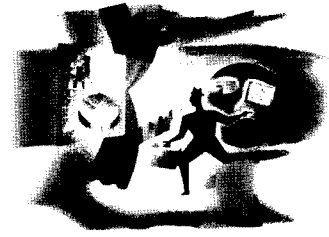
老百姓的数学

统计——数字的科学

盛立人 编著



安徽教育出版社

图书在版编目(CIP)数据

老百姓的数学.统计:数字的科学/盛立人编著.
合肥:安徽教育出版社,2001.9
ISBN 7-5336-2256-1

Ⅰ.老… Ⅱ.盛… Ⅲ.①数学—普及读物②数理统计—普及读物 Ⅳ.01-49

中国版本图书馆 CIP 数据核字(2001)第 062708 号

责任编辑:严云锦 王冰平 装帧设计:李 静
出 版:安徽教育出版社(合肥市跃进路 1 号)
网 址:http://www.ahep.com.cn
经 销:新华书店
排 版:安徽飞腾彩色制版有限责任公司
印 刷:合肥义兴印刷厂
开 本:880×1230 1/32
印 张:3
字 数:70 000
版 次:2001 年 9 月第 1 版 2001 年 9 月第 1 次印刷
印 数:2 000
定 价:5.00 元

发现印装质量问题,影响阅读,请与我社发行部联系调换
电 话:(0551)2651321 邮 编:230061

序

数(shù)起源于数(shǔ),量(liàng)起源于量(liáng)。在有文字历史之前人类就有了数和图形的概念。几千年来,数学由人类生产和社会实践的需求而产生和发展。她不仅被用于科学和技术各领域,也渗透到经济和管理领域以及老百姓的日常生活之中。她不仅是一种工具和语言,也是人们重要的思考方式。她是一种文化,是人类文明的一个重要组成部分。

人类生产和社会实践的需求是数学产生和发展的根本动力。与此同时,数学还有自身内部逻辑完善和追求数学美的强大内部动力。这种内部动力的意义和作用往往不被人们所正确理解,被视为"抽象的游戏"。此外,数学的术语和符号也不易看懂。从中学几何证明开始,数学论述的书写形式就被训练成以固有的逻辑推理为基础,而这种形式常常是探究和思考的真正数学思维方式的颠倒。这一切使人们对数学望而生畏,把数学看成是少数人的一种专门技艺。

综上所述,我们迫切需要用生动的语

言，把数学的进步和数学的思想方法通俗地介绍给大众，让更多的人认识到数学的作用和意义，在不同的工作领域中自觉地采用数学思维方式，使数学更贴近大众。在这方面，盛立人教授撰写的《老百姓的数学》丛书是一个很好的尝试。作者用轻松活泼的语言把人们带进千姿百态的数学世界，让人们领略数学在老百姓日常生活中所起的作用和影响。我相信这套书对于普及和传播数学知识和思想，让民众更加了解、掌握和运用数学，会起到促进的作用。

从某种意义上说，撰写通俗性数学普及读物比撰写专业数学著作和论文更为困难。它需要作者对数学和相关领域的深刻理解，也需要文采。我们需要更多的有识之士共同努力，做好数学的普及与传播这项艰巨而又神圣的事业。

冯克勤

识于 2000 年 3 月

作者的话

一

这是一套关于数学的通俗书籍。作者想通过本书来向读者证明两件事:第一,数学是无所不在的;第二,数学是童叟皆宜的。因此我们可以说,本书的读者对象就是普通老百姓——每一位完成九年制义务教育的老百姓。

作者长期以来一直在思考这么一个问题:面向21世纪,人们应当怎样在文明素质方面做好准备,以迎接世界范围高科技时代的到来?或者说,我们应当在学校和社会之间做些什么铺垫性的工作,才能真正实现科学的社会化和普及化。

一个理想的、倡导学科学和用科学的社会氛围应当具有这样的魅力:

第一,能使十分严谨但略显呆板的课堂内容变为活生生的社会常识。倘使我们能有办法让每个人在面对社会时,能像牛顿所说那样"在在慎思",真可谓善莫大矣!

第二,能让人们在各自工作领域中,自觉地养成理性思考的习惯,时时改善本职工作,以便人人都能成为高科技时代的参与者,而不是同路人。

但是造就这样一个"理想王国"不啻是在建立一座"乌托邦",因而不是一二个人,或是一朝一夕可以完成的。

二

1996年给作者带来一个极好的机遇。那一年作者给安徽大学的几个文科系(社会学、经济和管理等专业)开设了两门新的公选课:《社会选择》和《实用规划》,作为文科数学课的辅助课程。实践说明,这两门课深受文科学生,乃至不少理科学生的欢迎。追踪调查显示,学生们对于这种比传统数学课更灵活、更实用、更轻松的数学,是非常感兴趣的。

正是出于这样的效果,使作者心中燃起了勇气:倘使这些材料能在大学校园外也发挥它应有的作用,也许正是为我们上面提到的那个理想氛围在添砖加瓦。或许那时就产生了写一部书的想法。

但是此事想的容易,做来极难。作者既要时时克服犹豫、彷徨等心理上的障碍,还要不断索求大量资料。幸而我得到的却都是支持和勉励:在南京大学、中国科学技术大学和安徽大学的师长、朋友和同事们的不断鼓励下,才毅然动起笔来。

我最初确定选材的标准是:有实用价值,有深厚背景,有现代意识;同时,材料应不与现有许多通俗书籍重复;方法应强调实用且易于掌握。经多方考虑,将本套书分成五个分册。

《千姿百态的几何世界》

《数学家走近社会学》

《数学家走进管理学》

《计算机的数学故事》

《统计——数字的科学》

在撰写过程中,取材最多的是圣 Olaf 学院,Lynn A. Steen 教授领衔主编的(参与写作的专家达15位之多)

For all practical purposes: Introduction to contemporary mathematics。

除此之外，采集和借鉴的来自报章书刊的材料，为数极多。对于有关作者和译者，在此一并表示我的谢意，不再一一列出。

三

本书从开始撰写到刊行，先后经历5年时间，其间最深切的体会是：比起撰写一部学术专著来，撰写通俗材料其实更为困难。由于受作者能力的限制，选材虽然很用心思，实际仍不能免俗；行文力求流畅生动，写来却时有晦涩；内容纵有考虑，难度仍难以驾驭。这就造成本套书目前这个样子。不足之处，还希望宽容的读者多多包涵才是。

清华大学冯克勤教授对本书表示了自始至终的关注，并愿为之作序，在此表示我衷心的感谢。

最后，在严肃著作的出版和商品经济尚未协调运行的今天，安徽教育出版社毅然支持本书出版，令人感动，特陈数语，以表谢意。

盛立人

安徽大学数学系

2000年10月

目 录

第一章 数据采掘

§1 引 言

读者们一定非常习惯于新闻媒体里许多有关数字的消息。例如我们常常可以看到播音员在新闻联播中宣布说,某省今年的工人下岗率已经下降到7%。在西方国家,更有一些社会咨询部门常常公布就大众关心的某一问题的调查结果。例如说,根据调查,近45%的美国人由于害怕暴力经常晚上呆在家里,不敢出去。有段时间美国媒体炒作得最厉害的事莫过于关于总统克林顿的绯闻的报道。由于总统的威信受到威胁,所以媒体几乎天天报道选民对于总统的支持率。例如某日报导说克林顿的支持率只有46%,过了几天又报导说支持率又从46%回升到54%,等等。但或许很少有人去想,这些数字到底从哪儿来的呢?很明显,在大多数情况,没有人会逐个采访每个公民,以便确定他们是否继续支持总统或是否失业;同样,一般的社会调查只可能是询问了居民当中的一小部分人晚上是否呆在家里。

让我们从另一方面来看一下这个问题。仍以美国为例,美国国家事务局社会劳工部大约每个月要访问60 000个家庭,但实际上美国全国约有9 000万个家庭。可见,上述这些百分比或别的数字来自不到千分之一的小团体。如果这样的数字能够代表整个国家的意见,这就是一件很值得注意的事。因此,我们自然会问:怎么样才能从一个小团体采得的信息得出有关整体信息的结论呢?

如同新闻广播一样，在日常生活中也会出现有关数字的信息。例如，当某医学单位宣布将向市场投放一种治疗爱滋病的药物时，关于它的安全和可靠性便是建立在对于少数患者的临床试验上。而公众对新药的可靠性和安全性，正是建立在这种来自临床数字的统计结论上，专家们常称之为**统计推断**(inferences)。

显然，任何一种数字的运用有赖于正确的数字采集。当然，计算下岗率或民意意向的最好办法是向每一个个体进行直接询问，但这样做起来的工作量非常之大，显然是很不切实际的。代之而用别的办法，应当是仅仅从小群体的信息以获取整个信息。这种想法在统计上常称之为**抽样**。

人们常常用抽样方法来得出整体结论。我们最常见的试验是，从一匙汤来品尝一盆汤的味道。这是因为一盆汤显然十分均匀，因而一匙便可以代表一盆。可是另一些考察却不行，例如对于治安和对于药品的意见就不能如此容易获得，因为这些意见必然因人而异和因地而异。

与各式各样的变化情况打交道，可以说是统计的中心任务。这里第一步是采集代表一个大群体的个体数字，为此必须小心地弄清楚我们到底要从哪一个群体里获取信息。这样的群体在统计上便称为**样本**。

例如，如果我们要统计下岗情况，就必须确定我们将要讨论的样本。应当把什么样的年龄组计算在内？非城镇户口和锒铛入狱的阶下囚算不算在内？为了了解每个月的下岗情况，调查单位便必须回答这些问题。现行的统计办法是把劳动局登记在册的所有工人作为统计对象。

新药物的临床试验则需用一些特殊的样本。样本可能由一些已受感染的个体组成，也可能是一些患特殊病情的病人，或一定年龄的病人，等等。制订临床试验的样本计划(称为协定)常常是药物试验的第一件工作。

§2 抽 样

样本未必都由人组成,也可以是动物或任何物体。例如,我们要测验的样本是新生产的电热丝,目的是测试电热丝在连续通电的情况下是否会烧坏。这时样本中的每一个成员(即每一根电热丝)的情况好坏可能使整个样本失去意义。在另一些情况下,例如当我们要清点一个大仓库中的全部库存物资,一一清点自然会因为倦怠发生错误。在这种情况下,如同民意测验和临床试验的样本一样,人们将只从少量对象中获取信息。用来做出结论的那部分样本成员,便称为**采样**。

如何来选择一个采样使它能真正地作为样本的代表?选择采样的最省事的办法是首选近在身边的个体。如果我们要想知道全国现在有多少人在工作,我们就得站在街口询问每一个过路人是否已被雇用。这种最简单的从样本中就近选取采样的办法称为简单抽样。但简单抽样常常产生无代表意义的数字。当我们从过路人中获取雇员数量信息时,我们总是专找那些服装整齐和中等收入的对象加以询问,避免去采访那些衣着不整、不够友好或相貌凶恶的客体。此外,一个刚刚被雇用的人一般说来是不会大白天上街信步漫游。换言之,这种街头采访的办法不大可能精确地反映出一个较大范围的下岗率。

单凭友善与否的外表或收入水平来决定抽样个体,不管是有意识或无意识,我们将会失去具有代表意义的样本成员。例如以街头访问而言,可能是选取中产阶层或富有家庭太多,而选取蓝领阶层或需要脱贫的家庭太少。这种由于采样不同所得出的不同的结果,以及由此得到的样本的结论之间的差别,常称为**偏袒**(bias)。为了采集到精确的数字必须采取一些特殊的步骤以消除偏袒。在实际中,统计学家们通过大量的努力来消除采样过程中

出现的这种人为因素。让我们举个例子。

例 电话投票

电视观众应邀用电话登记它们的意见。电视台设有公用电话,通常是节目开始时提出问题,在节目结束时回答结果。观众们拨打几个特殊号码中的一个电话告知自己的意见。拨打一个号码表示同意,拨打两个号码表示反对。这个办法可以避免对话,只要比较电话登记就可以知道答案。

其实这种电话表决仍然会出现偏袒。显然,没有安装电话的家庭便被自动摒弃在外(在美国,93%的人家安装了电话,但分布并不均匀:阿拉斯加州有四分之一家庭未装电话,密西西比州也有十五分之一家庭未装电话)。此外,虽然拨打特殊的带有前缀的电话所花无几,但低收入家庭的人员仍然会考虑是不是值得给电视台拨打电话。

电话投票中出现偏袒最重要的原因是志愿应答——即应答者选择自己为询问对象。只有那些既有时间,又无烦恼和花得起电话费的人才被包括在内。任何一种由志愿应答所选择的抽样都带有一定的感情色彩:不是正面的就是负面的。例如,在询问成人是否因为害怕暴力而晚上呆在家里时,对暴力特别憎恨的人更愿意向电视台表示自己的否定意见。志愿应答可以说是电话投票中最常有的重要的偏袒。

电话投票还可能出现虚假操作。如亚洲某些国家的暴力问题常出于政治原因引起,这时一个政党完全可以唆使它的成员整个晚上拨打电话,或者干脆用计算机设置一个程序就行了,因为根本不需要说话。

因此,人为选择是抽样中产生偏袒的常见原因。不管是用当面采访的方式还是志愿应答的方式,人为选择都可能使调查结果失去意义。因此要消除偏袒,重要的是克服人为选择的影响。统计学家消除人为偏袒的办法是用机会均等来进行采样。

§3 随机抽样

设想我们有一个装有千把颗珠子的盒子,这些珠子完全匀等,只是颜色有别:大部分为淡色,少量为深色。这些珠子便构成一个样本。我们也可以设想,这些珠子全体表示的是全体职工,而深色珠表示下岗人员。

我们的意图是想,不用每次数数而能估计出深色珠子在样本中所占的百分比。假设已将珠子在箱子里搅混,然后每次倒出 50 颗。这时要注意的是,每一颗珠子与别的珠子有着同样的选择机会,而每 50 颗的一组也与别的 50 颗的一组有着同样的选择机会。这便是一个简单的幅度为 50 的随机抽样。

一个幅度为 n 的**简单随机抽样**,是一个这样的抽样方法,其时样本中每一个成员个数为 n 的抽样具有同样的被选机会。

既然每一颗珠子具有同样的被选机会,所以简单随机抽样可以消除抽样带来的偏袒。当我们将这种方法用到社会问题调查时,简单抽样方法便赋予每个人——富人或穷人,男人或女人,就业者与失业者——以同样的机会。

如果我们从珠子盒里第一次抽样得到 12 颗深色珠,即这个抽样有$\frac{12}{50}$,即 24%的深色珠子,则由于简单随机抽样能避免偏袒,我们就用这个抽样的百分比来估计整个样本含深色珠的百分比。换句话说,我们对于整个箱子里深色珠的百分比,也是估计为 24%。这里已经包含一个重要的统计学技巧:要估计一个样本的特征,采取简单随机抽样方法,并用这个抽样的特征来作出整个样本的估计。

原则上我们已经知道如何来估计一个较大范围的下岗率。例如将所有工人的名字都输入电脑,将这些名字次序排乱后,取出其

中的 60 000 个来，由于这是一个简单随机抽样，所以这 60 000 人中的下岗率，便可以用来估计全部工人的下岗率。当然，如果样本太大，采用记名字的方法不是一个好主意。例如可以将样本的每一个成员用数码来标记，其时一个抽样即可以从数码的集合中选出，选择时要求每一数码具有同等的被选机会。如果样本相对较小，则编号码与选择抽样就可以通过表 1－1 那样的一张随机数表进行。对于大样本的情形，计算机可以提供数据以生成相应的随机数表，并打出所需的抽样。

表 1－1　随机数表

101	03918	86495	47372	21870	28522	99445	38783	83307
102	10041	35095	66357	64569	08993	20429	28569	63809
103	43537	58368	80237	17407	89680	04655	24678	61932
104	64301	47201	31905	60410	80101	33382	95255	10353
105	43857	42186	77011	93839	28380	49296	63311	49713
106	91823	39794	47046	78563	89328	39478	04123	19287
107	34017	87878	35674	39212	98246	29735	09924	27893
108	49105	00755	39242	50472	39581	44036	54518	46865
109	72479	02741	75732	99808	02382	77201	44932	88978
110	84281	45650	28016	77753	39495	41847	19634	82681
111	61589	35486	59500	20060	89769	54870	75586	07853
112	25318	01995	87789	41212	74907	90734	31946	24921
113	40113	37395	51406	93099	43023	70195	70713	72306
114	58420	43526	15539	24845	15582	16780	95286	69021
115	18075	45894	09875	42869	20618	07699	80671	54287
116	52754	73124	93276	71521	59618	44966	37502	11570
117	05255	53579	08239	99174	75548	95776	42314	13093
118	76032	35569	28738	38092	74669	00749	17832	64855
119	97050	31553	32350	51491	53659	89336	36912	05292
120	29030	43074	84602	95131	22769	44680	68492	33987
121	28124	29686	63745	12313	15745	11570	20953	17149

122	97469	41277	90524	36459	22178	63785	20466	67130
123	91754	40784	38916	12949	76104	20556	34001	59133
124	84599	29798	57707	57392	91757	76994	43827	69089
125	06490	42228	94940	10668	62072	58983	10263	08832
126	30666	02218	89355	76117	75167	69005	42479	79865
127	87228	15736	08506	29759	74257	85594	75154	48664
128	45133	49229	32502	99698	68202	44704	39191	73740
129	55713	98670	57794	64795	27102	83420	26630	95009
130	20390	38266	30138	61250	07527	02014	43972	49370
131	13400	68249	32459	41627	56194	93075	50520	96784
132	08900	87788	73717	19287	69954	45917	80026	55598
133	86757	47905	16890	99047	78249	73739	97076	00525
134	19862	54700	18777	22218	25414	13151	57954	80615
135	96282	11576	59837	27428	60015	40338	39435	94021
136	17463	26715	71680	04853	55725	87792	99907	67156
137	44880	55285	95472	57551	24602	98311	63293	58110
138	61911	78152	96341	31473	58398	61602	38143	93833
139	07769	22819	58373	88466	71341	32772	93643	92855
140	73063	63623	29388	89507	78553	62792	89343	27401
141	24187	60720	74055	36902	22047	09091	79368	35408
142	06875	53335	91274	87824	04137	77579	54266	38762
143	23393	37710	46457	03553	58275	11138	18521	59667
144	00980	73632	88008	10060	48563	31874	90785	78923
145	46611	39359	98036	25351	88031	72020	13837	81321
146	56644	79453	49072	30594	73185	81691	29225	70459
147	98350	36891	04873	71321	29929	37145	95906	41005
148	17444	61728	86112	76261	92519	61569	65672	95772
149	45785	21301	89563	23018	60423	50801	70564	45398
150	54369	08513	36838	19805	67827	74938	66946	01206

一张随机数表是一张数字 0,1,2,3,4,5,6,7,8 及 9 的数串，由下面办法构成:每一数字的选择具有相等的 10 种可能性，并且