

漫画趣知系列

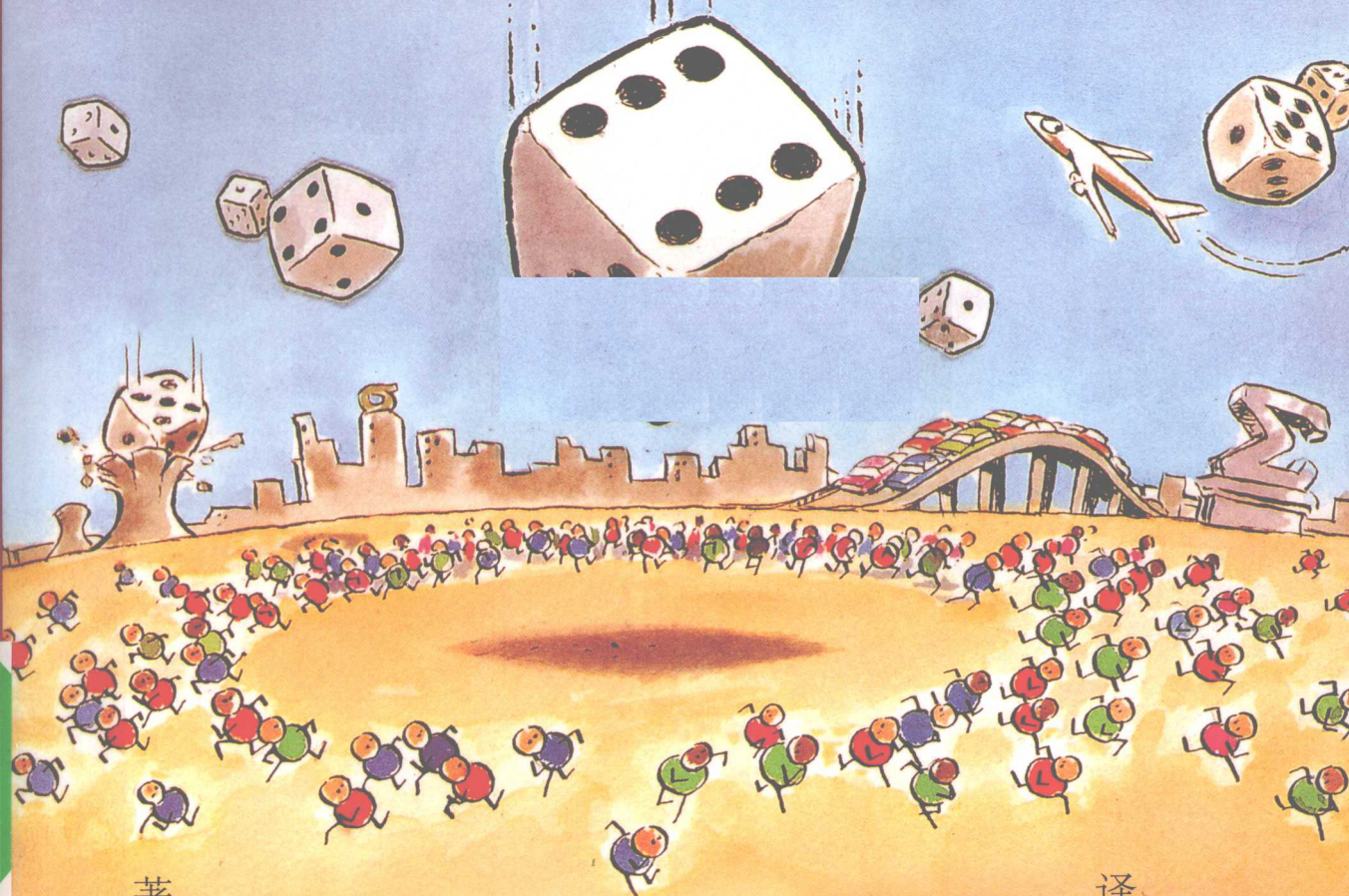
漫画



THE CARTOON GUIDE TO

STATISTICS

玩转统计学



著

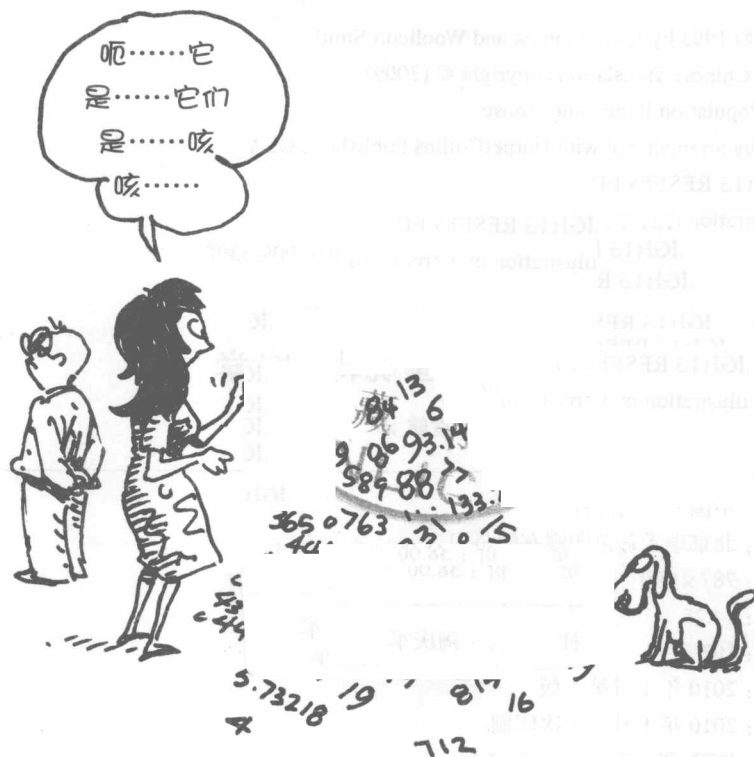
拉里·戈尼克 LARRY GONICK 沃克特·史密斯 WOOLLCOTT SMITH

译

袁航

中国人口出版社

漫画玩转统计学



沃克特·史密斯

拉里·戈尼克

袁航

著

译

中国人口出版社

图书在版编目 (CIP) 数据

漫画玩转统计学 / (美) 戈尼克 (Gonick, L.), (美) 史密斯 (Smith, W.) 著; 袁航译.

—北京: 中国人口出版社, 2010. 1

(漫画趣知系列)

书名原文: The Cartoon Guide To Statistics

ISBN 978-7-5101-0306-3

I. 漫… II. ①戈…②史…③袁… III. 统计学—通俗读物 IV. C8-49

中国版本图书馆 CIP 数据核字 (2009) 第 214483 号

THE CARTOON GUIDE TO STATISTICS by Larry Gonick and Woollcott Smith

Copyright © 1993 by Larry Gonick and Woollcott Smith

Simplified Chinese Translation copyright © (2009)

by China Population Publishing House

Published by arrangement with HarperCollins Publishers, USA

ALL RIGHTS RESERVED

Cover illustration by Larry Gonick

著作权登记号: 01-2009-3390

漫画玩转统计学

[美] 拉里·戈尼克 沃克特·史密斯 袁航 译

出版发行: 中国人口出版社

印 刷: 北京瑞禾彩色印刷有限公司

开 本: 787×990 1/16

印 张: 14.75

字 数: 192 千字

版 次: 2010 年 1 月第 1 版

印 次: 2010 年 1 月第 1 次印刷

书 号: ISBN 978-7-5101-0306-3

定 价: 38.00 元

社 长: 陶庆军

电子邮箱: chinapphouse@163.net

电 话: (010) 83519390

传 真: (010) 83519401

地 址: 北京市宣武区广安门南街 80 号中加大厦

邮政编码: 100054

版权专有 侵权必究 质量问题 随时退换

◆ 目 录 ◆

第 1 章	-----	1
什么是统计学		
第 2 章	-----	7
数据描述		
第 3 章	-----	27
概 率		
第 4 章	-----	53
随机变量		
第 5 章	-----	73
双分布记		
第 6 章	-----	89
抽 样		
第 7 章	-----	111
置信区间		
第 8 章	-----	137
假设检验		
第 9 章	-----	157
比较两个总体		
第 10 章	-----	181
实验设计		
第 11 章	-----	187
回 归		
第 12 章	-----	211
总 结		
参考文献	-----	221
索 引	-----	224

◆ 第 1 章 ◆

什么是统计学

我们一生都在根据**不完整的信息**应付各种选择……

我应该点汤吗？其他的都贵极了，再说，也不知道最后谁结账……不知道搞统计的是不是都很小气？我还从来没和这类人约会过……不过，我倒是认识一个很大方的会计师……

我应该点汤吗？36次中我有27次点了汤，它的味道真不错……但是周一晚上的厨师上班吗？如果所有的空气分子突然涌向天花板，会发生什么事情？



大多数人都能接受一定程度的不确定性。

请给我来份汤。

呃……能给我拿一个计算器过来吗？



统计学的独特之处在于它能够**量化**不确定性，使其精确化。从而，统计学家可以对他们所做判断的不确定程度作出**明确的断言**！

点得好！我 95% 地确信，今晚的汤非常可口的概率在 73% 到 77% 之间！

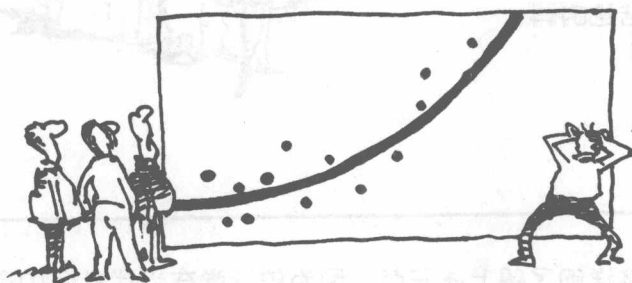


这不仅仅是点不点汤的问题！统计学也常常攸关生死……

嘿！你以前在这样的夜晚点过汤吗？

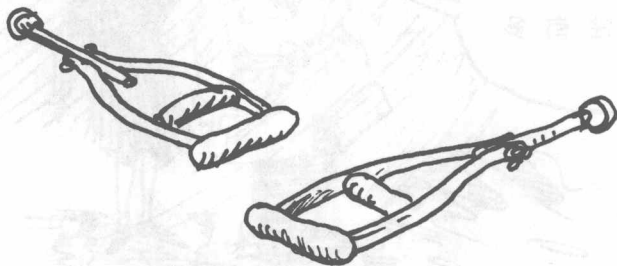


例如，1986年，“挑战者号”航天飞机失事，7名宇航员遇难。当天气温仅为华氏29度，做出发射决定前工作人员却并未对低于这个温度条件下的性能数据做出简单分析。



喔……曲线的那一段！

再说一个正面的例子——脊髓灰质炎疫苗。1954年，约40万名儿童接种了脊髓灰质炎疫苗，接种过程得到了严格控制，排除了存在偏差的结果。精确的统计分析有力地证实了疫苗的效果，因而今天，脊髓灰质炎已几乎不为人知。



要完成这样的数学戏法，统计学家依赖于三个相关原则：

数据分析

数据的搜集、显示与归纳；

概率

赌场内外的随机性规律；

统计推论

运用概率知识，从特定数据中得出统计学结论的科学。



在本书中，我们将详细了解上述三点，因为统计学在当代社会中有着至关重要的作用。



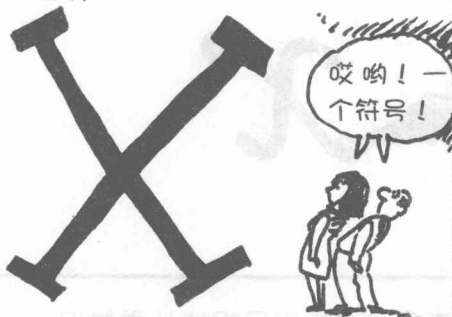
在第2章中，我们要来看一个简单的数据集：一群大学生的体重数据。



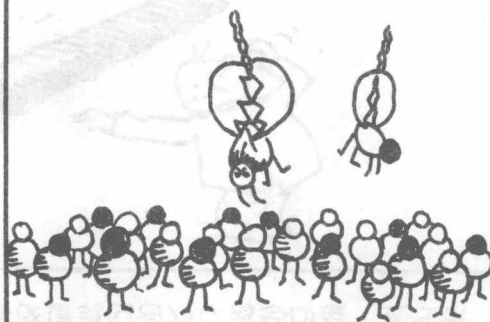
在第3章中，我们要到概率论的发源地——赌场——去研究它的规律。



第4和第5章将介绍如何运用**随机变量**的概念，以**概率模型**来描述我们的世界。



第6章介绍了一种基本的统计过程——从总体中**取样**。

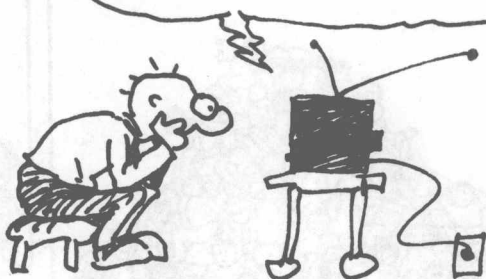


第7章及其后章节，我们将结合各种现实场景（如**选举民意调查**，**生产质量控制**，**医疗检测**，**环境监测**，**种族歧视**，以及**法律**）来解释统计推论。

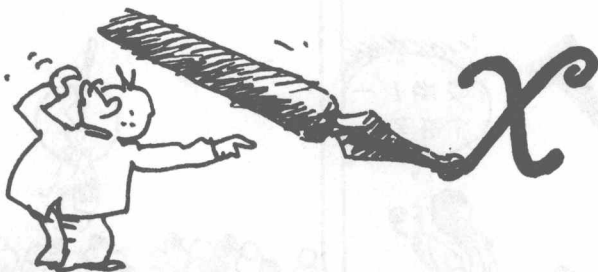


最后，在讨论统计学的过程中不能不提到另一个问题：当今社会广泛存在的对统计学的不信任。每个人都知道“统计数字会说谎”，在日常生活中鲜见好的统计学分析。我们该怎么办？

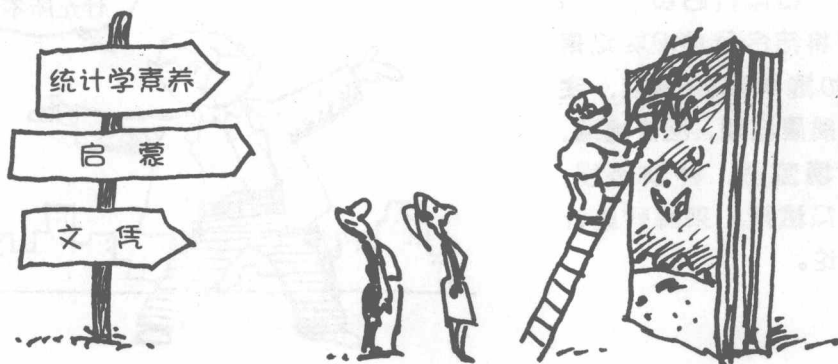
75% 的医生建议不要相信任何以“75% 的医生建议……”之类的语句开头的论断。



依我们的拙见：多学一点统计学知识也许并没有坏处……这也是我们写本书的目的！



接下来，我们会努力以尽可能直观和形象的方式来呈现统计学知识。你所要做的只是有一点耐心，多一些思考，以及对于代数的容忍——如果不具备这几项，那么还是去修一门专门的课程吧！



◆ 第2章 ◆

数据描述



数据是统计学家的素材，是我们用来解读现实的基础。所有的统计学问题不外乎对数据的搜集、描述以及分析，或者是对这些过程的思考。

我搞清楚这东西的机会是 92%……



8
4
1881997
 $\sqrt{5}$
2676
7.3
3.14159
11043
19.7
516
9.99
4.33
75
85
95
101
15.1

本章关注于数据描述。我们如何有效地呈现数据？我们如何从一堆原始数字中找出隐藏的规律？我们如何归纳出数据的基本形态？

怎么才能在它搞定我们之前搞定它？



2² 6⁴ 13
14.1 8 15 3 21
19 08 35 19.9 8 66
08 04 61 28 54
15.3 44 219 06 7 18 22
15 19 19.7 10.01 517 4 4 1 0.1
6 21 10 8 21.37 19 6 53 12 5 7 18 3
66 5 5 5 30.6 55 56 99 9 6
1 1 7 > 11

呃，为了描述数据，首先你必须找到一些真实的数据！



同学们！

这里是一些真实的数据：作为课堂实验的一部分，92名来自宾夕法尼亚州的学生报告了他们的体重，结果如下：



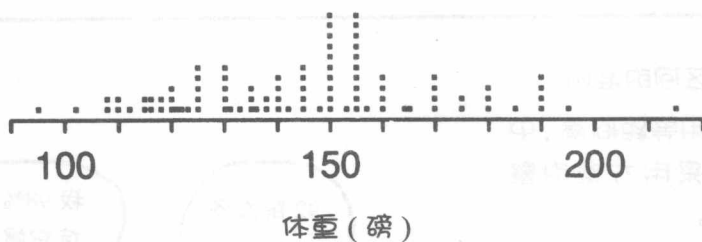
男性：

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180 135
170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160 135 160
130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

女性：

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135 125
118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

言归正传，我们绘制出一张点阵图：每个学生的体重对应一个点。



你可能会发现一个问题：体重**150磅**和**155磅**的人数最多。学生倾向于按**5磅的增量**报告他们的体重。在这样的现实情况下，这类舍入误差可能导致数据的规律被隐匿……对此，我们也有相应的解决措施。

我们可以用一个**频率表**来汇总数据。将数据划分为若干个区间，并对每个区间内学生的体重进行计数。**频率**是每个区间的计数值。**相对频率**是每个区间体重所占的比例，也就是说，它等于频率除以学生总数。

等级区间	中点	频率	相对频率
97.5-102.4	95	2	.022
102.5-117.5	110	9	.098
117.5-132.4	125	19	.206
132.5-147.4	140	17	.185
147.5-162.4	155	27	.293
162.5-177.4	170	8	.087
177.5-192.4	185	8	.087
192.5-207.5	200	1	.011
207.5-222.4	215	1	.011
总计		92	1.000

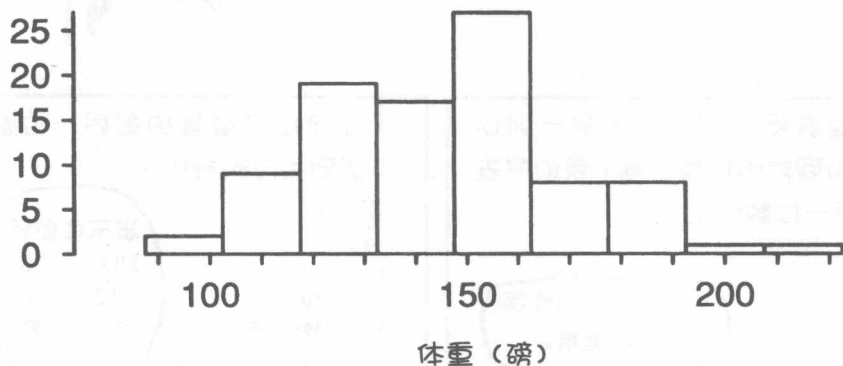
注：我们有意使间隔避开那些棘手的 5 磅增量，这样做可以解决学生的报告偏差问题。

构建等级区间的准则：

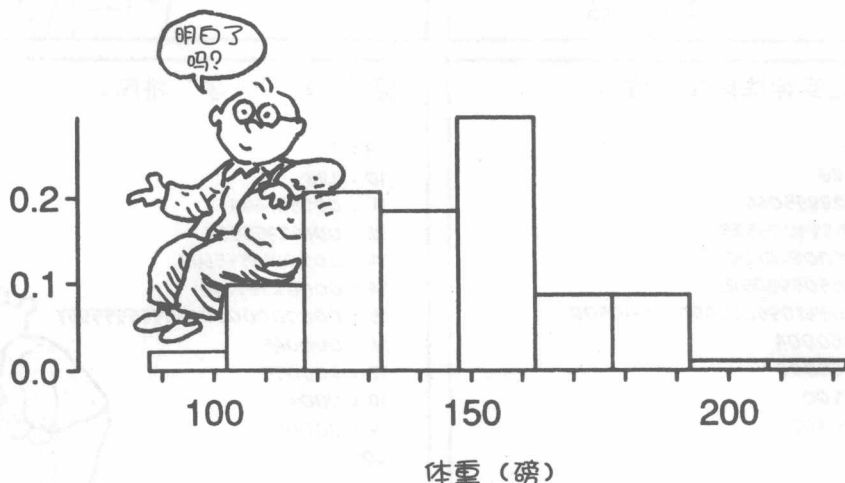
- 1) 使用等距间隔，中点采用方便的整数。
- 2) 对于小数据集，使用小的间隔。
- 3) 对于大的数据集，使用更多的间隔！



在频率表中，我们展示了每个数值“周围”有多少个数据点。我们也可以图示这一信息。柱形图的结果被称作**直方图**。每个长条代表一个间隔，并以中点为中心。长条的高度代表该区间的数据点数量。



我们也可以绘制一张**相对频率直方图**。除了纵坐标单位不同以外，它看起来和上图完全相同。



统计学家约翰·塔基发明了一种能够保留独立数据点的快速汇总数据的方法，它被称为**茎叶图**。



以体重数据为例，茎就是一列以10磅为间隔的体重数据（我们舍去了最后一位数字）。

9
10
11
12
13
14
15
16
17
18
19
20
21

即，90磅，100磅，
以此类推。



现在将每个重量的最后一位数字放置到相应的行中：

茎：叶

9 :
10 :
11 : 628
12 : 0155005
13 : 080015
14 : 05
15 : 0
16 :
17 :
18 :
19 :
20 :
21 :

表示体重分
别为：116、
112、118、
120，等等。



全部填写完毕后是这样的：

9 : 5
10 : 288
11 : 628855060
12 : 01553005525
13 : 8500850600153
14 : 05505580502
15 : 50537055055050500500
16 : 050004
17 : 055000
18 : 0500
19 : 00500
20 :
21 : 5

最后，对“叶子”排序。

9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 00002555558
15 : 00000000035555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5



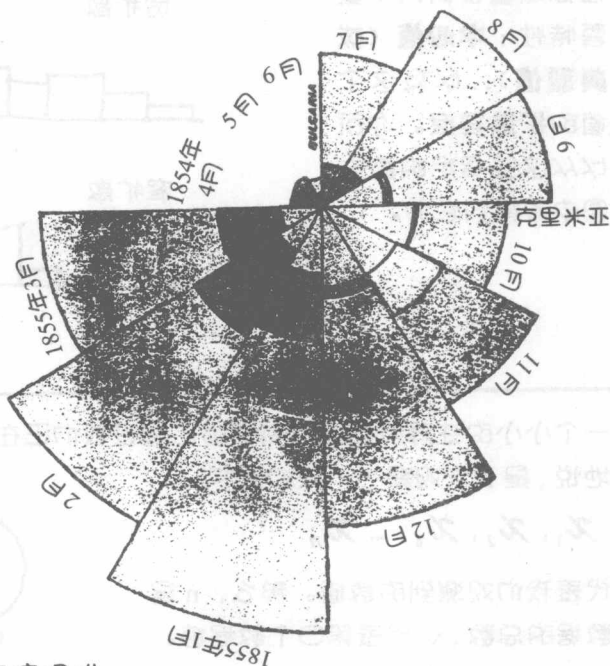
数值0和5清晰地显示出学生的报告偏差！

好的图示，部分是艺术，
部分是科学。



有时候甚至还与政治
有关！

战地护士**弗洛伦斯·南丁格尔**依照英国军队医院的**死亡率统计**，制作了令人震惊的直方图：径向轴代表在克里米亚战争中英国士兵在医院和战场上的死亡人数。



她在统计上所做的努力直接促进了医院条件的改善以及死亡率的下降。



统计学救了
我的命！