社会科学研究方法系列

世纪出版集团 上海人民出版社

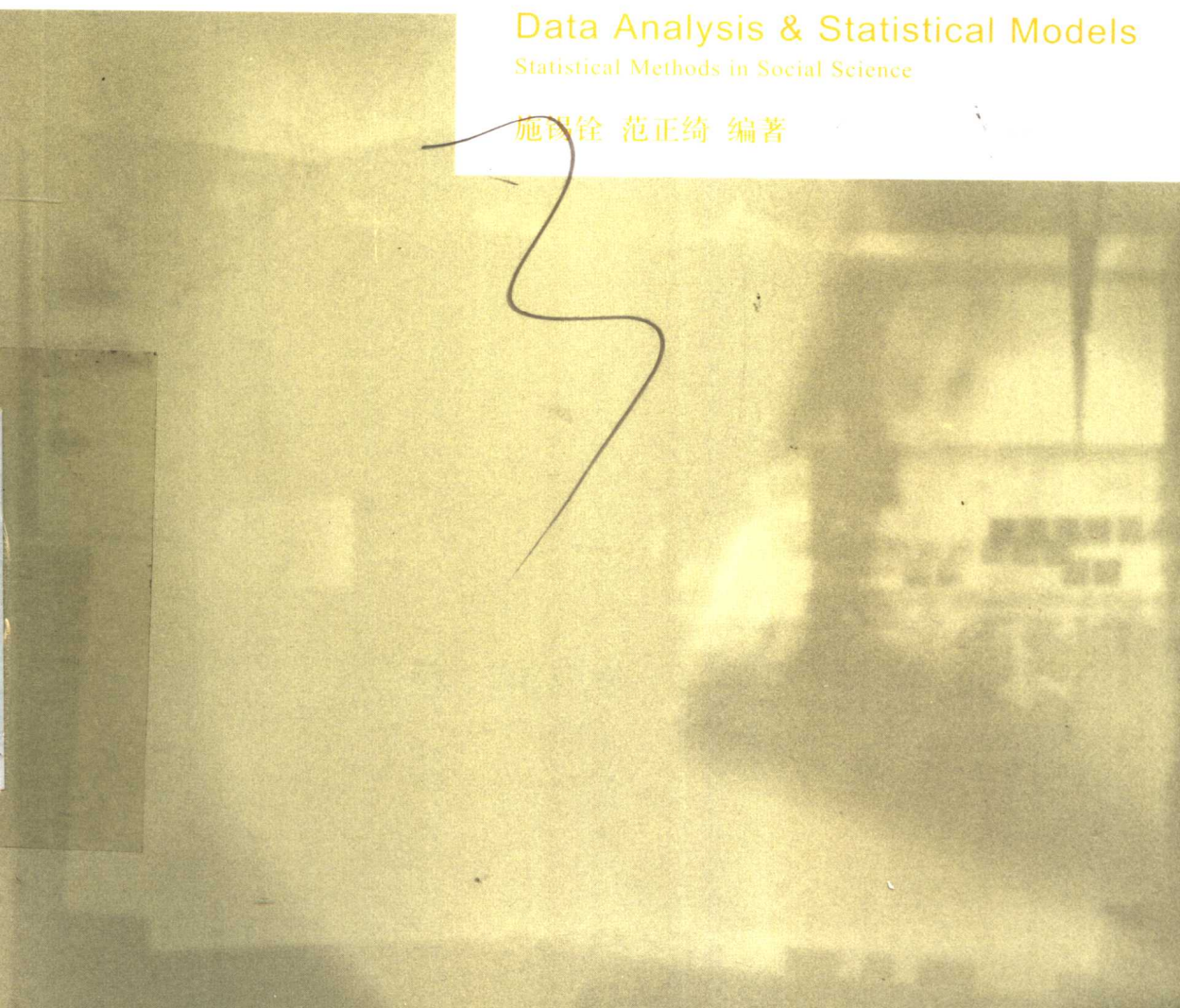
数据分析与统计建模

社科研究中的统计学方法

Data Analysis & Statistical Models

Statistical Methods in Social Science

施国铨 范正绮 编著



社会科学研究方法系列 | 世纪出版集团 上海人民出版社

数据分析与统计建模

社科研究中的统计学方法

Data Analysis & Statistical Models

Statistical Methods in Social Science

施锡铨 范正绮 编著

图书在版编目（CIP）数据

数据分析与统计建模：社科研究中的统计学方法/施锡铨，范正绮编著．—上海：上海人民出版社，2007
（社会科学研究方法系列）
ISBN 978-7-208-07362-3

Ⅰ．数… Ⅱ．①施…②范… Ⅲ．社会统计—统计方法
Ⅳ．C81

中国版本图书馆 CIP 数据核字（2007）第 135369 号

责任编辑　钱　敏
封面设计　人马艺术设计工作室·储平

·社会科学研究方法系列·
数据分析与统计建模
——社科研究中的统计学方法
施锡铨　范正绮　编著

出　版　世纪出版集团　上海人民出版社
（200001　上海福建中路 193 号　www.ewen.cc）

出　品　上海世纪出版股份有限公司高等教育图书公司
www.hibooks.cn
世纪高教（上海福建中路 193 号 24 层　021-63914988）

发　行　世纪出版集团发行中心
印　刷　上海商务联西印刷有限公司
开　本　787×1092 毫米　1/16
印　张　26
插　页　2
字　数　432,000
版　次　2007 年 9 月第 1 版
印　次　2007 年 9 月第 1 次印刷
ISBN 978-7-208-07362-3/F·1668
定　价　39.00 元

前 言

最初，这本书的工作被纳入学校某工程的计划之内，我们受命为财经类的研究生编写一本有关统计建模的工具书或教材。按照常理，人们都认为这只不过是件抄抄编编、拼拼凑凑的工作，很快就可以解决问题的。偏巧最近较长时期内，我们为各种层次的经济管理类学生讲授或深或浅的统计学方法，多少积累了一些统计教育方面的心得体会（当然，讲这种话，多少有点“自以为是”的味道，但是，既然自己觉得有些体会，还是希望得到读者的支持）。于是，很想把这些内容整理一下，做一个归纳总结，奉献给大家，以便互相交流，对统计思想有更深入的理解。正是出于这个设想，编写的过程几经推倒重来，在进程上违反了“计划”的要求。众所周知，当设想与计划产生矛盾的时候，博弈的结果自然“胳膊拗不过大腿”，“断奶”成为必然的结果。这给本书的编著工作多少带来冲击。无可非议，这要怪我们自己的“慢动作”太多，然而从内心还是盼望着人间自有真情与温暖。上海世纪出版集团在此过程中给予我们帮助，使得本书得以面世。倘若读者能觉得这本书真的在处理数据时很有用，总算这几年我们没有白白地苦熬苦度。

统计学方法是一门较难讲授的课，因为它所研究的是不确定事件，讨论如何收集数据、整理数据以及分析数据，在此基础上做出科学的推断与预测。从局部推断整体，从历史与现在试图预测未来，不可避免地会产生误差，为决策带来风险。因此它的讲授和学习与确定性的数学存在着极大的差异。况且，一个人即使学完了所有基本的统计方法，一大堆“杂乱无章”的数据放在他面前，也未必能顺利地建立起模型来。统计学的建模方法并不像放在超市货架上的商品，你想派什么用处，就到货架上去找相应的东西就可以了。所谓统计建模，应当是各种方法的巧妙综合利用，甚至你还得想出新点子。所以同样的数据，不同的分析人员也许可以建立不同的模型，只要估计与预测的误差小，都算找到了一个描述数据的好模型，正符合“不管白猫黑猫，逮着老鼠的就是好猫”的准则。要达到这种境界，在讲授时必须要强调统计的“idea”，这构成了统计学在教学与科研中最显著的特点。而要在书中完美地体现出统计思想，对于我们来说，却是一件极其困难的任务。然而，困难并不单纯地在于此。由于我们面对的对象是财经类的人

员，我国的教学体制将这些人员基本上划为“文科”性质，因此数学分析的功底相对比较浅薄。要对这批读者从理论上讲清楚得到数学强烈支撑的统计学原理，难度真可比喻为“上天揽月”。深思熟虑下来，我们决定编写一本几乎不用数学证明的教科书（当然，适当的公式还是需要罗列），因为统计学与日常生活的联系相当密切，许多统计的基本概念几乎都来自于人们平平常常的生活。例如，平均数是一个人人都知道并且几乎天天使用的指标。即使复杂一点的回归模型，当人们预测某个 2.20 米的篮球运动员其儿子成年之后的身高时，大多数人会比较小心地预测一个小于 2.20 米的高度，而当预测一个仅有 1.30 米左右特别矮小的人其儿子成年之后的身高时，大多数人宁愿拿出一个大于 1.30 米的数字。这种现象称为“回归自然”的效应。许多人尽管不一定知道建立一个父子身高的回归模型，但却会“回归自然”地做出预测……倘若将这些人们意识中的东西融入相应的教学内容，即使不学习深奥的数学手段，也能使人理解统计学方法的某种奥妙。这种做法在国外被大量地引入应用类的统计教材。我们也试图在本书中尽可能地“深入浅出”。细心的读者可以从本书中领略到，对某些概念的叙述，作者确实是有所体会才这样描述的。

我们在编写过程中，还遇到一些问题，发现近年来国外的一些著名教材中的提法与我们自己以前学到的并且在实际中做的有所出入。例如，在线性回归模型的诊断过程中，“常数项是否为零”这一点是否要检验？我们以前从老师那儿学的，并且在以后的教学中依样教学生的做法是：在所有解释变量都完成了 t-检验后，就应对常数项做同样的检验。但是国外的某些教材在例题中居然对此“置之不理”。是他们疏忽了，还是发生了其他的准则引入？为了对读者负责，为了对科学负责，我们不得不翻阅大量国外最新著作，钻进了许多专业性的网站提问探索，利用现代化的通讯工具向朋友、专家请教。最后得到了一个尚能自圆其说的说法，收入在本书的相关部分之中。两个作者都已经年过六旬，“活到老，学到老”，这本书的编著过程也折射了作者不断学习的过程。

既想避开繁复的数学，又想让学生学会统计，计算机统计软件的引入显得格外重要。让学生先依样画葫芦地做例题，然后再回过头来一一详细地解释计算机的输出内容，对于理解统计思想有着“事半功倍”的效果。本书对每一种方法都进行列举，对几乎每一个例题都讲授 Excel 或 Minitab（或甚至两种都介绍）的详细操作，有时详细到连数据表如何排列，操作的每一步如何进行，有时会发生什么样的问题，不厌其烦地罗列在读者面前。我们自己也笑称这本书似乎成了

一架“傻瓜照相机”，照着书本在计算机上一步一步地走，读者会发现无穷的乐趣。有个管理者使用了我们所教的模型，在计算机上预测了公司在第二年的销售情况，我们至今还记得他讲述这件事的那份喜悦，“每年的销售计划不用拍脑袋的形式去编造了”。

用于建模与处理数据的统计软件有许许多多，本书只介绍 Excel 与 Minitab 两种，这大概是数据分析中最简单的软件了。教与学，只能用简单一些的，毕竟我们的目的是先学会各种统计方法及其原理，在计算机上操作软件只不过是重要的辅助手段。对于经济管理中的大量数据的处理，或者要建立比较复杂的统计模型，Excel 与 Minitab 显然是不够的，掌握了统计手段的人自然会去寻找更合适的统计软件，原理是一样的。你也可以知道去找什么样的人才帮助你做这件工作，你的“知道”当然借助于你现在所获得的统计知识。

两个老人，磨磨蹭蹭了好几年，总算拼凑了这么一本书，是不是会被人讥讽为“老黄牛拖破车”呢？我们怀着惴惴不安的心情等待着读者的反馈。

施锡铨、范正绮

2007 年 6 月

目　录

第三篇 属性数据分析

第四篇 非参数统计

第五篇 时间序列分析

第一篇　数据分析的图方法

众所周知，数据分析是运用统计理论并借助计算机对收集到的数据进行整理、归纳、分析，据此进行推断或预测的一门科学，因此，在操作过程中，会出现大量的数据列表和计算机输出结果。为什么还需要引进“图”呢？无非基于如下原因：

首先，图形能给人强烈的直感。无可厚非，人们的视觉系统也许是最高级的信息处理系统。直观的图像可以提供给人们大量的他们感兴趣的信息。例如，使人们或多或少可以对数据的结构有粗略的印象，可以发现数据呈现出来的形态以及提示变量之间大致的相互关系，从而通过视觉所获得的信息抽象出数据或图形的显著特征。

具体地说，通过直观的图像，我们可以探索时间序列是否有平稳的迹象；两个变量之间是否可能存在某种线性关系；我们也可以推断数据是否凝聚在“中心”附近或者向四周离散；两组数据（或两个母体）的分布是否存在显著的差异；通过统计质量控制图，我们可以分析产品的质量是否处于正常的控制范围之内；多元分析的图形可以提示我们有关产品之间的差异主要体现在哪几个质量指标上；等等。

其次，图形是统计工作人员向人们解释数据分析结果最有效的辅助工具。尤其在管理学领域，无论在试图说服上层管理人员了解某种情况，或者向属下的工作人员布置工作时，图形兴许都是不可或缺的“武器”。

并不是所有的管理层领导都能理解数据分析。特别在我国，不熟悉数据分析的管理者为数不少。然而，现代管理又少不了用数据来讲话，每当需要向某些人解释原始数据所反映的信息以及由此产生的推断时，图文并茂的形式有时会收到意想不到的效果。

统计的“图表示”通常分为两类：一是对原始数据的描绘，二是关于衍生数据的图像。对原始数据的描绘常用于探索性数据分析，而对于衍生数据的图表示常常作为数据统计分析的辅助手段。譬如，我们常用残差图来对线性回归模型中的若干假定进行验证。本书的第一部分将对这两类“图表示”分别进行介绍，同时也将原始数据分为一维和多维的情况个别处理。

第 1 章

一维样本数据的分布

1.1 各国人均 GNI 数据

研究世界经济，避免不了 GDP、人均 GNI、汇率等指标以及采集到的相关数据，譬如我们手头拥有近几年来约 184 个国家或地区的人均 GNI 数据：

表 1.1 近几年世界大多数国家或地区的人均 GNI(单位：美元)

编号	国家或地区		1998 年	1999 年	2000 年	2001 年	2002 年
1	Albania	阿尔巴尼亚	890	970	1 220	1 400	1 450
2	Algeria	阿尔及利亚	1 560	1 540	1 580	1 660	1 720
3	Angola	安哥拉	520	430	470	530	710
4	Antigua and Barbuda	安提瓜岛及巴布达岛	8 430	8 810	9 150	9 560	9 720
5	Argentina	阿根廷	8 230	7 780	7 690	7 200	4 220
6	Armenia	亚美尼亚	570	600	650	700	790
7	Australia	澳大利亚	21 240	20 860	20 080	19 850	19 530
8	Austria	奥地利	27 040	26 130	25 730	24 230	23 860
9	Azerbaijan	阿塞拜疆	510	570	610	660	710
10	Bahamas, The	巴哈马群岛	12 940	14 000	14 860	..	..
11	Bahrain	巴林	9 610	9 560	10 160	10 410	10 500
12	Bangladesh	孟加拉国	360	370	380	380	380
13	Barbados	巴巴多斯岛	8 230	8 650	9 130	8 980	8 790
14	Belarus	白俄罗斯	1 560	1 410	1 380	1 300	1 360
15	Belgium	比利时	25 590	25 060	24 900	23 530	22 940
16	Belize	伯利兹	2 700	2 660	2 880	2 940	2 970

（续表）

编号	国家或地区		1998年	1999年	2000年	2001年	2002年
17	Benin	贝宁湾	390	390	390	380	380
18	Bhutan	不丹	450	470	510	560	600
19	Bolivia	玻利维亚	990	980	980	940	900
20	Bosnia and Herzegovina	波斯尼亚和黑塞哥维那	1 190	1 240	1 270	1 280	1 310
21	Botswana	博茨瓦纳	3 290	3 020	3 040	3 170	3 010
22	Brazil	巴西	4 610	3 930	3 650	3 090	2 830
23	Bulgaria	保加利亚	1 270	1 450	1 580	1 650	1 770
24	Burkina Faso	布基纳法索	250	260	250	240	250
25	Burundi	布隆迪	140	120	110	100	100
26	Cambodia	柬埔寨	290	290	290	300	300
27	Cameroon	喀麦隆	600	600	580	570	550
28	Canada	加拿大	20 000	20 600	21 720	21 930	22 390
29	Cape Verde	佛得角	1 300	1 340	1 320	1 280	1 250
30	Central African Republic	中非共和国	290	280	280	270	250
31	Chad	乍得湖	220	210	190	200	210
32	Chile	智利	4 890	4 730	4 810	4 600	4 250
33	China	中国	740	780	840	900	960
34	Colombia	哥伦比亚	2 410	2 190	2 050	1 910	1 820
35	Comoros	科摩罗	410	400	380	380	390
36	Congo, Dem. Rep.	刚果民主共和国	110	90	90	90	100
37	Congo, Rep.	刚果共和国	530	450	510	570	610
38	Costa Rica	哥斯达黎加	3 590	3 580	3 820	3 970	4 070
39	Cote d'Ivoire	科特迪瓦	780	750	690	640	620
40	Croatia	克罗地亚	4 690	4 500	4 430	4 330	4 540
41	Cyprus	塞浦路斯	12 110	12 220	12 460	12 320	..
42	Czech Republic	捷克	5 160	5 120	5 250	5 260	5 480
43	Denmark	丹麦	32 770	32 240	31 510	30 470	30 260

（续表）

编号	国家或地区		1998 年	1999 年	2000 年	2001 年	2002 年
44	Djibouti	吉布提	790	810	830	840	850
45	Dominica	多米尼加	3 270	3 240	3 270	3 350	3 000
46	Dominican Republic	多米尼加共和国	1 870	1 970	2 140	..	..
47	East Asia & Pacific	东亚和太平洋地区	800	810	860	910	960
48	Ecuador	厄瓜多尔	1 800	1 480	1 330	1 370	1 490
49	Egypt, Arab Rep.	埃及	1 270	1 370	1 490	1 530	1 470
50	El Salvador	萨尔瓦多	1 860	1 930	2 020	2 070	2 110
51	Equatorial Guinea	赤道几内亚	1 060	820	700	930	..
52	Eritrea	厄立特里亚	220	210	180	190	190
53	Estonia	爱沙尼亚	3 490	3 540	3 790	3 930	4 190
54	Ethiopia	埃塞俄比亚	110	110	110	110	100
55	Fiji	斐济	2 300	2 300	2 050	2 100	2 130
56	Finland	芬兰	24 750	24 640	25 000	23 940	23 890
57	France	法国	24 770	24 400	23 990	22 880	22 240
58	French Polynesia	法属玻利尼西亚	16 920	16 540	16 150	..	..
59	Gabon	加蓬	3 870	3 220	3 120	3 100	3 060
60	Gambia, The	冈比亚	330	340	330	320	270
61	Germany	德国	26 630	25 740	25 150	23 540	22 740
62	Ghana	加纳	380	380	330	290	270
63	Greece	希腊	12 130	11 910	11 700	11 450	11 660
64	Grenada	格林纳达	3 150	3 460	3 720	3 630	3 530
65	Guatemala	危地马拉	1 660	1 690	1 700	1 700	1 760
66	Guinea	几内亚	520	500	450	420	410
67	Guinea-Bissau	几内亚比绍共和国	140	150	160	140	130
68	Guyana	圭亚那	860	870	860	860	860
69	Haiti	海地	430	480	500	480	440
70	Honduras	洪都拉斯	740	770	870	910	930
71	Hong Kong, China	中国香港	24 820	25 580	26 830	25 780	24 690

（续表）

编号	国家或地区		1998年	1999年	2000年	2001年	2002年
72	Hungary	匈牙利	4 480	4 620	4 770	4 820	5 290
73	Iceland	冰岛	27 390	28 390	29 920	28 660	27 960
74	India	印度	420	440	450	460	470
75	Indonesia	印度尼西亚	670	590	570	680	710
76	Iran, Islamic Rep.	伊朗	1 650	1 600	1 650	1 690	1 720
77	Ireland	爱尔兰	20 630	21 750	22 970	22 950	23 030
78	Israel	以色列	16 730	16 470	17 020	16 760	16 020
79	Italy	意大利	20 560	20 350	20 170	19 490	19 080
80	Jamaica	牙买加	2 450	2 610	2 710	2 710	2 690
81	Japan	日本	33 780	33 170	35 400	35 920	34 010
82	Jordan	约旦	1 590	1 620	1 720	1 750	1 760
83	Kazakhstan	哈萨克斯坦	1 350	1 260	1 250	1 350	1 520
84	Kenya	肯尼亚	360	360	350	350	360
85	Kiribati	基里巴斯	1 150	1 080	1 060	1 070	960
86	Korea, Rep.	韩国	8 500	8 530	9 010	9 490	9 930
87	Kuwait	科威特	17 390	15 280	16 280	16 700	16 340
88	Kyrgyz Republic	吉尔吉斯斯坦共和国	350	300	280	280	290
89	Lao PDR	老挝	310	290	290	300	310
90	Latvia	拉脱维亚	2 430	2 570	2 940	3 260	3 480
91	Lebanon	黎巴嫩	3 670	3 910	4 000	4 000	3 990
92	Lesotho	莱索托	690	670	640	600	550
93	Liberia	利比里亚	110	110	130	140	140
94	Lithuania	立陶宛	2 700	2 850	3 110	3 340	3 670
95	Luxembourg	卢森堡公国	44 810	44 970	43 720	41 550	39 470
96	Macao, China	中国澳门	15 220	14 420	14 800	14 600	..
97	Macedonia, FYR	马其顿王国	1 920	1 830	1 830	1 690	1 710
98	Madagascar	马达加斯加岛	260	250	250	260	230

（续表）

编号	国家或地区		1998 年	1999 年	2000 年	2001 年	2002 年
99	Malawi	马拉维	220	190	170	160	160
100	Malaysia	马来群岛	3 630	3 370	3 390	3 410	3 540
101	Maldives	马尔代夫	1 950	2 050	2 130	2 130	2 170
102	Mali	马里	250	250	230	230	240
103	Malta	马耳他	8 790	9 270	9 300	9 280	9 260
104	Marshall Islands	马绍尔群岛	2 110	2 150	2 200	2 220	2 380
105	Mauritania	毛里塔尼亚	420	390	350	360	280
106	Mauritius	毛里求斯	3 760	3 720	3 690	3 850	3 860
107	Mexico	墨西哥	4 020	4 450	5 100	5 550	5 920
108	Micronesia, Fed. Sts.	密克罗尼西亚	1 910	1 940	2 030	1 970	1 970
109	Moldova	摩尔多瓦	470	410	390	400	460
110	Mongolia	蒙古	410	390	390	400	430
111	Morocco	摩洛哥	1 260	1 210	1 180	1 190	1 170
112	Mozambique	莫桑比克	200	220	210	200	200
113	Namibia	纳米比亚	2 160	2 140	2 100	1 970	1 790
114	Nepal	尼泊尔	220	220	230	240	230
115	Netherlands	荷兰	25 160	25 270	25 330	24 010	23 390
116	New Caledonia	新喀里多尼亚(岛)	15 750	14 820	14 030	..	..
117	New Zealand	新西兰	15 310	14 640	13 480	13 280	13 260
118	Nicaragua	尼加拉瓜	380	400	520	600	710
119	Niger	尼日尔	200	190	180	180	180
120	Nigeria	尼日利亚	260	260	270	300	300
121	Norway	挪威	35 240	34 530	35 660	36 960	38 730
122	Oman	阿曼	6 420	6 120	6 710	7 720	7 830
123	Pakistan	巴基斯坦	470	460	450	420	420
124	Palau	帕劳群岛	5 960	6 070	7 220	6 610	6 820
125	Panama	巴拿马	3 590	3 680	3 920	3 920	4 020