．英国的石油消耗量增加最多的一年是哪一年？

b．为什么在1957年英国的石油消耗量显著减少？

c．英国1961年的石油消耗比上一年增加多少万吨（近似值）？增加的百分数是多少？

d．英国1962年的石油消耗比1952年增加的百分数是多少？

2．美元购买力的下降情况如下页图所示。图中以1945年的美元价值作为基数，以后各年的美元按照实际购买力折合成1945年的美分(1美元＝100美分)。

a．哪类商品的价格增长最快？哪类最慢？

b．总类的价格何时增长最快？可能的原因是什么？

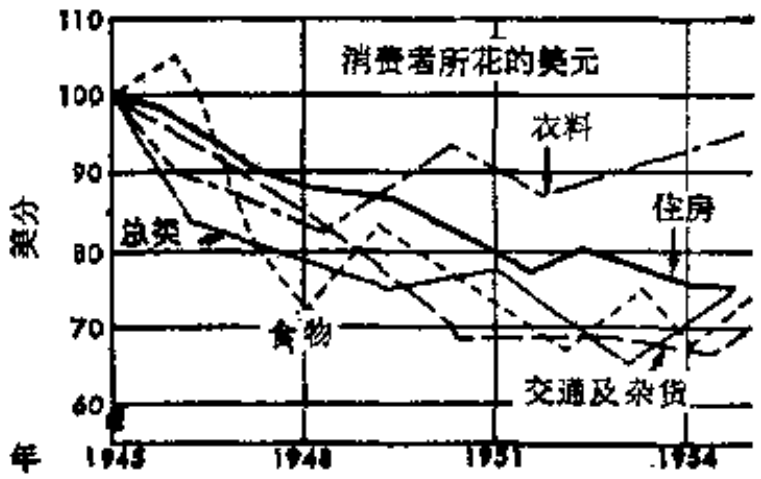
c．1945年花多少钱能买到1951年花一美元所能买到的衣料？

d．1945年价值90美分的衣料，在1951年要卖多少钱？

e．1945年价值一美元的衣料，在1951年要卖多少钱？

f．1945年造价为10,000美元的一所房子，在1954年的造价是多

美元购买力
1945-1956



少?

3. 根据下表中列出的 35 个测验分数,完成频数分布表,并制作表示这频数分布的直方图以及频数多边形。

21	26	21	20	23	24	22
19	24	26	25	23	26	29
21	24	19	25	26	25	22
25	27	23	26	24	25	30
25	23	27	24	28	28	28

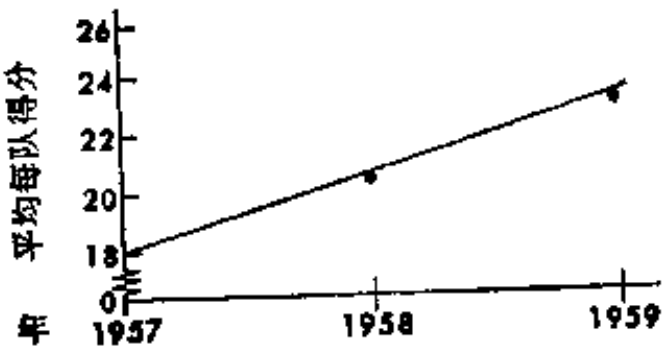
分数区间	组限	频数
29—30	28.5—30.5	2

3. "图形不会说谎,但说谎者利用图形"

虽然统计图的用处很大,但是它们有时也可能被用来制造假象。我们来看下面的例子。

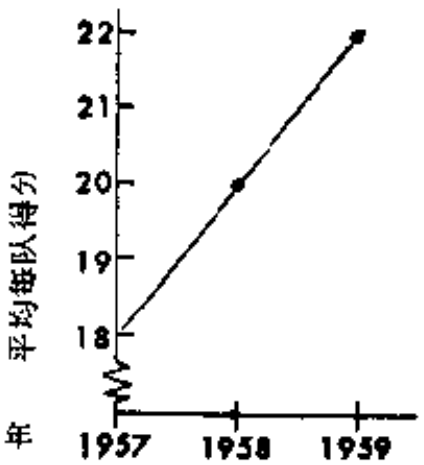
橄榄球比赛的历年平均得分是否大幅度提高了呢? 图 5、

图 6 所示的两幅统计图采用的是相同的数据，但却造成不同的印象。



得分增加——仅仅 10%

图 5



得分增加——不低于 10%

图 6

这两幅线形图，虽然都是说明橄榄球比赛的成绩在 1957—1959 年间每年递增 2 分，但因选择表示变化量的标度不同，所以画出两幅不同的图形，因而粗看上去似乎增长幅度有着显著的差别。这两幅图所用的标题也夸大了这种差别。

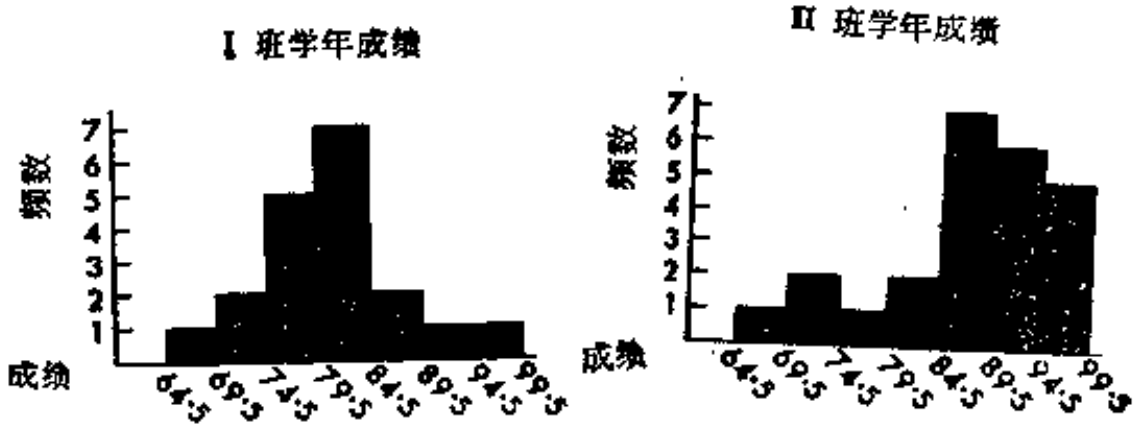
当使用统计图时，我们必须非常谨慎，以免误解。在达雷尔·赫夫所写的《统计学怎样被用来说谎》这本书里，举出了许多滥用统计图去说谎的例子。所以，你不仅应学会描画、识读统计图的方法，而且要学会正确地理解统计图的意义，以免被一些制作欠当的统计图弄糊涂。

练习 6　考察统计图

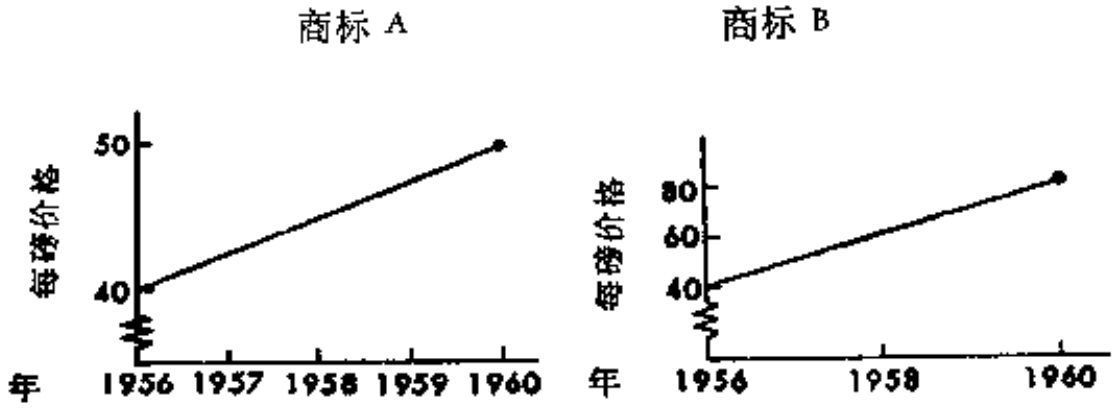
1．下列两幅统计图是否表明国家 A 比国家 B 花费更多的钱用于国防？

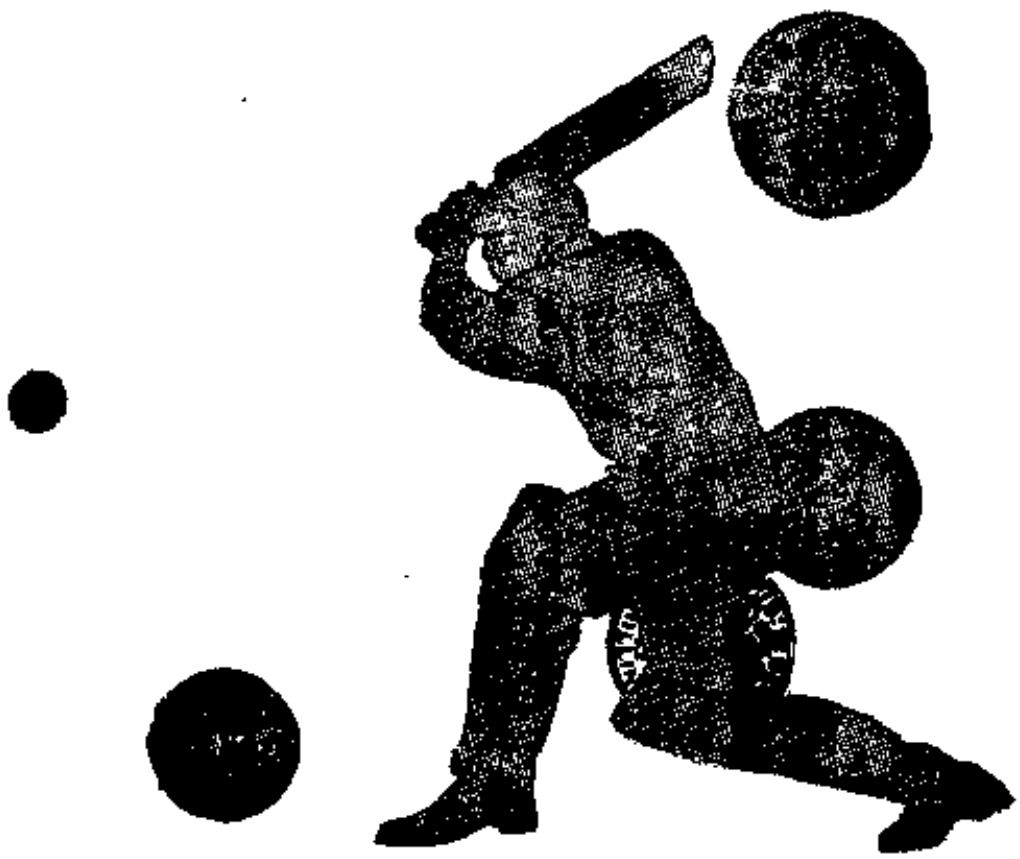
国防费占预算的比例

2．下列两幅直方图是否表明 II 班的学生比 I 班的学生成绩好些，为什么？



3．根据下列两幅图，说明哪种商标涨价更快？





三、数据的代表值

1. 水平指的是什么

“约翰在板球比赛中赢得 50 分．这个得分远远超过他的水平．”

“上一次数学测验中，我的成绩肯定达不到水平．”

“在英国，每周工资水平是 18 镑．”

在日常生活中，你会经常听到类似于上面的这些议论．这些议论中都提到“水平”这个词，但在上述三种场合下，这个词的意义恐怕未必相同．当看到“水平值”这个名词时，你大

概会联想起另外一些词语，比如“中等数值”、“可能性最大的数值”、“出现最多的数值”，等等。因为这些词语描述的都是我们所期望的最有可能出现的数值，所以有时把水平值称为“集中趋势的量度”。这样的数值“倾向”于刻划全部数据。正如口语中的“水平”一词有许多种解释一样，要求出水平值(或称代表值、典型值)也有几种方法。在统计学中，“水平”这个词一般是指三种不同的集中趋势的度量，即下面介绍的众数、中位数和算术平均数。

2. 众　　数

如果你要知道现在哪种牙膏最受顾客的欢迎，你就得去百货商店调查一下各种牙膏的销售情况。假设某百货商店上个月出售的各种商标的牙膏如表 8 所示。你从表 8 的数据，会得到结论：商标为 D 的牙膏现在最受顾客的欢迎，因为它出售的支数最多。在数据集合中，出现次数最多的那个数据叫做该数据集合的众数。

表 8

牙膏商标	出售支数
A	6
B	7
C	18
D	34
E	10

要找出一个数据集合的众数，只要把数据集合列成频数分布表，频数最高的那一项就是众数。

例如，表 9 列出了篮球运动员投篮得分的频数分布表，由表可见众数是8，因为有5个运动员得到8分，多于得到其它任何一个分数的运动员。

表 9

得　分	频　数
10	1
9	3
8	5
7	4
6	2

3. 中　位　数

某学校的校板球队的全体运动员，在一个季度里的得分见表10．哪一个分数是代表这个板球队"水平"的运动员的得分？

表 10

鲍勃　108	迈克　343	弗雷德　51
吉姆　92	乔　32	保罗　83
比尔　42	汤姆　50	戴维　112
哈里　47	沃尔特　71	

本例中如用众数作为水平值，恐怕是不行的．因为表 10 中的 11 个数据全都只出现一次，因而没有哪一个数据能够符合众数的定义．但是，如果把表 10 中的数据按照从大到小的顺序作如下排列：

343，112，108，92，83，71，51，50，47，42，32

然后把正中的一个数值（中间值）挑出来，我们就可找到能代表球队水平的运动员的分数了．因为这里共有 11 个数据，正中的一个是 71，它的两边各有 5 个数．这样选择的数据集合的代表值，即集中趋势的度量值，称为中位数．

要找出一组数据的中位数，首先要将这组数据按大小顺序依次排列，然后求出数据个数之半的值以确定中位数的位置．如果全部数据的个数是奇数，加 1 后除以 2，得中位数的位置是第几个数，然后从大数向下数，或从小数向上数，直至找到中位数．如果全部数据的个数是偶数，我们就取当中两个数值的平均数作为中位数．例如，13，14，15，18，21 这 5 个数据的中位数是 15．42，45，50，54，57，58 这 6 个数据的当中两个数据是 50，54，中位数就是它们的平均数，即 52．由此可见，中位数是容易计算的，而且它不受少数极端大或极端小的数值的影响．

4. 算术平均数

中学理科班的学生试验一种新型干电池．他们测试 9 只干电池的使用寿命，获得如下结果（单位：小时）：

20, 21, 22, 22, 22, 28, 29, 29, 32.

他们要表示被测电池的平均使用寿命. 他们不想用这组数据的众数或者中位数作为水平值. 他们认为, 平均使用寿命应是这样一个数, 这个数乘以 9, 应等于他们实验中所得 9 个数值的总和. 为了求得这个数, 他们将 9 个实验数据相加, 然后除以 9, 结果得到 25. 由于在求这个数值时, 将实验数据的每个数值, 都考虑在内, 所以, 同学们认为用它代表干电池的平均寿命比众数或中位数 (都是 22) 为好. 这种集中趋势度量值是简单的算术平均数, 通常就称为平均值. 平均值通常用字母 M 来表示.

如上所述, 要求数据集合的平均值, 只要将这些数据相加, 再除以数据的个数. 例如, 14, 17, 18, 19, 22 这 5 个数据的平均值是:

$$M = \frac{14 + 17 + 18 + 19 + 22}{5} = \frac{90}{5} = 18.$$

求平均值的公式是:

$$M = \frac{\Sigma X}{N},$$

这里的符号 Σ 是一个大写希腊字母, 读作"西格马". Σ 是一个求和的符号, 表示加法运算. X 通常用以表示单个数据, 如 14, 17, 18, 等等. 因此, ΣX 的意思是"求所有数据之和." N 表示该数据集合中数据的个数. 所以, 公式 $M = \Sigma X / N$ 的意思是: 数据集合的平均值等于所有数据之和除以数据的个数.

练习7　求几个简单的水平值

1．求下列各组数据的众数：

a. 2，3，4，7，5，7，3，3，10，4，9；

b. 24，21，20，25，21，27；

c. 3，3，1，2，2，3，2，4，5，4，3，4，2，3，3，4。

2．求题1中所给各组数据的中位数。

3．求题1中所给各组数据的平均值。

5. 由频数分布表求水平值

前面讲过，在整理大批数据时，频数分布表非常有用．所以，学会由频数分布求代表值，是很重要的．

对于数据不分组的频数分布，可用通常的方法去求众数和中位数．对于数据分组的频数分布，我们得用一些新方法去求众数和中位数．

在数据分组的频数表中，频数最大的区间的中点就是众数．

为了计算中位数，首先得找出中位数所在的那一组．然后通过数学计算，由这一组求出中位数的数值．

作为例子，我们来考察表11所给数据．

显然，众数是62，它是具有最大频数的那个区间的中点．

由于数据个数是25，所以中位数是第13个数．这个数在区间60—64内．从表的底部数起，我们发现在这区间之下

表 11

区　间	中　点	频　数
75—79	77	1
70—74	72	3
65—69	67	5
60—64	62	10
55—59	57	4
50—54	52	2
		$N = 25$

有 6 个数．须要再向上数 7 个数，才能达到第 13 个数．这意味着要达到第 13 个数，即须要达到这区间内 10 个数之中的第 7 个．这区间的长度是由 59.5 到 64.5，即 5 个单位．假设这 10 个数在这区间内是均匀分布的，那么为了达到第 13个数据，须要在这区间的下限 59.5 上加上区间长度的 7/10．所以，中位数的数值为：

$$59.5 + 5 \times \frac{7}{10} = 59.5 + 3.5 = 63.$$

对于数据不分组的频数分布，求平均值的简捷方法是将各数据分别乘以它的频数，将所得的这些积相加，然后除以数据的个数．表示这种方法的公式是：

$$M = \frac{\Sigma fX}{N},$$

式中 f 表示频数．

用这方法求平均值，可在表上计算，见表 12．

表 12

数据：8，7，8，9，6，10，8，9，6，9，8，7，10，8，7

数据 (X)	频数 (f)	fX
10	2	20
9	3	27
8	5	40
7	3	21
6	2	12
	$\Sigma f = N = 15$	$\Sigma fX = 120$

$$M = \frac{\Sigma fX}{N} = \frac{120}{15} = 8$$

对于数据分组的频数分布，在求平均值时把每组的各数据看做都与组中点的数值相同.表 13 就是这样的一个例子。

表 13

组 区 间	组中点 (X)	频数 (f)	fX
42—44	43	2	86
39—41	40	0	0
36—38	37	3	111
33—35	34	5	170
30—32	31	8	248
27—29	28	6	168
24—26	25	4	100
21—23	22	1	22
18—20	19	1	19
		$\Sigma f = N = 30$	$\Sigma fX = 924$

$$M = \frac{\Sigma fX}{N} = \frac{924}{30} = 30.8$$

当应用上述方法计算分组数据的平均值时，我们不能期望所得的结果与将各个数据相加后再除以数据个数所得的结果正好相同。因为我们曾假定各组数据都与组中点的数值相同，或者各组数据在组区间内是均匀分布的，而实际上这两个条件都是很难满足的，所以用这种方法计算分组数据的平均值，所得的结果只是近似值。但是，这样得到的近似值一般还是足够准确的，所以并不影响这种方法的应用。

练习8 求水平值

1．求下列各分布的众数：

a. 86, 82, 78, 93, 86, 84, 81, 90, 85, 79, 86, 85, 88, 81, 87;

b.

数据	频数
24	2
23	3
22	5
21	8
20	4
19	3
18	1

c.

组区间	组中点	频数
60—64	62	1
55—59	57	3
50—54	52	2
45—49	47	5
40—44	42	8
35—39	37	10
30—34	32	7
25—29	27	4

20—24	22	3
15—19	17	0
10—14	12	1

2. 求题 1 中所给各分布的中位数.

3. 求下列各分布的平均值($M = \Sigma X/N$ $N = ?$ $\Sigma X = ?$ $M = ?$):

a. 9, 12, 7, 6, 8, 11, 3;

b. −5, 3, −1, 0, 4, −7, −4, 2, −3, 6;

c. 在题 a 所给分布之后添上数据 32. 这时平均值变为多少? 这说明一个很大的数据将使平均值明显增大.

4. 应用公式 $M = \Sigma fX/N$ 求题 1 所给各分布的平均值:

a. $\Sigma fX = ?$	b. $\Sigma fX = ?$	c. $\Sigma fX = ?$
$N = ?$	$N = ?$	$N = ?$
$M = ?$	$M = ?$	$M = ?$

6. 三种水平值的比较

在调查一家工厂的工资水平时,这家工厂的年薪2,700镑的工厂主回答说:"我厂的工资水平是每年 934 镑."代表该厂工人的工会负责人说,工资水平是每年 800 镑. 而税务检

表 14

年薪 (X)	得到这种年薪的人数 (f)
2,700 镑	1
2,000 镑	1
1,500 镑	2
1,000 镑	3
900 镑	18
800 镑	23
700 镑	2

查员说，工资水平是每年 850 镑．这三种不同答复的根据都是表 14 的数据．那么他们各自所说的工资水平是指哪一种水平值呢？哪个值更有代表性呢？

我们来算一下三种集中趋势的度量值：

$$\text{年薪平均值} = \frac{\Sigma fX}{\Sigma f}$$

$$= \Big(1 \times 2{,}700 + 1 \times 2{,}000 + 2 \times 1{,}500 + 3 \times 1{,}000 + 18 \times 900 \div 23 \times 800 + 2 \times 700\Big) \Big/ \Big(1 + 1 + 2 + 3 + 18 + 23 + 2\Big)$$

$$= 934,$$

$$\text{年薪众数} = 800,$$

$$\text{年薪中位数} = \frac{800 + 900}{2} = 850.$$

由此可见，厂长讲的是年薪平均值，工会负责人讲的是年薪众数，而税务检查员讲的是年薪中位数．

这个问题中的三个度量值的具体意义可这样来解释：平均值或算术平均数表明，如果把工资总额平均分摊，则每个职工可得年薪 934 镑．众数表明，该厂人数最多的工资是每年 800 镑，而中位数则表明，有一半职工年薪在 850 镑以上，另一半职工年薪在 850 镑以下．这个例子说明，平均值、众数和中位数不仅意义不同，而且数值也往往不同．所以，我们在看到报道的某项水平值时，应想一想，“这是哪一种水平值？”“包括哪些对象？”以及“这数据准确到何种程度？”例如在上例中，

厂长等少数人的高工资使工资平均值显著提高．因此，如果指一般职工的工资水平，中位数或众数要比平均值更有代表性．

水平值的主要用途，在于从数据集合的所有数值中指出一个有典型意义的数值．

算术平均数或平均值，通常被认为是最佳集中趋势度量值，它的用途无疑也是最广的．它便于用一个公式来表示，并且在计算中考虑到每一个数据．但如果数据集合在一端含有少数极端数值（相对地说特别大或特别小的数值），平均值就可能没有代表性了．

例如：某中学一个由 5 人组成的集合，其平均年龄是 25 岁．其中四人是学生，一人是教师．试问 25 岁是否具有代表性？如果学生年龄都是 16 岁，教师年龄是 61 岁，那么取哪一种水平值比较有代表性？在这里我们看到，个别太高的数值对结果产生了显著的影响．

在科学研究与现实生活的许多方面都要用到平均值，例如：

在气象学中，为了得到平均温度或平均雨量；

在医学中，为了发现一种疾病的平均疗程；

在人类学中，为了测定某一群人在某一方面的平均特征；

在工商业中，为了估计平均工资、平均价格、平均指数，等等．

众数由于出现的频数最高，所以往往被认为是数据集合中最典型的一个数据．但是确定众数时，并不考虑数据集合

中其它数据的数值大小，这是用它作为代表值的缺点．确定众数一般是容易的，但是常常会遇到数据集合中有几个数据同时符合众数定义的情形，这时众数也就失去了作为代表值的意义．但是在有些情况下，只有众数才能作为合适的代表值．

例如：某一家服装店，它的设备只能制造一种尺码的男式运动衫，因此，它必须选择合适的尺码．起初它选用顾客购买的运动衫尺码的平均值作为它的产品的标准尺码，结果销路不大．为了扩大销路，它只好改用众数值（即最常见的尺寸）作为它的标准尺码．

中位数是位于中间的一个数值，它不受数据中高于或低于它的其它数值的影响．也就是说，只要某个数值低于中位数，对于中位数而言它究竟低多少是无关紧要的．但是，如果数据集中成明显不同且差异很大的几组，这时中位数可能就不适于作为集中趋势的度量值了．

例如：在一次满分为 30 分的测验中，某小组的成绩是 5 个 20 分，3 个 26 分，1 个 29 分．如果采用中位数作为集中趋势的度量值，那我们就得说这个小班的水平是 20 分．但是，由于这次测验的分数分布得很特殊，我们觉得用这个分数来代表这个小班的水平是不恰当的．在这情况下，平均值可能更适宜作为集中趋势的度量值．

在好些统计调查中要用到中位数，例如：

在人寿保险中，统计正常寿命多长；

在药物研究中，了解一种药品的效能；

在工业中，检验一种产品的质量。

练习 9　选择合适的水平值

1．由 9 家不同工厂出产的 6 英两重的罐装白漆的价格分别是：1/8（1 先令 = 12 便士，1/8 代表 1 先令 8 便士，以下类推），1/8，1/10，2/-（代表 2 先令），2/-，3/-，3/2，3/6，3/9．这种白漆价格的平均值和中位数分别是多少？你选择哪个代表值作为平均价格？为什么？

2．一所私立小学院，为了募集复兴基金，与 30 个往届毕业生联系后募集到的捐款如下：

钱数（镑）	频数
1,000	2
100	2
80	2
50	3
30	4
10	6
5	8
0	3

a．计算这群数据的三个代表值；

b．“通常的”捐款是多少？

c．哪个代表值最能代表这群数据？为什么？

3．让 20 个学生参加一次数学测验，以便确定哪些人编入快班，哪些人编入慢班。

a．如果得到的分数是：66，67，67，69，70，70，72，73，74，76，85，86，88，88，90，92，94，97，98，99，那么，选择哪个代表值作为编班的标准最好？

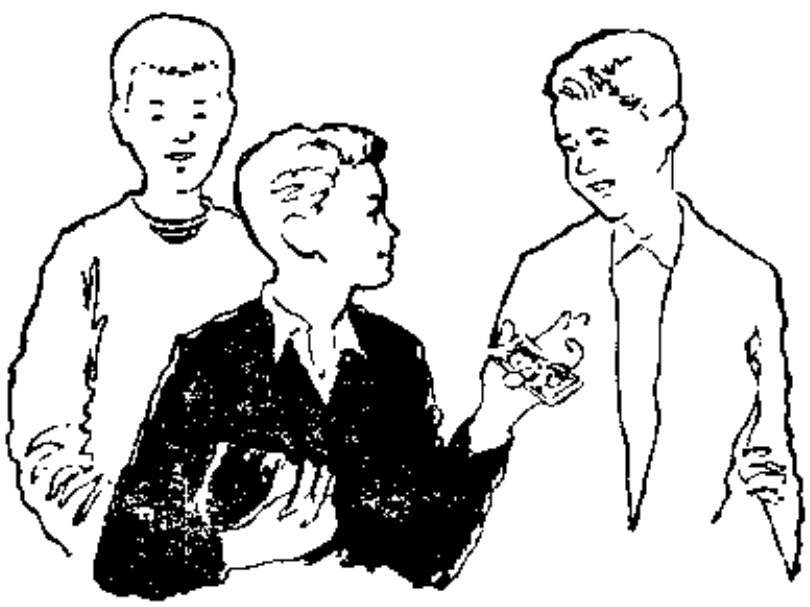
b．如果得到的分数是：62，63，66，66，66，66，67，68，68，69，70，87，89，90，92，95，98，98，99，100，那么，选择哪个代表值作为

编班的标准最好？

4．某轮胎厂检验它所生产的24只轮胎，以便确定能行驶多少英里不损坏，检验后得到下列结果：

英里	轮胎数
60,000	1
55,000	1
35,000	6
30,000	8
25,000	8

为了说明该厂生产的轮胎的平均寿命，选用哪个代表值最合适？为什么？



7. 怎样排名次

罗伯特身长5英尺.这次数学考试得56分．上星期六他去做工，挣得10个先令．那么，罗伯特的个儿算不算矮？他的数学成绩好不好？他做工得到的报酬低不低？仅仅根据上

述数据，这些问题都是无法回答的．上述数据必须和水平值或标准值相比较之后才有意义．比较的标准通常是由这个人所属的类别提供的．如果罗伯特只有 9 岁，那他应被认为是个儿高的．如果他那个班级的中等数学成绩是 60 分，那么，他这次数学测验的成绩就不算好．如果上星期六他仅仅在花园里劳动半个小时，那么，他所得的报酬应说是很高的了．

一个量值，只有当它与一群同类的量值作对比时，才能看出它自身的意义．给数据排名次，就是一种对比的方法．例如我们说比尔的成绩在他班上是第七名，这就是排名次的典型例子．但是，如果我们不知道全群的总数，这个名次就没多大意义了．如果比尔是在总共只有七人的小班里，比尔就是末了一名．如果比尔是在总共有 350 人的大班里，他的这个第七名简直可算是名列前茅了．因此，为了表示比尔的相对地位，最好是用比尔的名次与他所在班级的总人数作某种比较．让我们采用这样一种比较方法：将位于比尔之后的人数除以比尔所在班级的总人数．

如果比尔在具有 350 名学生的班级中名列第七，那么位于比尔之后的学生人数是 350 — 7，即 343．将 343 除以 350，得 343/350，即 49/50．我们把这个比值表示为百分数，得 98/100，即 98％．这时我们说，比尔在班上的超前百分数[1]是 98，意思是说在比尔的班上，每 100 个人中有 98 人名列比尔之后，或者说比尔在 100 人中超过 98 人．因此所谓超前百分

---

1) 原文是 percentile rank.——译者

数，是指列于某一个数值或等级之后的人在全群中所占的百分数。

假设罗伯特在一次数学测验中得 56 分，又参加这次测验的人数为 72．如果有 18 人的分数低于 56，则罗伯特的超前百分数是 $18/72 \times 100 = 25$，意思是说有25％的人的得分低于 56。

练习 10　百分数和超前百分数

1．全班 20 个学生在一次历史测验中的得分如下：2，5，6，7，8，9，10，10，11，11，11，12，12，12，13，13，14，16，17，18．

a．14 分的成绩在班上名列第几？

b．得分在 14 分之下的人数占全班的百分之几？

c．14 分的超前百分数是多少？

d．什么得分的超前百分数是 25？

e．什么得分的超前百分数是 40？

2．学生简的班上有 24 人，简名列第 5．问简在班上的超前百分数是多少？

3．学生查尔斯的班上有 24 人，查尔斯的超前百分数是 25．

a．有多少学生的成绩低于查尔斯？

b．查尔斯在班上是第几名？

4．下表列出了某个班级的学生身长（单位：英寸）的频数分布．

身长(英寸)	频数
66	1
65	2
64	1
63	3
62	5

61	8
60	7
59	9
58	4
57	5
56	1
55	3
54	1
	总数50

a. 身长64英寸名列第几?

b. 身长64英寸的超前百分数是多少?

c. 高于62英寸的学生占总数的百分之几?

d. 马克身长56英寸,马克的超前百分数是多少?

e. 高于58英寸且低于62英寸的学生占总数的百分之几?

5. 在共有100名学生的班级中,巴巴拉的超前百分数是75. 在同一个班级中,约翰是第75名. 这两个人中哪一个在班级中的名次较高? 抄写下面的表,并填充空白:

	名次	超前百分数
巴巴拉		75
约 翰	75	

6. 在共有100名学生的班级中,马乔里是第1名,彼得是第100名. 求他们的超前百分数并填写下表的空白:

	名次	超前百分数
马乔里	1	
彼 得	100	

7. 约翰和朱迪不在同一个学校里. 约翰说,他在班上是第10名. 朱迪说,她在班上是第20名. 如果你想判断约翰的在校成绩确实比朱迪好,你还必须知道一些什么其它情况?

四、预测结果

1. 抽样难

你曾经取一块鸡肉来尝尝鸡的味道如何吗？在吃饭之前，你不可能先把一只整鸡吃光，甚至不可能吃完一整份鸡肉，所以，你得取一个样本。同样，汤姆·索耶为了判断西瓜有没有熟，在西瓜上挖下一小块先尝一尝。一小块鸡或西瓜的样本，常能表明整只鸡或整个西瓜的味道怎么样。与此类似，

政府部门、工厂、学校、商店都采用样本对从中抽取样本的群体进行统计计算，预测结果，估计产品质量。你也许听到过关于电视节目评分、歌曲评奖、民意测验之类的活动吧，这些都是某些团体为了收集公众意见而进行抽样调查的例子。如果你仔细想一想，你就会发现我们所有的知识几乎都是通过样本获得的。我们知道山是什么样的，那是通过观察几座山，而不是全部山得到的。我们知道诗是什么样的，那是通过阅读某些诗而不是全部诗得到的。我们学会欣赏古典音乐，那是通过听一些音乐或唱片才学会的。

至此，我们的大部分统计作业都是对极为完整的数据进行描述、图示、编排和说明。在我们学过的大多数例题中，我们都假设我们有一份所考察的数据的完整清单。其实，这在许多情况下是不可能的。所以，统计工作者经常遇到选取样本的问题，以使样本能代表无法完全得到的更大的数据集合。

当我们根据样本获取情报时，总得冒导致错误结论的风险。样本并非总是与整体相符。比如，你取的鸡样可能只是一块胸肉；又比如读诗，你选的诗可能正好是语言无味的蹩脚作品；而你听到的音乐也可能演奏得并不带劲。但是在通常情况下，我们宁愿冒因抽样而导致错误结论的风险。没有什么人愿意为了判断煮熟的鸡是否味美，把整只鸡一下子吃掉！我们抽样是为了节约金钱、时间和人力。

所以在统计学中，我们得研究如何抽取样本以尽量减少导致错误结论的风险，并估计导致错误结论的风险究竟有多大。为了减少风险，我们用几个样本，用大样本（含有较多数据

的样本)，尤其注意用随机样本。随机样本是指凭机会抽取的样本。例如，你把你班上的每一个同学的名字分别写在不同的卡片上，然后把这些卡片放进一只纸盒里，把纸盒摇几摇，使卡片充分混杂。现在你不用看，伸手从纸盒里抽出5张卡片来，就能得到由你班上五个人名组成的随机样本。又如，你们在打纸牌时每次都要洗牌，目的在于让每个打牌的人都得到一个随机样本。同样，如果你要知道你所在市镇上的居民对某个问题的意见，你得先从该镇居民的全部名单中搞一个随机样本。否则，如果你仅从街上行人中，或从商店顾客中，或从在教堂做礼拜的人中抽取样本，那么，这样的样本就是某一类型居民的样本，而不是一般居民的样本或随机样本。如果你要了解哪种汽车最受公众欢迎，仅从卖福特牌汽车的人那里征求意见是不行的。这时你的样本应包括具有各种不同职业和不同经历的人们，既要包括拥有几辆不同牌号汽车的人，也要包括没有汽车的人，这样你才能听到各种不同的意见，从而使你最后得到的结论更有代表性。

因此，当你看到在抽样调查基础上得到的任何结论时，你必须问这样几个问题：

这样本有多大(即包含多少个数据)？

这样本是随机抽取的吗？

你能否把样本结论作为关于总体的结论？

2. 当心样本是否可靠

统计使用不当的一个常见的原因，是选取了不适当的样本作为统计调查的基础.

如果抽取样本时采用了不适当的方法，往往得到错误的结论.

例如：1936 年一家美国杂志预言，戈维纳·兰登将会当选美国总统. 但结果恰巧相反，他的竞选对手，富兰克林·D·罗斯福赢得了多数. 这家杂志的错误出在什么地方呢？它的结论是根据选前组织的一次民意测验得到的. 而参加这次民意测验的人是从电话簿上随机抽取的，但是 1936 年是"萧条"的一年：仅仅有钱人才能拥有一架电话机，所以这次民意测验依据的"样本"只是拥有电话机的有钱人的代表，而不是全体选民的代表.

如果把样本结论应用到与抽取样本的总体不同的另一个总体上，这时也可能产生错误的结论.

例如：假如根据在你校抽取的随机样本，说明有 76% 的学生想要上大学，由此能断定你县所有中学生的 76% 都想上大学吗？

如果所调查的情报属于人们不易记忆或不愿提供的那一类，这时也会得到不准确的样本结论.

例如：如果你去打听一家人在去年一年里一共买过多少夸脱(约合 1.14 升)的牛奶，你能得到准确的数据吗？要是你

去向家庭主妇打听她们的丈夫在赌博中输了多少钱呢?

练习11 关于样本

1. 找出几则利用样本作宣传的广告。

2. 结合你所在的市镇、县或民族，举几个根据样本得到结论的例子。

3. 为什么下列样本不是随机样本?

a. 足球队的成员;

b. 在一个商店里的人;

c. 住在你们市镇上的居民;

d. 任何大学的大学生。

4. 为什么以下列方式收集的样本不能包括各种意见的代表?

a. 向学校音乐队的全体队员征询意见;

b. 邀集商店老板进行调查;

c. 发函征求意见，并请函复;

d. 中午去旅馆访问所遇到的任何一个人。

3. 可能性和概率

"我们的橄榄球队平均每场比赛得 20 分。在星期六的比赛中，他们能得 20 分左右。"

"我的弟弟在数学统考中得 86 分。这意味着和他年龄相同的同学中仅有5%的人在数学方面超过他。"

在一个大城市中，询问了 10,000 名不同学校的学生，发现其中有 400 名学生要求额外补习。因为这城市的各个学校共有 50,000 名学生，教育局决定组织上夜课，给大约 2,000 名

学生补习。

象上述这一类的叙述是十分常见的。这些都是应用统计数据作决定或作预报的例子。

如果橄榄球队平均每场比赛得 20 分，这能意味着他们在下一场比赛中肯定会得 20 分吗？当然不能肯定。但是我们能够说，出现这种情况的机会是很多的。

一个学生在数学统考中得 86 分的成绩，能否意味着在全国同年龄学生中正好有 5% 的人超过他呢？ 当然不能肯定。但是根据英国历年来进行数学统考的统计资料来看，大概会出现这种情况。

如果在上面说的那个城市的各个学校作调查，能否表明正巧有2,000个学生需要补课呢？当然不一定如此。但是，如果学生样本选取得很适当，需要补课的人数很可能是2,000个左右。

机会，大概，可能——这三个词都表示不确定性。事实上我们已经看到，由于采用数值资料在许多情况下可能产生不确定性。我们注意到在使用数据样本时是要冒风险的，因此，概率或机会是极为重要的统计学课题。

如果将一枚硬币掷 100 次，正面朝上的会有多少次呢？当然，这实际上是无法得知的，但是我们认为在这一百次中出现五十次正面朝上的可能性最大。我们说掷得正面的可能性是 100 次中出现 50 次或 2 次中出现 1 次。我们知道，并不是每掷 2 次硬币就一定出现 1 次正面的，但多次抛掷时我们可预期会得出这样的结果。

可能性在数学中用概率来量度．概率的定义在数学中可以更严格地表述．这里只就一种特殊类型的概率[1]作一简单介绍．设某一试验有有限多个(个数用 t 表示)可能的结果，这些结果各不相同但发生的可能性是相等的．若某事件 A 由其中若干个（个数用 f 表示）结果组成，则事件 A 出现的概率 $P(A)$ 为：

$$P(A) = f/t$$

在掷硬币时，投掷一次有两个可能的结果（$t = 2$），但是对于通常的硬币来说，出现正面的事件包含其中一个结果（$f=1$）．所以，投掷一枚硬币时出现正面的概率是 1/2 或 0.5．如果使用一种两面都是正面的特殊硬币，那么掷得正面的概率就是 2/2，即 1．概率为 1，说明事件肯定发生．如果使用一种两面都是背面的特殊硬币，那么掷得正面的机会就是 0/2，即 0．概率为 0，说明没有成功的机会，即事件肯定不会发生．一般地，任何一事件发生的概率，都是大于等于 0 且小于等于 1 的一个数．即

$$0 \leqslant p \leqslant 1.$$

投掷一粒骰子[2]，一次掷得 3 点的概率是多

---

1) 这类概率在概率论(数学的一个分支)中称为古典型概率．由于结果的等可能性，它可以不经试验而直接由计算求得．在实际问题中严格的"等可能"是很难遇到的，即使拿常用的掷钱币、骰子的例子来说，由于各面花纹不同，"等可能"的假设也只是近似的．至于其它类型的概率，读者如有兴趣可以参阅概率统计方面的有关书籍．

2) 骰(音投)子，也叫色子，是一种骨制的小正方体，各个面上分别刻有 1 至 6 点．骰子是旧社会流行的一种赌具，数学书中常用它作为说明概率的模型．——译者

少？对于一粒骰子来说，投掷一次有 6 个可能的结果，出现 3 点是其中一个结果，因此，掷得 3 点的概率是 1/6.

一组分别标有 1，2，3，4 的四张卡片放在一只箱子里，另一组同样标有 1，2，3，4 的四张卡片放在另一只箱子里．如果你不用眼看，从每只箱子里各取一张卡片，那么你取得的两张卡片的数字和为 4 的概率是多少？要解决这个问题，我们必须知道抽出数字和为 4 的两张卡片有几种方式，以及总共可能有几种方式抽出两张卡片．表 15 对此作了说明．可能从第一箱中抽出的卡片列在表的左侧栏，可能从第二箱中抽出的卡片列在表的上部栏．由两箱中抽出的卡片的可能结合列在表内．

表 15

第一只箱子 \ 第二只箱子	1	2	3	4
1	1,1	1,2	1,3	1,4
2	2,1	2,2	2,3	2,4
3	3,1	3,2	3,3	3,4
4	4,1	4,2	4,3	4,4

由此可知，两张卡片的可能的结合方式共有 16 种，即从两个箱子里各取一张卡片，共有 16 个等可能的结果．其中数字和为 4 的可能结合方式是 (3，1)，(2，2) 和 (1，3)，即有 3 种方式．所以，抽得数字和为 4 的概率是 3/16.

上例中抽得的两张卡片的数字和为 3 或 4 的概率又是多

少呢？两张卡片的各种可能的结合方式仍然是16，不同的只是现在事件成功的条件是数字和为3或者4，由表15可以看到，数字和为3或4的可能结合方式有5种，即(1，2)，(2，1)，(3，1)，(2，2)和(1，3)。所以，这时的概率是5/16，或约0.31。由此可知，抽得的数字和为3或4的机会是百分之三十一，即31%。

练习12 概率的计算

1．某个班级有18个男生和12个女生。把这个班级的所有学生的姓名分别写在卡片上，每张卡片写一个名字，并把卡片放在箱子里充分混和。然后不用看伸手取出一张，取到女生姓名的概率是多少？

2．掷两枚硬币，可能的结果是正面-正面，正面-背面，背面-正面，背面-背面。试问：

a．两枚硬币都出现正面的概率是多少？

b．两枚硬币中出现一枚正面、一枚背面的概率是多少？

3．掷三枚硬币，有哪几种可能的结果？试逐一列举出来（总共有八种可能的结果）。

a．三枚硬币都出现正面的概率是多少？

b．三枚硬币中出现两枚正面、一枚背面的概率是多少？

4．将一粒骰子投掷一次。

a．结果可得多少种不同的点数？

b．仅有一点的面有几个？

c．仅有一点的面数与可能得到的结果的总数的比值是多少？

d．投掷结果得到一点的概率是多少？

e．投掷结果得到一点以外的其它点的概率是多少？

5．袋中放着3个白球和4个黑球，这些球除颜色之外其它方面都相同。你伸手从袋中摸出一个。

a．袋中球的总数是多少？

b. 白球数目与球的总数之比是多少?

c. 一次摸到一个黑球的概率是多少?

6. 从52张一副的扑克牌中抽到一张7方块的概率是多少?

7. 从52张一副的扑克牌中抽到任意一张方块的概率是多少?

8. 在52张一副的扑克牌中, A牌占总数的比率是多少? 从一副牌中抽一张抽到A牌的概率是多少?

9. 在52张一副的扑克牌中,人面牌(包括K、Q、J三种)的比率是多少? 因此,从一副牌中抽一张抽到人面牌的概率是多少?

10. 同时投掷两粒骰子,总共得到5点的概率是多少?

11. 同时投掷两粒骰子,总共得到3点或4点的概率是多少?

4. 衡量分散程度的统计量

诸如一组数据的平均值、中位数和众数之类的集中趋势度量值,只给出一个数值,这通常还不足以概括这组数据的全貌. 我们可用下面的例子来说明这一点,试考察表16的年薪数据,并求出各厂的平均年薪.

由计算可知,表中所列四个工厂的平均年薪都是975镑. 但是, 这四种分布明显地存在着巨大的差别. 最显著的差别是有的集中,有的分散. 因此,需要有一个衡量分散程度的统计量来说明一个分布的分散程度.

衡量分散程度的最简单的统计量是*极差*. 所谓极差, 是指数据集合中最大数据与最小数据的差. 它是分散程度的最粗略的度量值,因为它只取决于两个极端的数据,不能反映其余数据的状况. 例如, 在表16中, 工厂 A、B、D 的最高年薪与最低年薪是相同的,都分别是1200镑与750镑,因此,它们

表 16

年　薪	工厂 A 的职工	工厂 B 的职工	工厂 C 的职工	工厂 D 的职工
1,200 镑	1	5		3
1,150 镑	1			1
1,100 镑	1			1
1,050 镑	1			
1,000 镑	1		5	
950 镑	1		5	
900 镑	1			
850 镑	1			1
800 镑	1			1
750 镑	1	5		3

的极差也相同，都是 450 镑．但是你可以看到，这些工厂的年薪的分散状况存在着很大的差别，极差并没有反映出这种差别．虽然如此，极差在统计学中仍有许多用处，例如在气象资料中，经常引用温度的极差．

因为用极差来衡量分散程度有一些缺点，所以需要用一个较好的方法．由于平均值是最常用的集中趋势的度量值，因此我们想到试着搞一个与平均值有关的衡量分散程度的度量值．

假如你和你的朋友分别在人数相同的两个班级里，如果两班分别进行代数测验，结果各班平均分数都是 50 分．如果你得 70 分，而你的朋友得 74 分．那么，你的朋友在他班上的成绩是否比你在你班上的成绩名次更高呢？先弄清楚各个班级的全部得分关于平均值的分散程度，将有助于你回答这个

问题．如果你班上的大部分得分都在平均值附近，而你的朋友班上的大部分得分都离开平均值很远，那么你的 70 分在班上的名次恐怕要比你朋友的 74 分的名次高呢．

怎样才能得到数据集合关于平均值的分散程度的度量值呢？这里介绍一个很有用的方法．首先，列出数据集合的各个数据 (X)，并算出数据集合的平均值 (M)．然后将各数据减去平均值 $(X-M)$，得到每个数据离开平均值的差，叫做*离差*．因为有些离差是正数，也有些离差是负数，如果将全部离差相加，必然导致正的离差和负的离差相抵销．为了克服这个矛盾，先将各离差平方 $(X-M)^2$，再求出离差平方和并除以数据的个数．这样就得到各离差平方的算术平均数 $(\sum(X-M)^2/N)$．最后，为了消除平方后数值增大的影响，再求出这个算术平均数的平方根 $\left(\sqrt{\frac{\sum(X-M)^2}{N}}\right)$．这个值表示出该数据集合中的各数值偏离平均值的程度．

衡量分散程度的这种度量值叫做*标准差*．它给出数据集合中各数值偏离平均值的趋势的大小．如果标准差相对来说比较小，表明这群数据大多集中在它的平均值的附近；如果标准差相对来说比较大，表明这群数据总的来看离开平均值的距离较大，比较分散．

标准差通常用一个小写希腊字母 σ（读作西格马）来表示，因此求标准差的公式为：

$$\sigma=\sqrt{\frac{\sum(X-M)^2}{N}},$$

式中：X 代表每一个数据，M 代表数据的平均值，N 代表数据

的个数。

现在让我们根据表 16 提供的数据，来计算工厂 A、B、C、D 的职工年薪的标准差。为了方便，我们用符号 σ_A、σ_B、σ_C、σ_D 分别表示工厂 A、B、C、D 的职工年薪的标准差，先求它们的平方值，然后再开方：

$$\because\ \sigma_A^2 = \Big[(1200-975)^2+(1150-975)^2+(1100-975)^2$$
$$+(1050-975)^2+(1000-975)^2\Big]\Big/\Big[10\Big]$$
$$+\Big[(950-975)^2+(900-975)^2+(850-975)^2$$
$$+(800-975)^2+(750-975)^2\Big]\Big/\Big[10\Big]$$
$$=\frac{(225^2+175^2+125^2+75^2+25^2)\times 2}{10}=20{,}625;$$

$$\therefore\ \sigma_A=\sqrt{20625}=143.6;$$

$$\because\ \sigma_B^2=\frac{(1200-975)^2\times 5+(750-950)^2\times 5}{10}$$

$$=\frac{225^2\times 5+225^2\times 5}{10}=225^2$$

$$\therefore\ \sigma_B=\sqrt{225^2}=225;$$

$$\because\ \sigma_C^2=\frac{(1000-975)^2\times 5+(950-975)^2\times 5}{10}$$

$$=\frac{25^2\times 5+25^2\times 5}{10}=25^2$$

$$\therefore\ \sigma_C=\sqrt{25^2}=25;$$

$$\because\ \sigma_D^2 = \Big[(1200-975)^2\times 3+(1150-975)^2$$

$$+(1100-975)^2+(850-975)^2+(800-975)^2\Big]\Big/$$

$$\Big[10\Big]+\Big[(750-975)^2\times 3\Big]\Big/\Big[10\Big]$$

$$=\frac{225^2\times 3+175^2+125^2+125^2+175^2+225^2\times 3}{10}$$

$$=\frac{(225^2\times 3+175^2+125^2)\times 2}{10}=39{,}625;$$

$$\therefore\ \sigma_D=199.1.$$

所以，A、B、D这三个工厂的年薪的极差虽然相同，但是标准差却有显著差异，工厂B的标准差最大，其次为工厂D，再次为工厂A，反映出这三个分布的分散程度是不同的．工厂C的标准差比这三个工厂要小得多，说明工厂C的年薪数据的分散程度最小．

在应用公式计算标准差时，采用列表法比较清楚明了．例如要计算下述数据：

11，12，13，14，15，16，17

的标准差，先求平均值：

$$M=\frac{\sum X}{N}=\frac{11+12+13+14+15+16+17}{7}=\frac{98}{7}$$

$$=14,$$

然后列出表17，计算离差的平方和．再计算标准差σ：

$$\sigma=\sqrt{\frac{\sum(X-M)^2}{N}}=\sqrt{\frac{28}{7}}=\sqrt{4}=2.$$

表 17

数据(X)	离差 $(X-M)$	离差平方 $(X-M)^2$
17	3	9
16	2	4
15	1	1
14	0	0
13	−1	1
12	−2	4
11	−3	9
		$\Sigma(X-M)^2=28$

在有些问题中，我们可以应用标准差进行比较。例如，把你的体重与同年龄的其他人的体重作比较，假设你比同年龄人的平均体重重 10 磅，那么，在 100 个和你同年龄的人中，大概还有多少人比你更重呢？你要回答这个问题，首先要知道和你同年龄的人的平均体重与标准差。如果和你同年龄的人的平均体重是 100 磅，标准差是 5 磅。那么你的 110 磅体重，相当于比平均值重两个标准差 (2 × 5 磅)。在你学过较多的统计学知识以后，你就会知道，这意味着在通常情况下，在 100 个和你同年龄的人中，仅有两人比你更重。

让我们应用标准差来讨论一些关于得分的问题。如果一次物理测验的平均成绩是 42 分，标准差是 6 分。如果你的得分是 48，你就比平均成绩高 6 分，即比平均值高一个标准差。假如在另一次化学测验中平均成绩是 25 分，标准差是 10 分。如果你在这次化学测验中得 35 分，你就比平均成绩也高一个标准差 (10 分)。那么，你可以说在这两次测验中你取得了一

样好的成绩，因为每一次都比平均值高一个标准差。

想一想：假设理查德在物理测验中得 30 分，那他就比平均成绩低 12 分，即低两个标准差。如果他在化学测验中的成绩象物理成绩一样差，那么他在化学测验中得多少分？简恩在物理测验中得 45 分，在化学测验中得 35 分，那么简恩在哪一次测验中成绩好些？

练习 13　分散程度的测度

1．计算下列两组数据的标准差：

a．1，2，3，4，5，6，7；

b．6，7，8，9，10，11，12，13，14。

2．伦敦学校在一次拼音测验中，平均得分是 54，标准差是 8．抄下并完成下面的表，把得分的离差换算成标准差的个数。

学生	得分	离差	以标准差为单位表示离差
马克	62	$62-54=8$	$\frac{8}{8}=1$
彼得	46	$46-54=-8$	$-\frac{8}{8}=-1$
苏普	70		
凯	50		
查尔斯	54		
简	66		

3．两个星期以后，同一个学校在另一次拼音测验中，平均成绩是 28 分，标准差是 10。抄下并完成下面的表。

学生	得分	离差	以标准差为单位表示离差
马克	38		
彼得	28		
苏普	43		
凯	18		

查尔斯	33
简	23

a. 就标准差说，哪个学生的成绩提高得最快？

b. 就标准差说，哪个学生的成绩下降得最快？

c. 就标准差说，哪个学生的成绩保持不变？

4. 用符号 f，X，M 和 N 写出求频数分布表的标准差的公式。

5. 下表是将满分定为 10 分的一次测验成绩分布表。抄下并完成这张频数分布表，并根据所给数据计算平均值和标准差。

得分 (X)	频数(f)	fX	$X-M$	$(X-M)^2$	$f(X-M)^2$
0	1	0			
1	0				
2	1				
3	1				
4	3				
5	3				
6	5				
7	8				
8	2				
9	2				
10	1				

5. 根据离差度量值求概率

某青年俱乐部筹措体育基金。这俱乐部里有一位青年认识儿童乐园的工作人员，并从他那里弄来一个装有一千余张圆片的袋子，每张圆片上有一个数码（从 1 到 7 的整数）。俱乐部打算把它用来摸彩。为了决定哪一数码应得最佳奖，他

们必须知道袋中数码的分布情况.但是他们得不到这资料,又不愿意察看每一个数码,所以他们决定进行统计试验,以了解数码的分布情况. 他们于是从袋中抽出 64 个数码, 这 64 个数码的分布如表 18 所示.

表 18

数 码	频数(各数码的个数)
1	1
2	6
3	15
4	20
5	15
6	6
7	1

表 18 表明,数码 1 有 1 个,数码 2 有 6 个,等等.

由计算可知, 这 64 个数码的平均值是 4, 标准差约等于 1.2.

如果俱乐部的人能够肯定, 这一个样本正好代表袋中数码的分布, 那么就可以利用样本分布去估计从口袋中抽到某一数码的概率. 例如, 上述数据表明, 从口袋中每次抽一数码,每抽64次只有 1 次机会抽到 7,所以抽到 7 的概率可估计为 1/64, 即 0.016. 同样, 抽到 1 的概率也是 1/64.

从口袋中抽到 6 或 7 的概率可以这样来估计: 样本分布表明,每抽 64 次有 6 次机会抽到 6, 有 1 次机会抽到 7, 因此能抽到 6 或 7 的机会共有 7 次,即抽到 6 或 7 的概率是 7/64,

即 0.11.

表 18 的数据可用直方图来表示，见图 7. 在图 7 中还画了频数多边形，是通过连结相邻两矩形的顶部中点作成的. 设想在另外的情形下直方图由很多矩形组成，每矩形的宽度很窄，那时画出的频数多边形将会更接近于一条光滑的曲线，图 7 中画的一条曲线就代表这样的曲线. 我们看到，这条曲线中间高、两边低，近似于钟形. 这种形状说明，在数据的平均值附近，频数最高；离开平均值愈远，频数也就愈低.

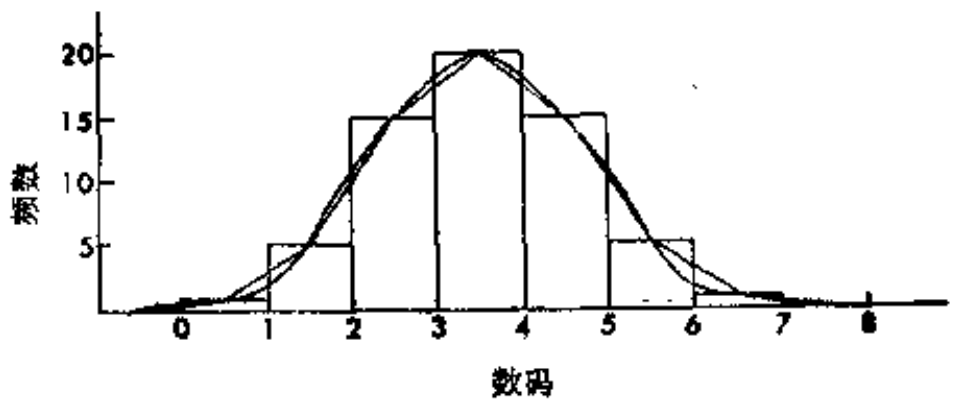


图 7

表示数据的无穷集合之分布的完美钟形曲线，叫做正态曲线. 正态曲线所表示的分布叫做正态分布（关于它的严密的定义和理论，在这本小册子里不作研究）. 表 18 的分布大体上接近于正态分布，因此，如果口袋中所有数码的分布和所抽取的样本的分布基本相同，那么它也近似于正态分布.

正态分布有着广泛的应用. 有许多实际问题的数据分布都很接近于正态分布的模型. 例如，如果你搜集到关于几千个六岁儿童的身长的数据，关于几千个苹果的每只苹果重量的数据，关于同一架机器生产的几千只同种零件的厚度数据，

或者关于参加统一数学测验的几千个学生的得分数据，那么你得到的所有这些恐怕都将是近似的正态分布．不过你同时应当记住，并不是所有分布都遵循正态分布的模型，除了正态分布之外还有其它分布．

关于正态分布和正态曲线，有许多重要的理论和性质，但是由于比较复杂，这里不作详细讨论．但是，通过考察本节开头提出的数码圆片的例子，我们能够学到关于正态分布的一些基本概念．在这个例子中，我们看到一个数码离开平均值的差距愈大，即离差的绝对值愈大，那么抽得这个数码的概率愈小．例如，抽得 7 的概率只有 1/64，换句话说，抽得 7 仅有约 2% 的机会，而抽得 6 或 7 的概率是 7/64，即大约有 11% 的机会．数学家们已经发现，在正态分布或近似于正态分布的分布中，得到分布中某数据以上部分数据的概率与这数据的离差（以标准差为单位）之间存在着一定的关系．表 19 反映了正态分布的这一性质．

表中第一行说，如果一数据的离差是两个标准差，那么，这数据或比它大的数据出现的概率是 0.02．第二行说，如果一数据的离差是一个半标准差，那么，这数据或比它大的数据出现的概率是 0.07，等等．在数码圆片的例子中所得的结果与表 19 所列数据是相似的．

如果你知道某一数据分布近似于正态分布，并知道这组数据的平均值与标准差，你就可以利用表 19 判断得到某些数据的概率．回到前节中关于体重的那个例子，假定某一年龄儿童的体重数据构成正态分布，那么，超过平均体重两个标准

表 19　正态分布表

一数据的离差(以标准差为单位)	大于或等于这数据的部分数据出现的概率
2(平均值之上 2 个标准差)	2%
$1\frac{1}{2}$ (平均值之上 $1\frac{1}{2}$ 个标准差)	7%
1(平均值之上 1 个标准差)	16%
$\frac{1}{2}$ (平均值之上 $\frac{1}{2}$ 个标准差)	31%
0(距平均值 0 个标准差)	50%
$-\frac{1}{2}$ (平均值之下 $\frac{1}{2}$ 个标准差)	69%
−1(平均值之下 1 个标准差)	84%
$-1\frac{1}{2}$ (平均值之下 $1\frac{1}{2}$ 个标准差)	93%
−2(平均值之下 2 个标准差)	98%

差以上的体重出现的概率仅有 2%. 换句话说，如果你的体重比起年龄和你相同的人的平均体重来正好超出两个标准差，那么在年龄和你相同的人中，仅有 2% 的人与你等重或比你更重. 这句话也可以用另一种方式表达为：在年龄和你相同的人中，你的体重超前百分之九十八，或你的体重的超前百分数是 98. 正态分布表也可用超前百分数予以说明，这样列出的正态分布表如表 20 所示.

如果在一次全国统一的数学考试中，所有考生的得分形成正态分布，平均得分是 42，标准差是 6，那么，这次考试中 48 分的超前百分数是 84. 这就是说，大概有 84% 的考生名列于得 48 分的考生之后，大概有 16% 的考生名列于这考生之前.

表20 正态分布表

一数据的离差(以标准差为单位)	这数据在正态分布中的超前百分数
+2.0(平均值之上2个标准差)	98
+1.5(平均值之上1.5个标准差)	93
+1.0(平均值之上1个标准差)	84
+0.5(平均值之上0.5个标准差)	69
0(与平均值之距离为0)	50
−0.5(平均值之下0.5个标准差)	31
−1.0(平均值之下1个标准差)	16
−1.5(平均值之下1.5个标准差)	7
−2.0(平均值之下2个标准差)	2

请注意表19与表20的相似之处。表19表明，48分或高于48分的成绩出现的概率大约是16%或0.16，因此，低于48分的成绩出现的概率是84%或0.84，这与从表20查得的结果一样。

练习14 正态分布与概率

1. 依据练习十三中的题2和题3所提供的数据，求出伦敦学校的这些学生在拼音测验中各人的超前百分数（假定拼音测验的全部得分形成正态分布，利用表20）。

学生　第一次测验中的超前百分数　第二次测验中的超前百分数

马克

彼得

苏曾

凯

查尔斯

简

2．利用表 19 求出在某个班级中体重在某一重量以上的学生大概有多少．假定全部学生的体重数据形成正态分布，且平均值是 100，标准差是 5．抄录下表并填满空白．

重量	离差	以标准差为单位表示离差	体重等于或大于这重量的学生出现的概率
100	$100-100=0$	0	50%
105			
110			
95			
90			
$102\frac{1}{2}$			
$92\frac{1}{2}$			

3．在测量中为了提高精确度，经常需要重复测量．当比较测量所得的两组量数之间的差数时，我们可以认为这些差数形成正态分布．假设两组量数的差数平均值是 8，标准差是 2，我们可用正态分布表去估计下表中差数的概率．

抄录下表并填充空白．

两组量数的差数	离差	以标准差为单位表示离差	这差数或更大差数出现的概率
10	$10-8=2$	$\frac{2}{2}=1$	16%
6			
8			
9			
5			
12			

4．观察下列数的模型：

1

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

注意每行的两端都是1，其它各数都等于肩上两数的和.

a. 继续写出这个模型的以下三行；

b. 用你写出的最后一行的各数作为一个频率分布的频数，并画出这个频数分布的直方图；

c. 你认为这一分布有些象正态分布吗？当你继续写出以下各行时，你看到最后一行的各数与正态分布的频数之间有什么关系？

5. 某城市一月份的每天最高温度形成一个近似正态分布. 这分布的平均温度是38°F (华氏38°)，标准差是6°F.

a. 这个城市一月份的每天最高温度在47°F以上的概率是多少？

b. 这个城市一月份的每天最高温度在26°F以下的概率是多少？

c. 这个城市一月份的每天最高温度在32°F与44°F之间的概率是多少？

6. 预测准不准

我们已经讨论了抽取数据样本的必要性和用途，并举例说明如何应用概率思想来解决一些统计问题，抽样和概率被列入统计学家最重要的课题之中. 我们在前面还讨论了应用样本、概率和正态分布进行预测的一些例子. 但是这些例子又提出了另一些必须解决的问题：样本能够真正代表从中抽样的总体的概率是多少？如果我们依据样本作预测，那么预测准确的概率是多少？换句话说，统计学家还必须解决这样

的问题："预测究竟有多准确?"

在应用随机样本对从中抽样的事物或数量的总体作预测时，统计工作者通常采用下列两种方法进行：(1) 先根据观察和经验，对总体的某些统计量(如平均值、极差等)作出估计或推断，然后从总体中抽取样本，以检验他的估计的准确程度；(2) 一开始就抽取样本，并计算出样本的一些统计量，比如算出平均值，然后用样本平均值去估计总体的平均值. 这时准确程度一般要比光凭观察得到的估计值要高.

让我们来看一个关于工厂生产中的质量控制的问题. 假定一个圆珠笔制造商要研究他所生产的圆珠笔的质量，他最关心的一个问题是"产品中有多少次品笔?"

他可能用方法 (1) 来解决这个问题，先根据观察和经验推断在每 90 支笔中有一支次品笔，并以此作为正常生产的平均次品率. 然后他从生产的笔中抽样测试 (他当然不想测试每一支笔)，检验的结果是每八十支笔中有一支次品笔. 应用统计方法可以判断，这结果是否接近到足以维持原来的推断 (每 90 支笔中有一支次品笔).

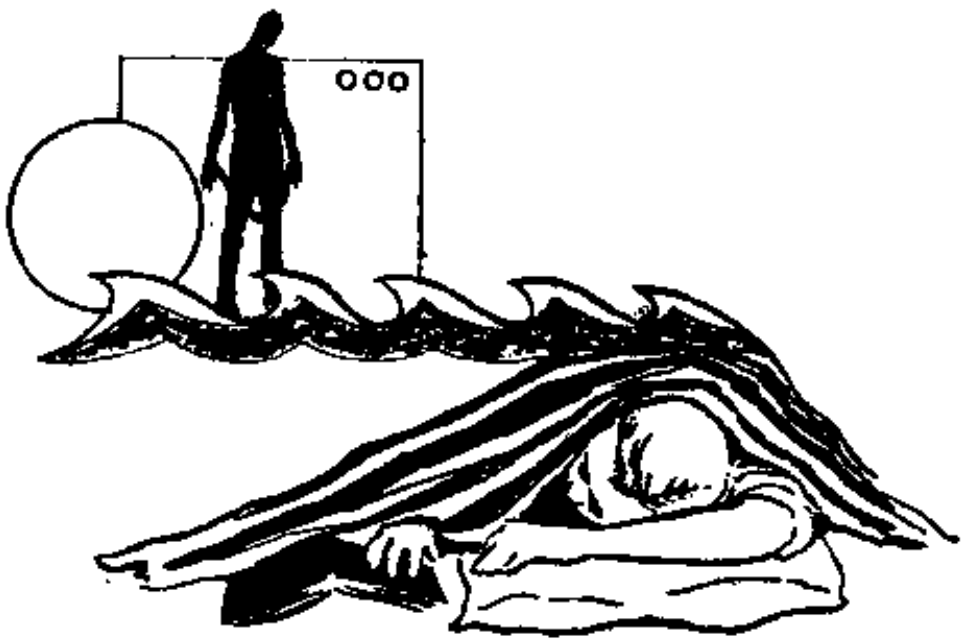
他也可能用方法 (2) 来预测次品率. 为此，他必须先抽样并计算样本中次品笔所占的平均比率. 然后再应用统计方法估计次品率的标准范围.

当然，在这两种情况下这位制造商都会遇到这样的问题：怎样适当地选取样本使它能给出最可靠的结果.

当他得到次品率的统计估计之后，他就可以用统计方法去回答另一些问题，譬如：这样的次品率能否使他的公司依

然保持产品质量高的声誉呢？

对于上面提到的各种统计方法，我们在这一本小册子里不可能详谈．我们只希望你能对于统计中的一些最基本的思想和最重要的方法有一个初步印象，并引起你进一步研究统计学的兴趣。



五、关系何在

——由迷信到科学

你的睡眠时间和你的数学成绩之间是否有关系？犯罪行为的增多是否由电视节目引起的？统计被用来寻找这类问题的答案.

前面我们所讨论的都是关于水平值、离差以及根据数据资料作预测等类问题，现在我们着手讨论关于寻找两组数据之间的关系这个重要问题.

科学上的大多数重要发现都是寻找数据之间关系的结果. 海潮的起落与日、月的位置之间的关系，就是这样一个典

型的例子. 只有当科学家们发现了这种关系之后，他们才弄清楚是什么原因引起了潮汐现象.当他们弄清楚这个原因了，他们就能预报潮汐发生的时间和大小.

但是,有时几种现象偶然一起发生了,却被人们误认为这几种现象之间有什么关系. 例如，一件不幸的事件正巧发生在一只黑猫在场的时候，有的人由此认为不幸事件与黑猫之间有一定的关系，产生了黑猫是不祥之兆的迷信心理. 在古代的印度还有这样的迷信思想，认为月亮的盈亏与天气状况是有关的. 我们一定要注意，不能认为同时发生的事件之间必然具有一种因果关系. 这就是说，当两件事情正好同时发生时,一件事情不一定是另一件事情的原因.

在数学中，我们经常发现必须将两组数据作对比才能找出它们的内在联系. 例如,表 21 是一些圆的直径长度与这些圆的周长之间的对比表.

表 21

直 径	周 长
1 英寸	3.14 英寸
2 英寸	6.28 英寸
3 英寸	9.42 英寸
4 英寸	12.56 英寸
5 英寸	15.70 英寸
6 英寸	18.84 英寸

表 21 说明，周长的数值随着直径的数值的增长而增长.

实际上，只要我们将每个圆的周长除以直径，我们就发现两个数值的比都等于或近似等于 3.14. 正如你所知道的，任何一个圆的周长与直径的比总是给出同一个数值，我们用符号 π 来表示这个常数. 我们用公式：

$$C \div d = \pi \text{ 或 } C = \pi d$$

来表示圆的周长与直径之间的关系. 以表 21 中的每组对应值为坐标，在直角坐标系内描点画图，就得到表 21 所列数据的图象，见图 8. 我们看到，代表对应的直径和周长的点都位于一条直线上. 公式和图象都表明，圆的直径和周长的比是常数. 我们常把这样的一种关系称为完全关系.

圆的直径与周长对比图

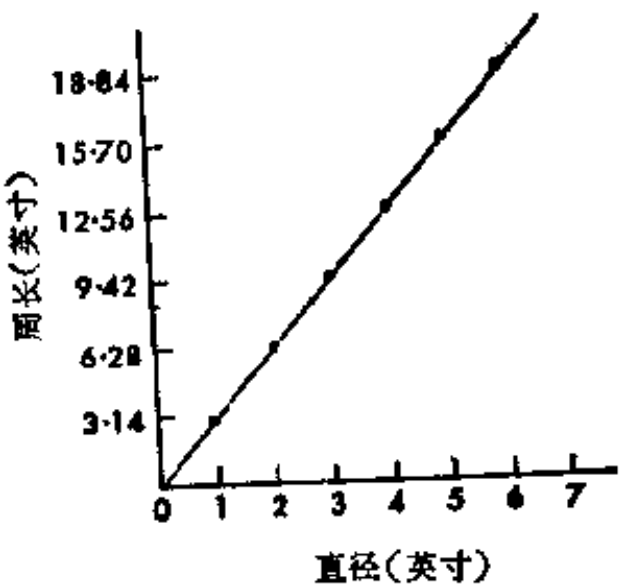


图 8

表 22 是 8 个男孩的体重与他们在一次英语测验中的得分之间的对比表.

以表 22 中的每组对应值为坐标在直角坐标系中找点，如

图 9 所示. 我们看到这些点的位置毫无规律，既不在一条直

表 22

体 重	得 分
90	22
100	20
105	26
110	15
115	8
120	25
125	28
135	10

线上，也不在一条曲线上. 图 9 和表 22 表明，在这两组数据之间没有明显的关系，所以无法用公式来表示.

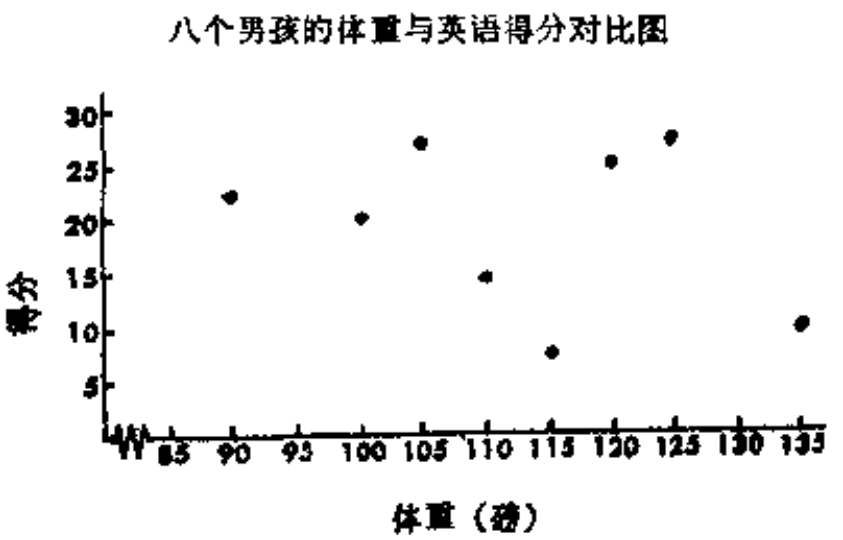


图 9

在举行足球比赛时，球场一般设有小卖部. 表 23 是一个季度内各场比赛卖出的茶的杯数与比赛当天气温的对比表.

表 23

温 度	杯 数
80°F	20
65	24
55	34
50	38
40	50
30	64

依据表 23 的数据画出图象，如图 10 所示.

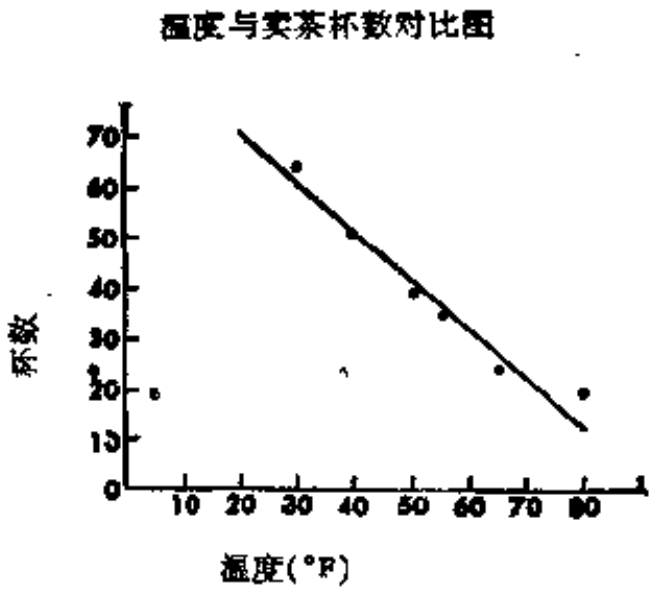


图 10

图 10 和表 23 都表明，在温度和卖茶杯数之间存在着某种关系，我们看到当气温降低时，卖茶杯数增加. 一个量的增加引起另一个对应量的减少，这两个量之间的关系就叫做逆关系(反之，一个量的增加引起另一个对应量的增加，这两个量之间的关系就叫做正关系). 温度和卖茶杯数之间的关系

就是逆关系，但不是完全逆关系，因为所比较的两个量的比值并不总是相等的．这在图 10 上反映为图象上的点并不正好位于一条直线上．这时我们找不到能准确地描述这种关系的公式．

我们在上面考察了将两组数据进行对比的三种不同的情况．对于每种情况，我们都力图确切地描述这种对比结果．虽然有时不可能用数学公式来描述两个量之间的关系，但是在统计学中，却总有可能用数字来描述两组数据之间的关系．不过所用的方法很复杂，这里不便介绍．但是，我们能够对这种方法的最后结果得到一些概念．

两组数据之间的关系叫做**相关**，用以衡量两组数据之间关系的数叫做**相关系数**．相关系数的数值在 -1 与 1 之间．没有关系用相关系数为 0 来表示；完全关系（完全相关）用相关系数 1 来表示；完全逆关系（完全逆相关）用相关系数 -1 来表示．正关系（正相关）的系数是正数；逆关系（逆相关或负相关）的系数是负数．所以，圆的周长与直径之间的相关系数是 1．第 2 个例子中体重与英语得分之间的相关系数是 0．第三个例子中的相关系数是 -0.8 或 -0.9．但是相关系数的计算比较复杂，本书不予研究．

练习 15　两组数据之间的关系

1．判断下列数量关系是哪种相关．用符号＋表示正相关，用符号－表示负相关（逆相关），用 0 表示没有关系．

a．降雨量与足球比赛的观众人数；

b. 一辆汽车的年龄与它的价值；

c. 谷物的价格与牛肉的价格；

d. 学识与生活中的成功；

e. 失业人数与工资数额；

f. 电影观众与电视机购买数量；

g. 一个工厂付的保险费与它的资产；

h. 相貌与才智；

i. 人口总数与在籍学生总数；

j. 圆的面积与它的半径；

k. 在公式 $LW=12$ 中的 L 与 W；

l. 吸烟数量与生肺癌的机会；

m. 行驶路程与消耗的汽油量.

2. 举出三个本书中没有的正相关的例子.

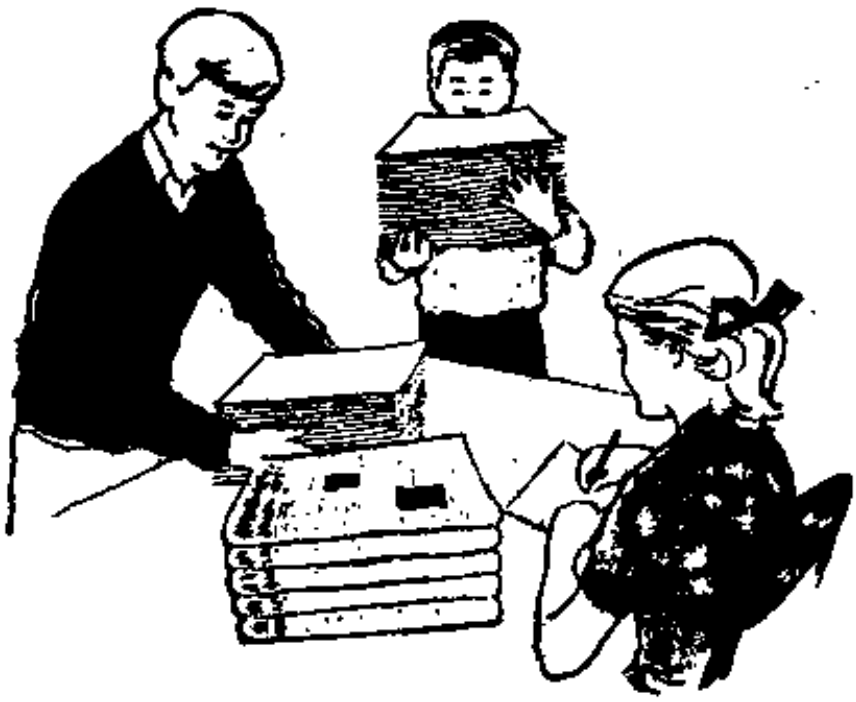
3. 举出三个本书中没有的负相关的例子.

4. 下列量数之间可找到正的相关系数，但是显然并不存在因果关系. 由此可见对相关系数应作具体分析. 试说出下列情形中相关系数较高的可能的原因.

a. 医院的病人数与汽车事故数；

b. 英国电视机的数目与大学教授的平均工资；

c. 烟草的消耗量与死于脊髓灰白质炎人数的减少数.



六、应用所学的知识

——调查数据的实践活动

如果你想要搜集关于各种现实事件的数据，下面的单子将向你提供某些可供调查的项目．你在搜集到这些数据以后，就可应用本书所讨论的一些统计原理对这些数据进行整理和分析．首先对你所搜集的数据确定一个完整的标题，写明搜集方法，然后列出频数分布表，画出数据的图示，并写出依据这些数据能够回答的问题．

1. 体育比赛的成绩记录．

2. 天气报告——温度，雨量，湿度，暴风雨．

3. 交通运输记录——事故，运输量，车辆数，停车场的数

目.

4. 学校中缺席人数和迟到人数.

5. 学费.

6. 学校点心店销售数.

7. 人口统计——出生,死亡,结婚,失业.

8. 文娱活动——听收音机,看电影,借阅图书杂志.

9. 市场行情——股票,谷物,牲畜.

10. 报纸——广告,图片,喜剧.

11. 商业状况——销售额,价格,银行存款,利率.

12. 税率和消费额.

13. 学校考试成绩.

14. 货币流通额,公债.

15. 脉搏,体温,每分钟呼吸次数.

16. 煤气费,水电费.

17. 一本书中每页的字母,单词,数字.

18. 衣服清单——颜色,式样,数量.

19. 身长,体重,鞋子的尺码.

20. 生日的分布,某个生日落入一定期间内的概率.

21. 硬币上的日期.

这里再列举一些活动项目，你在这些活动中可用到本书所讲的许多知识.

1. 进行一些实验以便试验概率法则. 可用来做实验的材料有: 硬币,骰子,卡片,死亡率表,自动售货机等.

2. 把日常生活中的统计实例罗列出来,也可画成图表.

3. 制作一个模型、图表或装置，来说明正态分布。

4. 访问一个保险公司统计室或统计学实验室，也可访问一个统计学家。

七、回顾和展望

学过一些统计的初步知识后，当你看到包含有数据的新闻报道、广告或者什么新发现时，你都要思考这样几个问题以检验其结论是否可靠：

文章的作者是什么人？这人有无偏见？他是否适合回答这个问题？

作者是怎么知道的？他是否应用了足以代表各种观点的随机样本？他怎样选取样本？样本有多大？取了几个样本？关系是否合理？

文章中缺少些什么？用以作比较的标准是什么？结论得以成立的概率是多少？

读过这本书并做过一些练习之后，你就能理解统计世界中的许多事物，你就能识读数据并知道这些表说明什么问题，你就不会被企图制造假象的统计图所迷惑，你就能看穿广告宣传的真实意义，你就懂得并会计算平均值、中位数、超前百分数和标准差等统计量，你就学到关于应用样本和概率去发现未知事物的一些知识．因为你知道了怎样应用统计去解决问题，你就会更加热心于学习统计．因为你懂得一些关于统计世界的知识，世界将会变得更加有趣．

在今天的世界上，统计方法的新应用正在不断地被发现．

应用电子计算机进行计算，使科学家能够解决过去认为不可能解决的统计问题．统计方法的应用将继续扩大，统计学将在未来的空间时代发挥重要的作用．如果你有志于从事统计工作，在你面前展现着广阔的前景．

练 习 答 案

练习 1

答案依赖于你所看到的新闻、广告和问题.

练习 2

1. 长途公共汽车行驶的里程数与消耗的汽油数.

2. 向学生调查,有多少人打算在校用膳.

3. 了解学生最近迟到的原因,包括起身时间,离校路程等,并列表分析.

1. 选择一些初一学生作为样本,要他们记录看电视的时间,至少要统计一个星期.

练习 3

1. 错在重复相减. 例如星期天,假日和睡眠时间有重复部分.

2. 1960 年飞行的飞机要比 1928 年多得多.

3. 可能吉尔西种乳牛得到特别的食物和特别的照顾.

4. 可能在法国行驶的汽车要比德国少.

5. 大多数人即使不服药,伤风以后七天也会自愈.

6 和 7. 应用课文中讲述的方法去分析新闻或广告.

练习 4

1. 36, 34, 33, 32, 31, 28, 28, 25, 23, 21, 20, 19.

中间一个分数是 28;

从上向下数第三个分数是 33;

小于32的分数有8个，占67%.

2.

分数	频数
97	1
96	1
95	2
94	0
93	3
92	3
91	3
90	5
89	1
88	2
87	2
86	1

3. a. 63—64英寸；b. 74.49；c. 63—64；d. 19.

4. b. 69.5—74.5，5，72；c. 70—79，69.5—79.5，74.5；d. 74—76，73.5—76.5.

练习 5

1. a. 1958；b. 苏伊士危机；c. 600,000吨，8%；d. 60%。

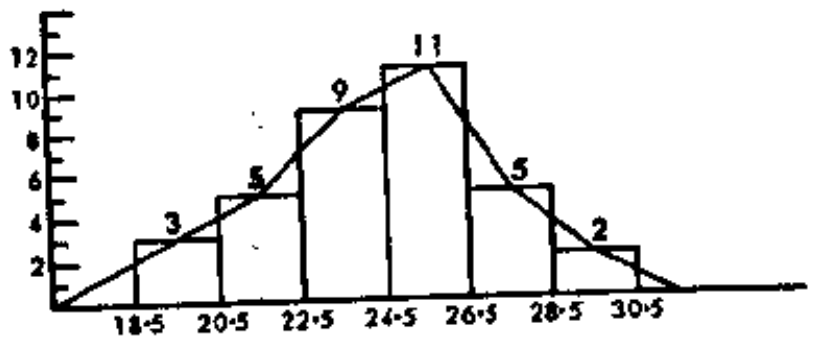
2. a. 交通及杂货，衣料；b. 1945—48，生产成本增加；
   c. 90美分；d. 1美元；e. 1.11美元；
   f. 约13,300美元.

3.

分数区间	组限	频数
29—30	28.5—30.5	2
27—28	26.5—28.5	5
25—26	24.5—26.5	11
23—24	22.5—24.5	9
21—22	20.5—22.5	5

19—20　　18.5—20.5　　3

直方图和频数多边形



练习 6

1．否．因为我们不知道这两个圆形图的面积是否和总的预算成比例。

2．否．因为没有说明这两班的试题难易程度是否相同。

3．商标B。

练习 7

1．a．3；b．21；　c．3。

2．a．4；b．22.5；c．3。

3．a．5.2；b．23；c．3。

练习 8

1．a．86；　b．21；　c．37。

2．a．85；　b．21；　c．38。

3．a．8；　b．—0.5；c．11。

4．a．84.7；b．21.2；　c．38.4。

练习 9

1．中位数是2先令，平均值是2先令6便士．选平均值作为平均

价格。因为已知数据形成差异较大的两组，中位数的数值偏低。

2. a. 众数是 5 镑，中位数是 10 镑，平均值是 91 镑；

b. 5 镑；

c. 中位数，因为它不受少数极端数值的影响。但是，如果要知道通常的捐款是多少，则用众数。如果要估计一个募捐活动能得到的总数，平均值就是最好的代表值。

3. a. 用中位数 (80.5) 或平均值 (81) 都可，但中位数较易计算；

b. 平均值 (79) 较好。已知数据形成差异较大的两组，中位数 (69.5) 的数值偏低。

4. 中位数 (30,000) 最合适。因为它不象平均值那样受少数过大的数值的影响，又这个分布符合众数定义的数据有两个，因此选众数不合适。

练习 10

1. a. 第 4; b. 80%; c. 80; d. 9; e. 11.

2. 79.

3. a. 6; b. 18.

4. a. 第 4; b.92; c. 14%; d. 8; e.48%.

5. 巴巴拉名次较高，第 25 名。约翰的超前百分数是 25.

6. 99, 0.

7. 还必须知道各人所在班级的人数以及排名次的标准。

练习 11

1 和 2. 视具体情况而定。

3. 所列各人群都是按照一定标准选取的，而不是随机抽取的。

4. a. 音乐队员不是随机样本；

b. 商店老板不是随机样本；

c. 仅有具备一定条件的人才会函复；

d. 中午在旅馆的人不是一般居民的随机样本。

练习 12

1. 2/5. 2. a. 1/4; b. 1/2.

3. 用H表示正面，T表示背面，有如下八种可能的结果：HHH, HHT, HTH, HTT, THH, THT, TTH, TTT. a. 1/8, b. 3/8.

4. a. 6; b. 1; c. 1/6; d. 1/6; e. 5/6.

5. a. 7; b. 3/7; c. 4/7. 6. 1/52. 7. 1/4.

8. 1/13. 9. 3/13. 10. 1/9. 11. 5/36.

练习 13

1. a. 2; b. 2.58.

2.

苏曾	16	2
凯	−4	−0.5
查尔斯	0	0
简	12	1.5

3.

马克	10	1
彼得	0	0
苏曾	15	1.5
凯	−10	−1
查尔斯	5	0.5
简	−5	−0.5

a. 彼得; b. 简; c. 马克.

4. $\sigma = \sqrt{\frac{\Sigma f(X-M)^2}{N}}$.

5.

得分(X)	频数(f)	fX	$X-M$	$(X-M)^2$	$f(X-M)^2$
0	1	0	−6	36	36
1	0	0	−5	25	0
2	1	2	−4	16	16
3	1	3	−3	9	9

4	3	12	−2	4	12
5	3	15	−1	1	3
6	5	30	0	0	0
7	8	56	1	1	8
8	2	16	2	4	8
9	2	18	3	9	18
10	1	10	4	16	16

平均值：$M=6$；标准差：$\sigma=2.16$。

练习 14

1.

学生	第一次测验中的超前百分数	第二次测验中的超前百分数
马克	84	84
彼得	16	50
苏曾	98	93
凯	31	16
查尔斯	50	69
简	93	31

2.

重量	离差	以标准差为单位表示离差	体重等于或大于这重量的学生出现的概率
100	0	0	50%
105	5	1	16%
110	10	2	2%
95	−5	−1	84%
90	−10	−2	98%
$102\frac{1}{2}$	$2\frac{1}{2}$	0.5	31%
$92\frac{1}{2}$	$-7\frac{1}{2}$	−1.5	93%

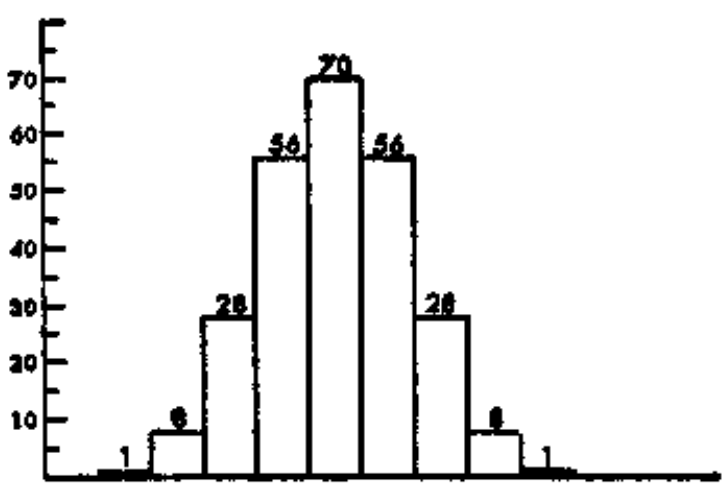
3.

两组量数的	离差	以标准差为单	这差数或更大差

差数		位表示离差	数出现的概率
10	2	1	16%
6	−2	−1	84%
8	0	0	50%
9	1	0.5	31%
5	−3	−1.5	93%
12	4	2	2%

4. a. 1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1

b.



c. 行数愈多，这分布就愈象正态分布。

5. a. 7/100; b. 2/100; c. 68/100。

练习 15

1. a. −; b. −; c. +; d. +; e. −; f. −; g. +; h. 0; i. +; j. +; k. −; l. +; m. +。

2 和 3. 自编例子。4. a. 两者都随人口增加而增加;

b. 两者都随近年来的经济增长而增长;

c. 烟草增加是由于人口增加，另一个降低是由于使用了脊髓灰白质炎疫苗。

[General Information]
书名=统计世界
作者=（英）D．A．约翰逊　　W．H．格伦
页数=88
SS号=10282705
出版日期=1981年05月第1版