欧姆社学习漫画

漫画统计学

〔日〕高桥 信/著
〔日〕株式会社TREND-PRO/漫画制作
陈 刚/译

科学出版社
www.sciencep.com

KindleDX出版署

✿前　言✿

本书是统计学的入门书。

我将读者对象预设为下列两大类：

● 写作毕业论文或在工作中须进行资料分析者

● 虽然现在没有分析资料的需求，但想一窥统计学的奥妙者。当然也非常欢迎对统计学稍有涉猎的读者。

统计学是与“生活”及“工作”有密切关系的一门学科。如果能够掌握统计学知识，那么你的生活将会变得更加方便，例如：

● 可预测校庆时推出的炒面可卖出几份

● 可预测资格考试可否通过

● 可比较投入药剂 X 和不投药剂 X 两种情况下的存活率

本书共分 7 章。各章原则上由下列部分构成。

● 漫画部分

● 补充漫画部分的解说

● 例题和解答

● 总整理

但是也有某些章节并不遵循上述构成方式。

读者即使仅阅读漫画部分，也可逐渐了解统计学概念。如果再阅读其他部分，则可增加知识掌握的深度。

“统计学可真是有趣而实用呀！”若各位在读完本书后能有这样的感受，我将感到荣幸之至。

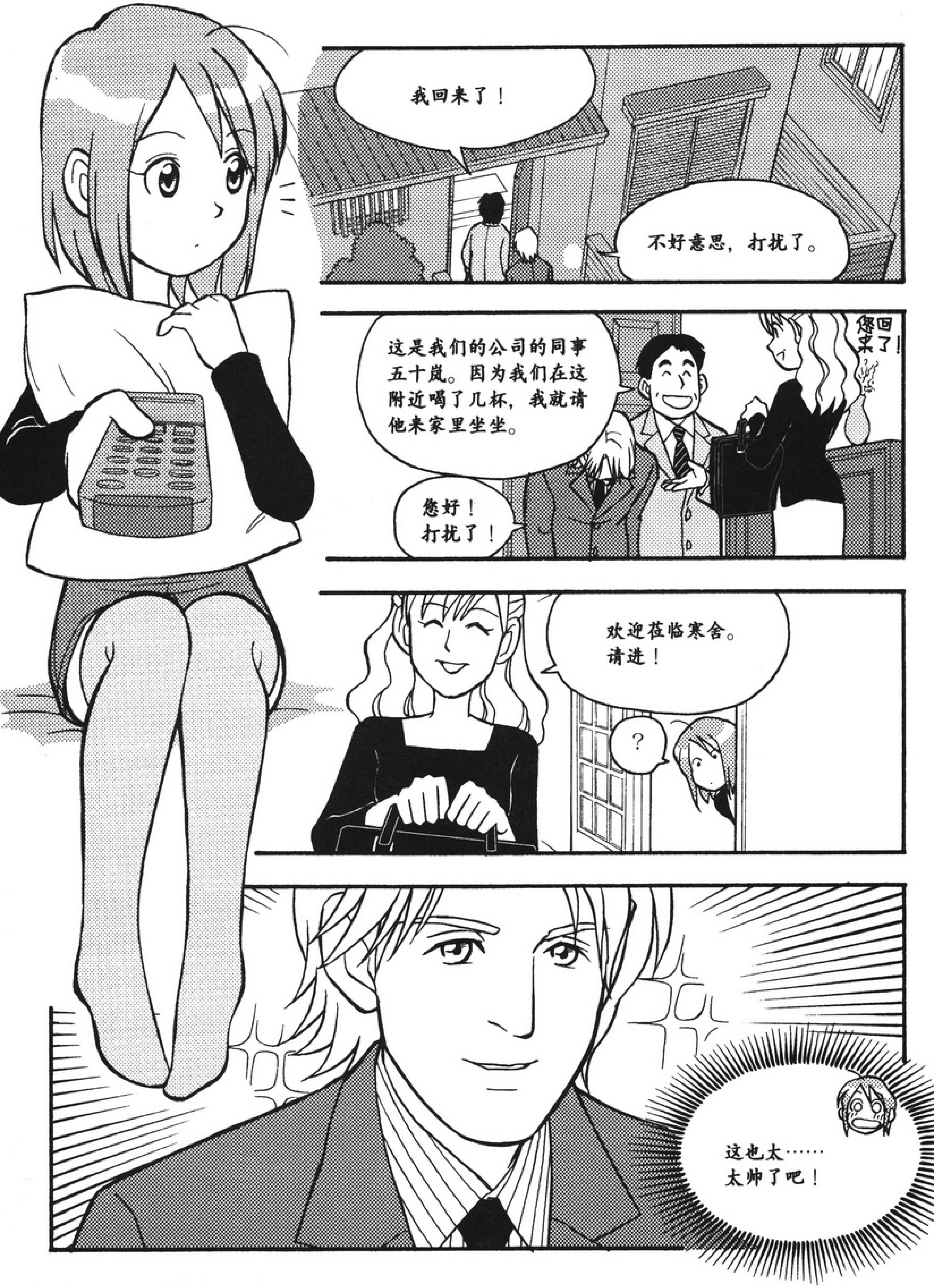
感谢欧姆社的各位编辑，能给我这次机会著作此书。同时也感 TREND-PRO 股份有限公司的各位漫画作者，有了他们的努力，我的原著才得以转换成漫画形式。另外，还有负责脚本创作的 re-akino，负责做画的 Inoue Iroha。此外，还要感谢在我著作之际，为我提供多方建议的日本立教大学社会学系的酒折文武老师。

高桥　信

✿目　录✿

序　章

令人悸动的统计学


我回来了！
不好意思，打扰了。
您回来了！
这是我们的公司的同事五十岚。因为我们在这附近喝了几杯，我就请他来家里坐坐。
您好！打扰了！
欢迎莅临寒舍。请进！
？
这也太……太帅了吧！

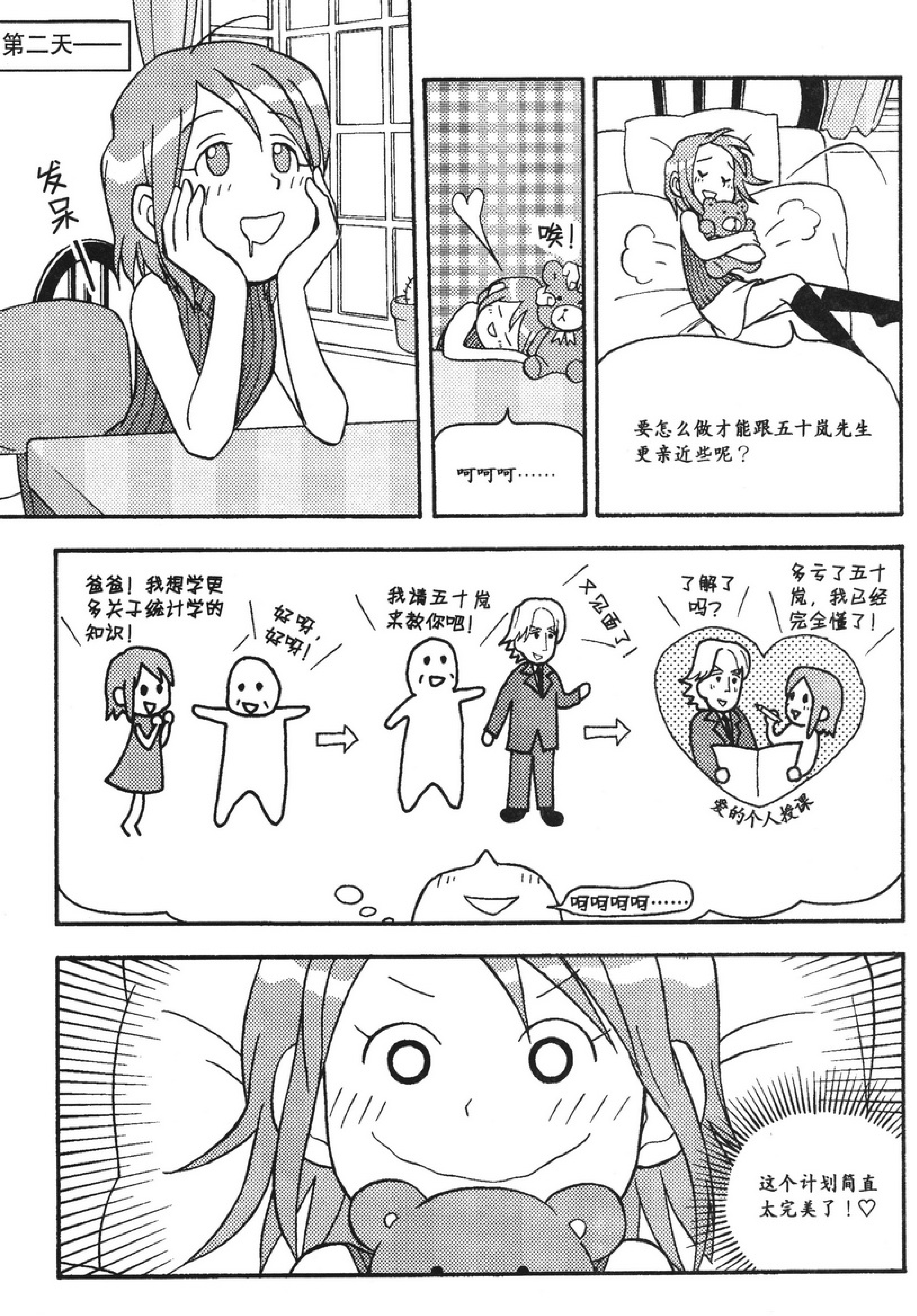
我回来了，琉衣，这是我们公司的同事，五十岚。
晚安！
请用！♡
请用！♡
谢……谢谢！
请慢用！♡
令媛真是可爱。
哎呀！
不要这样夸人家嘛……
虽然你说的是事实！
那么，五十岚先生，您在哪一个部门工作啊？
和……和令尊同一部门。
简单来说，就是负责营销工作！
营销


具体来说，就是利用统计学知识来做市场调查……
你还在上高中，即便我这么说，我想你应该还是不知道什么是“营销”，对吧！
嗯！的确不懂。
真是好直接啊！那么“统计学”呢？
嗯……
大概不是很清楚吧？所谓的统计学，粗略来说，就是从样本反应出的信息中推测总体状况的学问。
好像有点太难了！
喂，琉衣！
啊！有了。
正好今天的晚报有刊载内阁支持率的内容呢！


"根据朝每晚报的调查结果，内阁支持率为39%。"
这是怎么算出来的呢？
朝每晚报并没有来询问我支持谁啊？
高津先生呢？
没，
也没有人来问我啊！
嗯！
明明你们都没有直接接受调查，但朝每晚报却还能算出内阁支持率？
而且你们也都有投票权呀！是不是有点奇怪啊？
是的，此处就应用了统计学知识呦！
那……那么？
琉衣，日本大约有多少人有投票权呢？
JAPAN
嗯……很多！非常多！
是的！

1．总体：Population。　2．样本：Sample。


总之，以内阁支持率为例，总体即为“所有具有投票权的人”。
而这个调查似乎是以2000人为询问对象，因此，这2000人即为样本。
所有具有投票权的人
抽出
2000人
总体
样本
原来如此！
如果可能的话，当然希望调查总体。
100m²
在这片土地上生长得最茂盛的是……
然而，这在现实中是不可能的。真是令人困惑呀！
100m²
超麻烦啊！累死人了！
就算无法进行精细的调查，难道没办法尽可能准确地得知总体的状况吗？
100m²
嗯……
只调查1m²看看吧！
1m²
那么，解决这个问题的最佳方法就是统计学呦！
哦！再多教我一些吧！
哈哈。那就下次吧。
ニコッ
果然很帅啊！


第二天——
发呆
唉！
呵呵呵……
要怎么做才能跟五十岚先生更亲近些呢？
爸爸！我想学更多关于统计学的知识！
好呀，好呀！
我请五十岚来教你吧！
又见面了！
了解了吗？
多亏了五十岚，我已经完全懂了！
爱的个人授课
呵呵呵呵……
这个计划简直太完美了！♡


爸爸……♡
哦！谢谢。
那个……我想要请一位统计学的家庭教师。♡
因为我开始对爸爸的工作产生兴趣了。
激动
琉、琉衣……
好吧！那么，每周六就请家教来上课吧！
嗯！爸爸的同事也可以呦！比如五十岚先生……
包在我身上！
太棒了！
到了星期六——
叮咚！
嗨！辛苦你了！赶快进来。


来了！
喂，琉衣！
我带老师来了
哦——
好的。
タッタッタ……
你好……


琉衣，这是我的同事，山本守。
你好！
你……这家伙是谁呀！
爸，爸爸……五十岚先生呢？
嗯？山本住得比五十岚离咱家近呀！而且教得也比较好呦！
那么，你们好好上课吧！
哈哈哈！

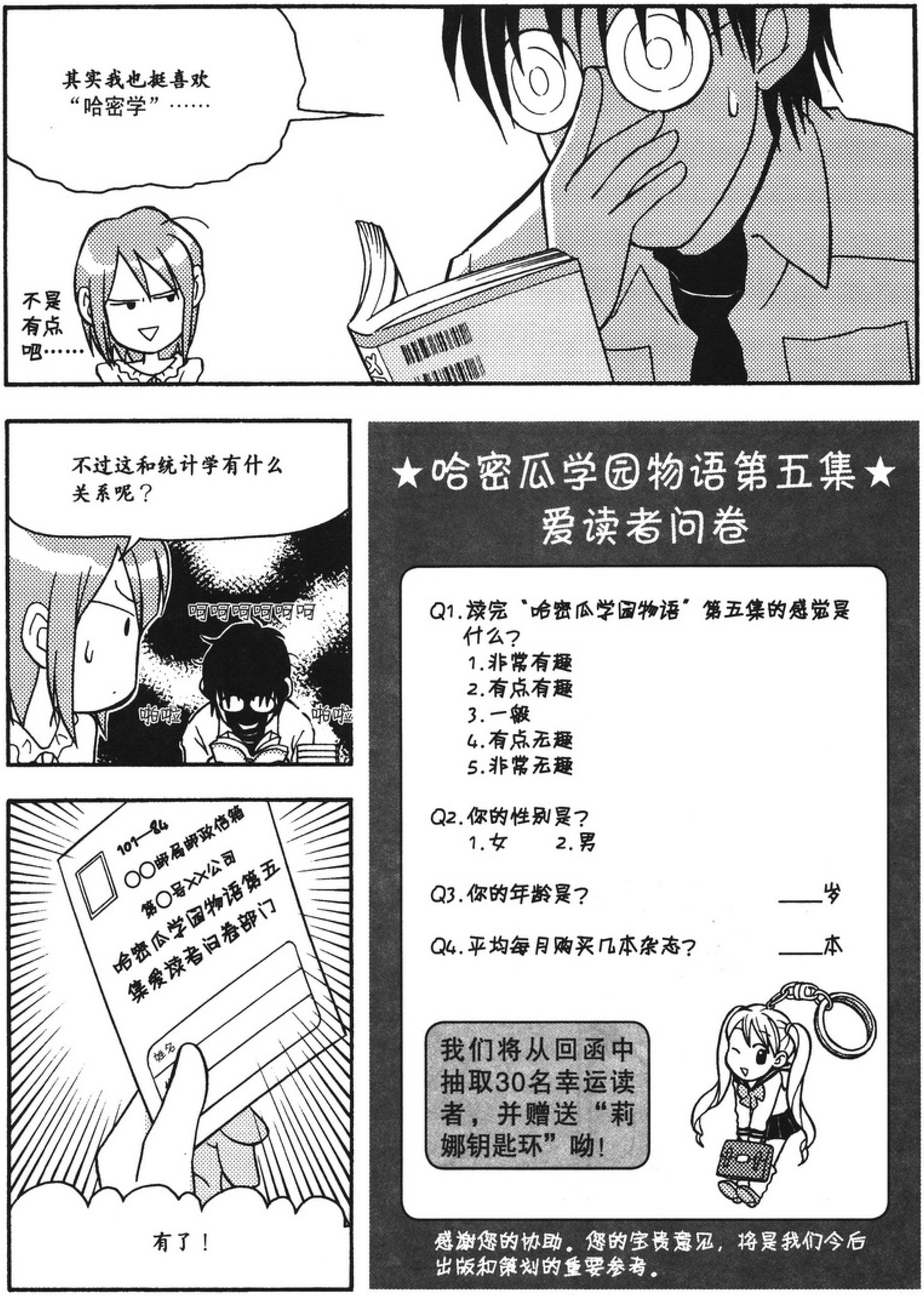

琉衣，请多多指教。
呵呵呵呵呵！
五十岚，琉衣很努力的念书呦！
太棒了！和我一起工作吧！
琉衣？
呵呵呵呵……
ばっ
？
我就踩着这家伙的肩膀来接近五十岚先生吧！
嘿
嘿
请多多指教。
请，请多多指教……
统计学课就这样开始了……

◆第1章◆

确认数据种类

✿ 1. 分类数据和数值数据 ✿

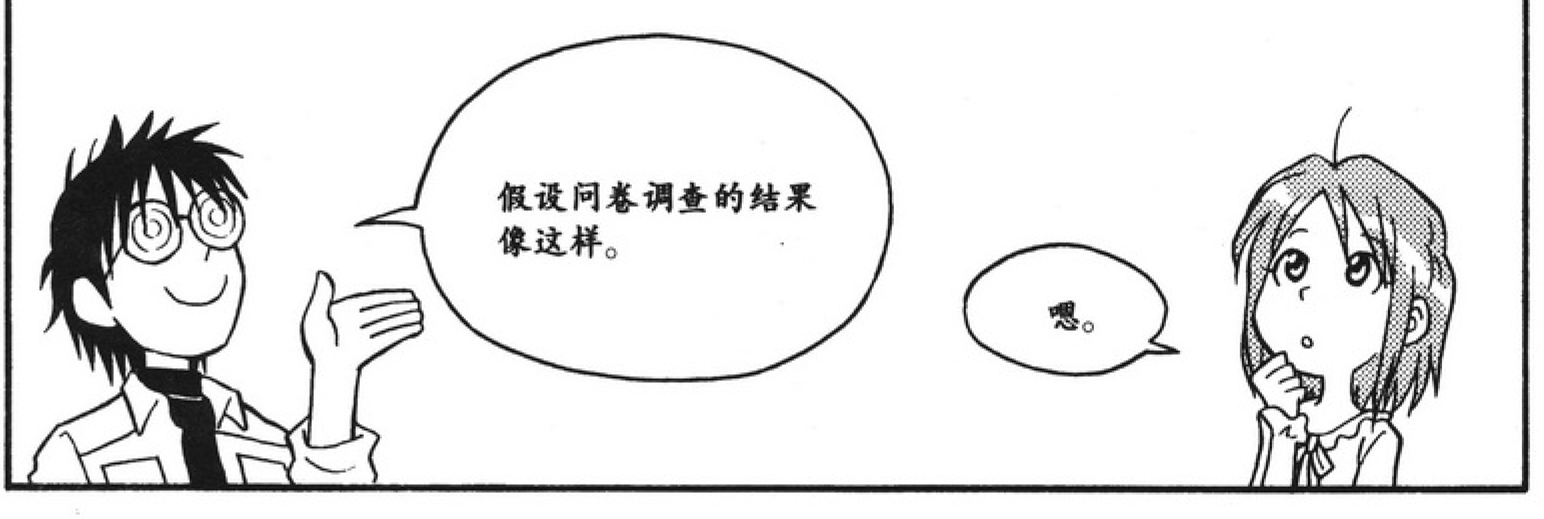
那么，
山本老师，我们从什么地方开始学起呢？
我想想——
既然才刚开始，就从简单的学起吧！
张望
琉衣，你喜欢“哈密瓜学园物语”吗？
プル
プル
嗯！
我收集了整套呢！


其实我也挺喜欢“哈密学”……
不是有点吧……
不过这和统计学有什么关系呢？
呵呵呵呵呵呵
啪啦
啪啦
★哈密瓜学园物语第五集★
爱读者问卷
Q1.读完“哈密瓜学园物语”第五集的感觉是什么？
1.非常有趣
2.有点有趣
3.一般
4.有点无趣
5.非常无趣
Q2.你的性别是？
1.女　2.男
Q3.你的年龄是？　____岁
Q4.平均每月购买几本杂志？　____本
我们将从回函中抽取30名幸运读者，并赠送“莉娜钥匙环”呦！
感谢您的协助。您的宝贵意见，将是我们今后出版和策划的重要参考。
101—84
○○邮局邮政信箱
第○号XX公司
哈密瓜学园物语第五集爱读者问卷部门
有了！


这个问卷是要做什么的呢？
寄出这个问卷的话……
推眼镜
笨蛋！亏我这么相信你。
我……会努力的。
我做了这个。
因为我喜欢你嘛！
（如果中奖）可以得到莉娜（女主角）的钥匙环哟！
原型制作：TAIYODO
超想得到耶……
啊！是……哦！
那么，我们进入主题吧！
已经开始上课了呀！
首先请回答这份问卷。
好吧！
做好了！
好的，非常感谢！

从读者处获得的数据

	Q1 读"哈密学"的感觉	Q2 性别	Q3 年龄（岁）	Q4 平均一个月购买的杂志数（册）
琉衣	非常有趣	女	17	2
A	有点有趣	女	17	1
B	一般	男	18	5
C	有点无趣	男	22	7
D	有点有趣	女	25	4
E	非常无趣	男	20	3
F	非常有趣	女	16	1
G	有点有趣	女	17	2
H	一般	男	18	0
I	一般	女	21	3

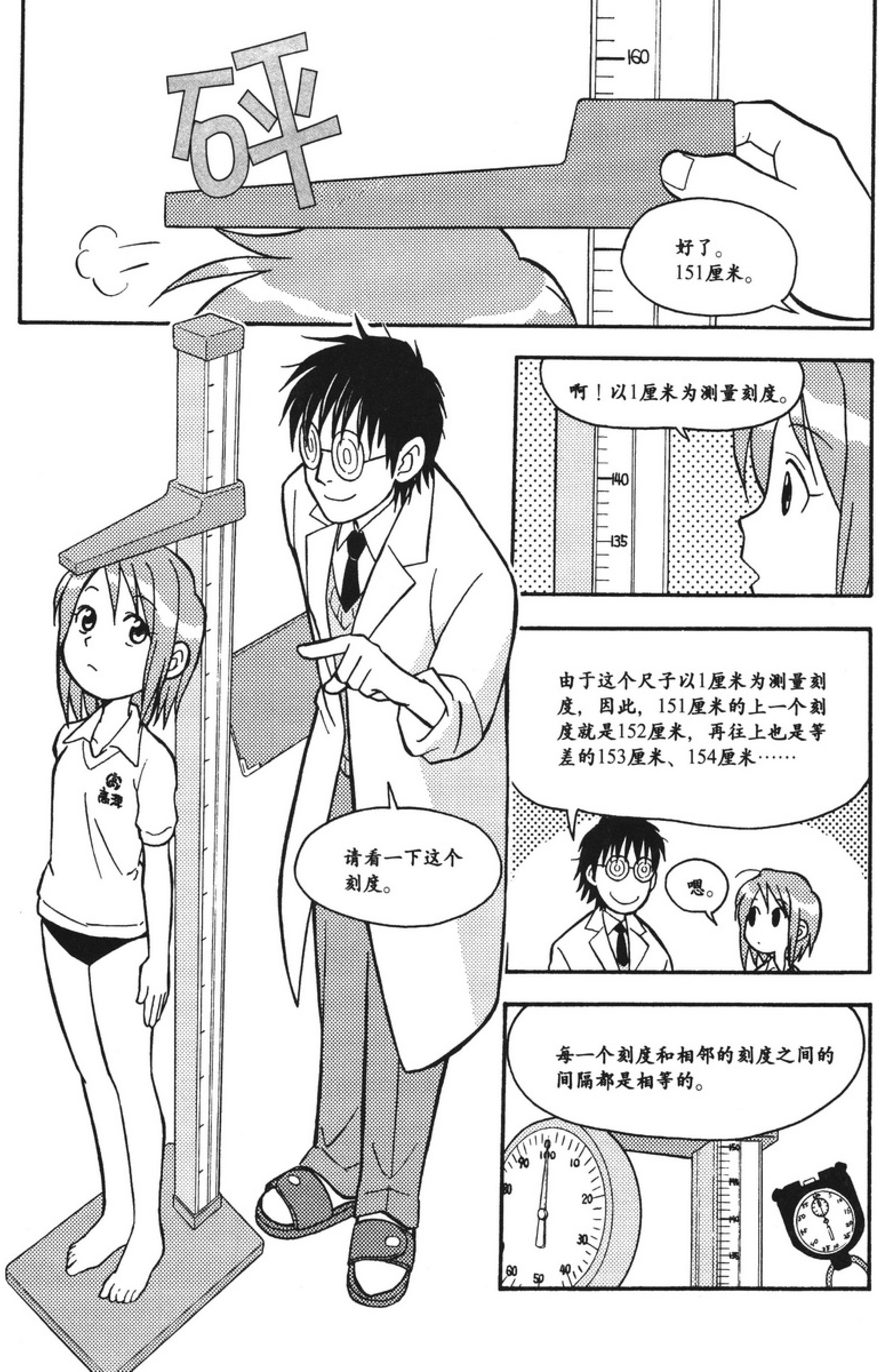



虽然我很在意"哈密学"的评价！
哈密学宅男
咚
但先不提这个，今天课程的重点在于数据分类。
什么意思？
数据分为"不可测量"的数据和
"可测量"的数据。
不可测量
可测量

1．分类数据：Category Data或Categorical Data。 2．数值数据：Numerical Data。

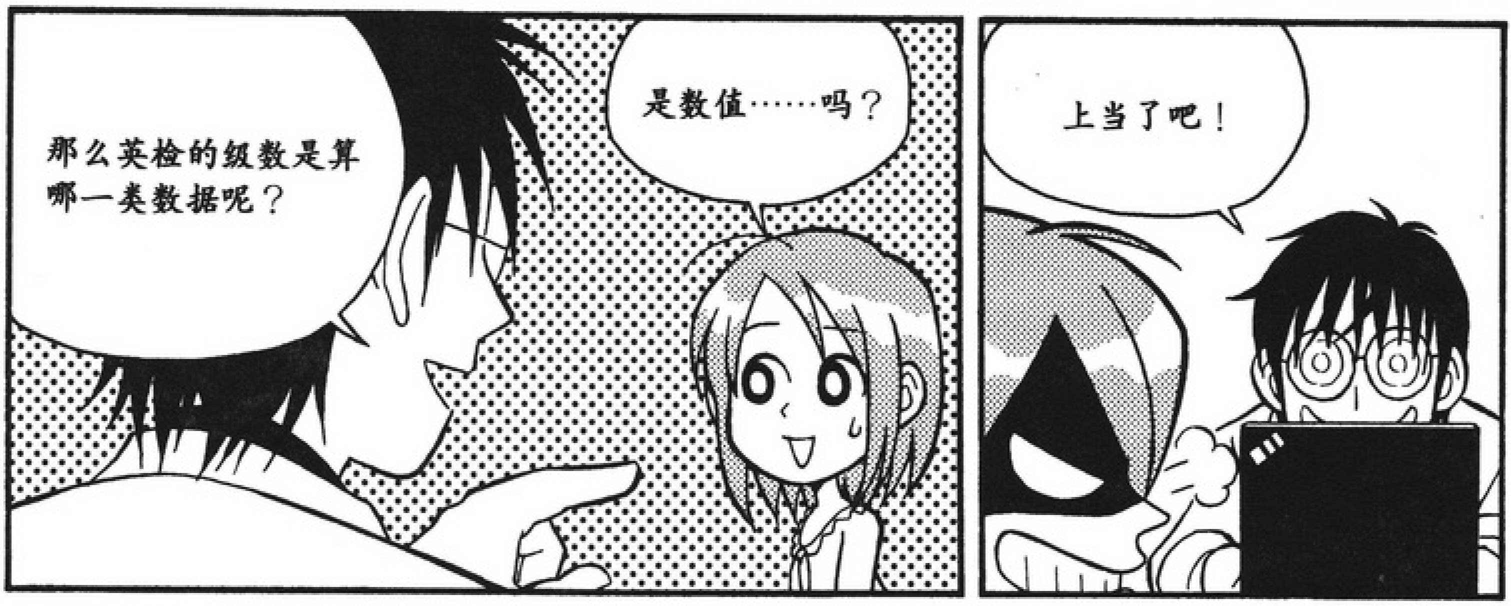
2. 分类数据注意事项举例


琉衣……
怎么了？
你体重多少？
くわっ
怎么可以问女生这种超级没礼貌的问题呢！
对不起！
那么身高多高？
151厘米。
先问这个不就好了吗？
太好了！你不介意……
我们使用尺子测量身高，对吧？
这还用说吗，那是当然了。



砰
好了。151厘米。
请看一下这个刻度。
啊！以1厘米为测量刻度。
由于这个尺子以1厘米为测量刻度，因此，151厘米的上一个刻度就是152厘米，再往上也是等差的153厘米、154厘米……
嗯。
每一个刻度和相邻的刻度之间的间隔都是相等的。

因此，“身高”是“可测量”的数据，也就是数值数据。
哦！
知道了！
再来看下一个例子吧！
什么？
什么？
かぱ
嗯……
Coogle
学園
物語♡
哈密学的桌面
琉衣应该已经通过财团法人日本英语考试协会的实用英语技能考试了吧？
高津告诉我的……
那是什么呀？
通称“英检”。
有！
我有准2级的证书。

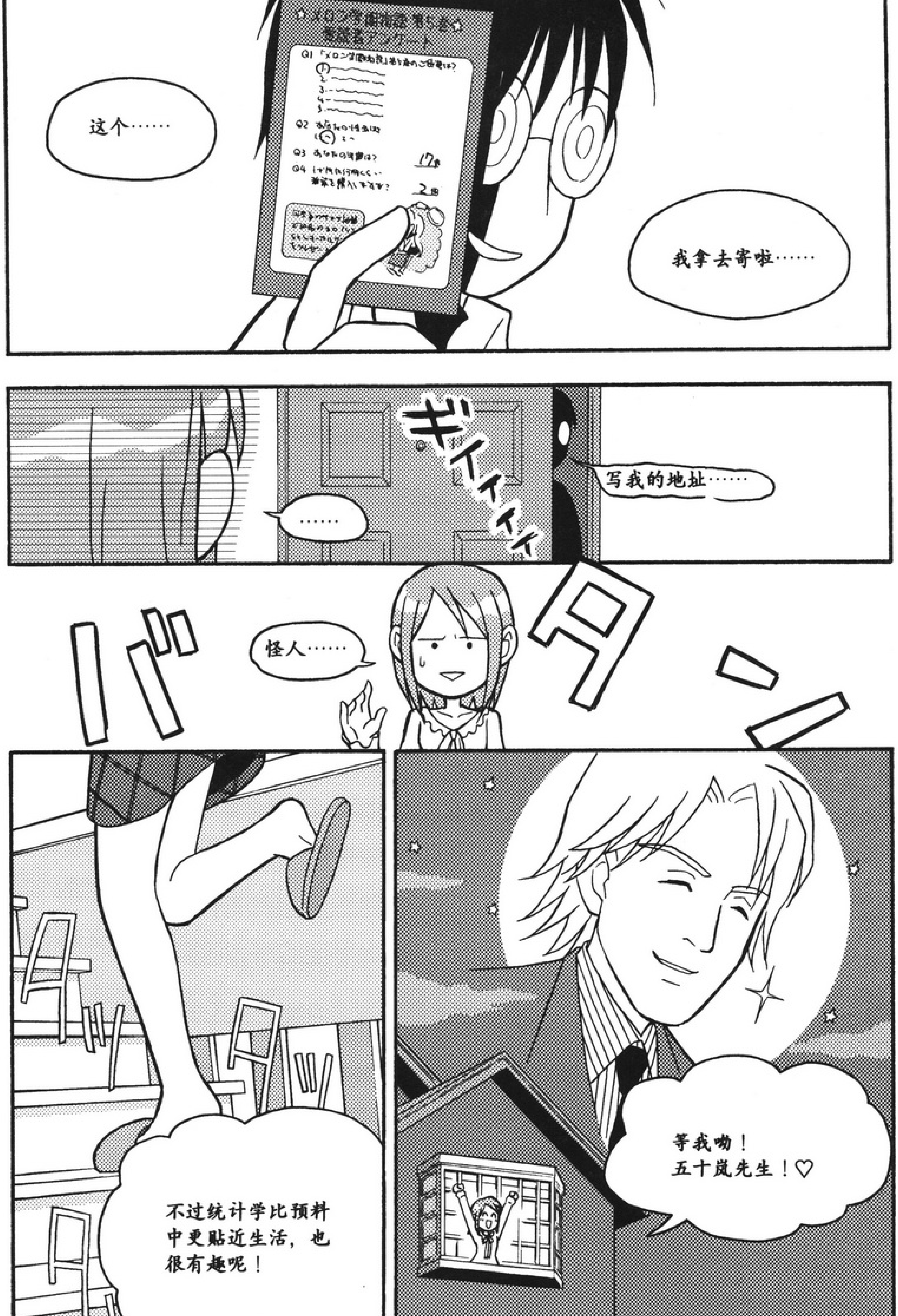


英检难易度的基准				
1级	2级	3级	4级	5级
大学 高级程度 （约 10,000～15,000 单词量）	高中 毕业程度 （约 5,100单词量）	中学 毕业程度 （约 2,100单词量）	中学 中级程度 （约 1,300单词量）	中学 初级程度 （约 600单词量）

（摘自财团法人日本英语考试协会http://www.eilken.or.jp）

所以级数之间的间隔并不相等。
也就是说……
英检的等级是"不可测量"的数据，也就是分类数据！
我明白了！
那么，你应该已经完全了解了吧！
★哈密瓜学园物语第五集★
爱读者问卷
Q1.读完"哈密瓜学园物语"第五集的感觉为……
1. 非常有趣
2. 有点有趣
3. 一般
4. 有点无趣
5. 非常无趣
Q2.你的性别是？
1 女 2. 男
Q3.你的年龄为？
Q4.平均每月购买几本杂志？
17 岁
2 本
这份读者问卷的Q1中，每个选项间的间隔相等吗？
嗯。不相等！
所以是分类数据。
那种东西因人而异吧！
没错！


那么，小测验开始啦！
气温是？
数值数据。
出生地是？
分类数据。
柔道的段位是？
分类数据。
体重是？
数值数据。
“哈密学”的发行集数是？
数值数据。
天气是？
分类数据。
非常好！
今天的课就上到这里。
谢谢你！
那么，下周见。
啊！琉衣……


这个……
我拿去寄啦……
写我的地址……
……
ギイイイ
怪人……
バタン
不过统计学比预料中更贴近生活，也很有趣呢！
等我呦！
五十岚先生！♡

✿3. 实务中“非常有趣”～“非常无趣”的运用✿

正如25页中所述，“*Q*1.读完‘哈密瓜学园物语’第五集的感觉为……”是分类数据。然而，实际的消费者问卷调查中，数值数据并不少见。也就是

非常有趣	⇨	5分
有点有趣	⇨	4分
一般	⇨	3分
有点无趣	⇨	2分
非常无趣	⇨	1分

或是

非常有趣	⇨	2分
有点有趣	⇨	1分
一般	⇨	0分
有点无趣	⇨	–1分
非常无趣	⇨	–2分

以这种方法解释数据的情况并不少见。

理论的世界和实际的世界，不，客套话的世界和真心话的世界也应该存在这样的区别。无论如何，希望各位知道，若观点不同，则数据的获得方式也有可能会不同。

例题和解答

例题

请注意下表:

	血型	对运动饮料X的评价	开空调令人感到舒适的室温(℃)	100米的短跑成绩(秒)
A同学	B	难喝	25	14.1
B同学	A	好喝	24	12.2
C同学	AB	好喝	25	17.0
D同学	O	普通	27	15.6
E同学	A	难喝	24	18.4
⋮	⋮	⋮	⋮	⋮

请将"血型"、"对运动饮料X的评价"、"开空调令人感到舒适的室温"、"100米的短跑成绩"分为分类数据或数值数据。

解答

"血型"和"对运动饮料X的评价"为分类数据。"开空调令人感到舒适的室温"和"100米的短跑成绩"为数值数据。

总整理

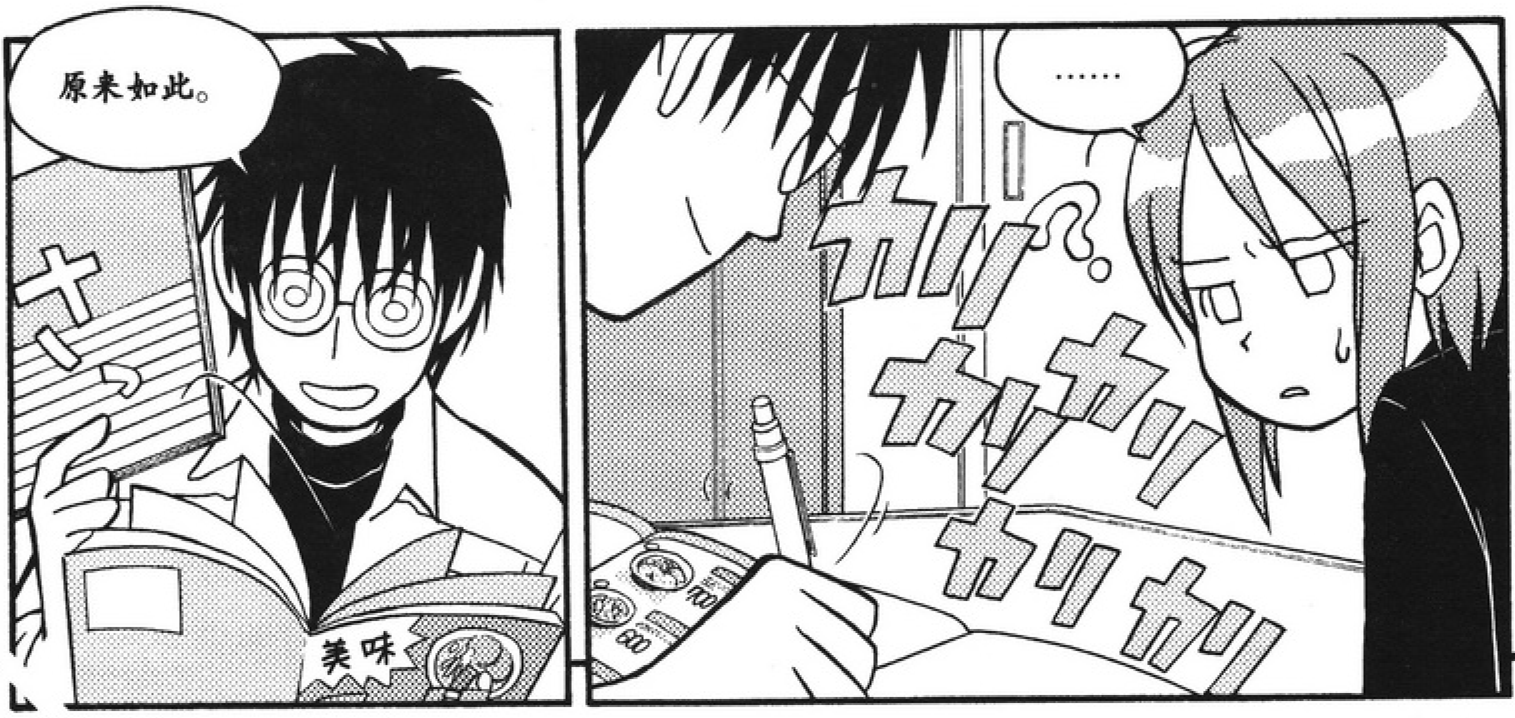
- 数据可分为分类数据和数值数据。
- "非常有趣"~"非常无趣"等，在理论上为分类数据。然而，在实务上，却经常将其视为数值数据。

◆第 2 章◆

掌握数据整体的状态

（数值数据篇）


1. 次数分布表和直方图
真
吼……
你好，琉衣。
美味
拉面
畅销前50
我超想吃呀！
看了这本杂志，不知道该选哪家，好难选哦！
喔！
你好！山本
咦？
你喜欢吃拉面呀？
看起来全部都超好吃的，对吧？
500
890
880
700
890

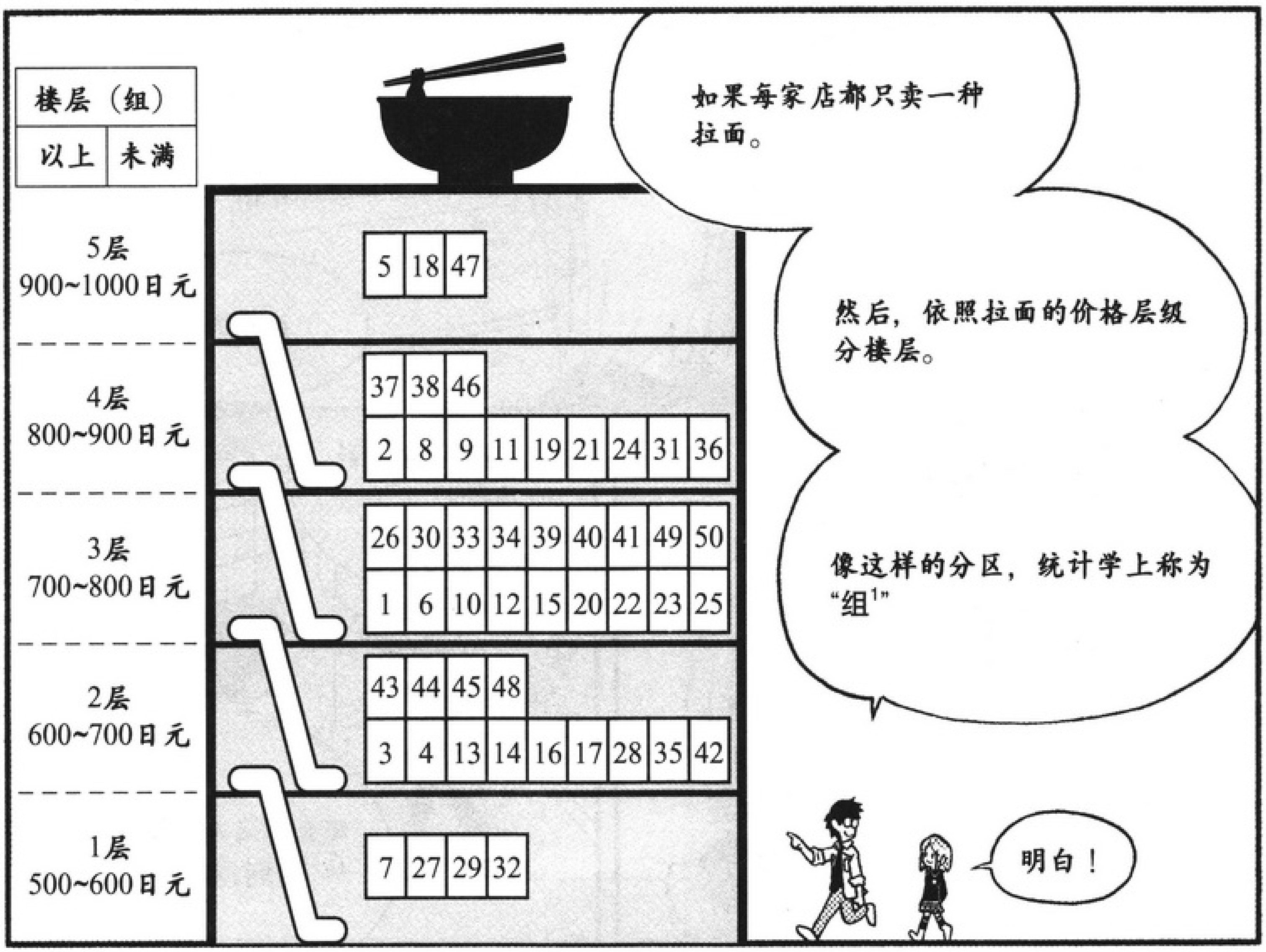


“美味拉面畅销前50”上刊载的拉面馆的拉面价格

	价格（日元）		价格（日元）
拉面馆1	700	拉面馆26	780
2	850	27	590
3	600	28	650
4	650	29	580
5	980	30	750
6	750	31	800
7	500	32	550
8	890	33	750
9	880	34	700
10	700	35	600
11	890	36	800
12	720	37	800
13	680	38	880
14	650	39	790
15	790	40	790
16	670	41	780
17	680	42	600
18	900	43	670
19	880	44	680
20	720	45	650
21	850	46	890
22	700	47	930
23	780	48	650
24	850	49	777
25	750	50	700


看了这个表格，有什么想法吗？
嗯……
真想吃拉面！
就没有其他什么想法吗？
耶——
即使你这样问，我也回答不出来……
我顶多只觉得价格很乱呀！
光看数字的排列，肯定无法了解更多信息。
要怎样才能让这个表格变得更容易理解呢？
嗯……
按某一标准重新整理一下吧！
没错！
为了方便整理，首先，以价格为标准进行整理！
嗯？


也就是说……
请想象一下，一家集结了50家拉面馆的百货公司吧！
拉面万岁
拉面
クイッ
太……太棒了！
怎么变成了电梯小姐？


楼层（组）
以上 未满
5层 900~1000日元
5 18 47
4层 800~900日元
37 38 46
2 8 9 11 19 21 24 31 36
3层 700~800日元
26 30 33 34 39 40 41 49 50
1 6 10 12 15 20 22 23 25
2层 600~700日元
43 44 45 48
3 4 13 14 16 17 28 35 42
1层 500~600日元
7 27 29 32
如果每家店都只卖一种拉面。
然后，依照拉面的价格层级分楼层。
像这样的分区，统计学上称为“组[1]”
明白！

1．组：Class。

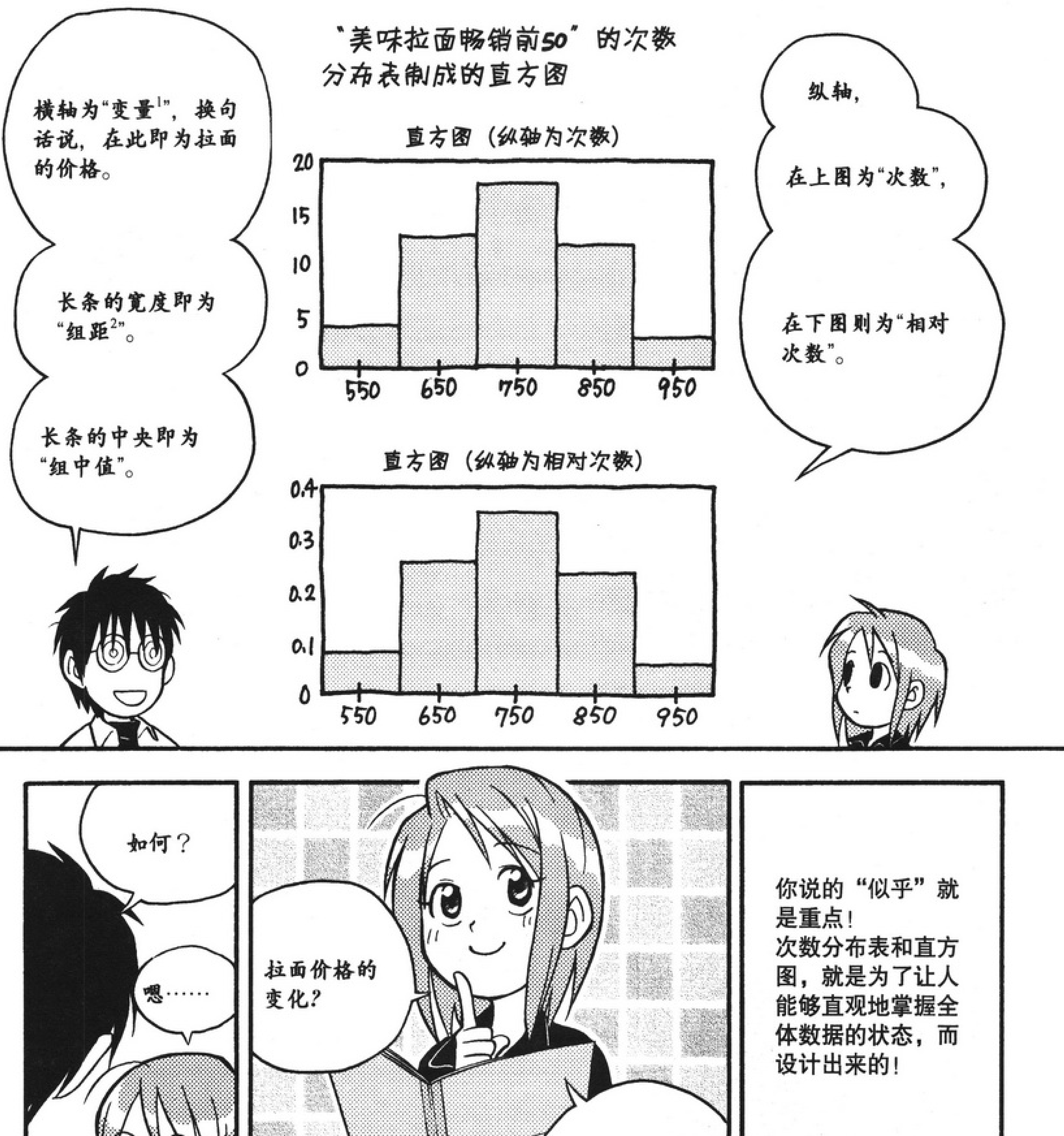
1．组中值：Class Midpoint。 2．次数：Frequency。 3．相对次数：Relative Frequency。


相对次数？
就像我们平常在使用的百分比一样。
将全体视为1的比例。
算式就是这样。
相对次数=所属各组的数据个数/数据总数
嗯……3楼有18家店铺，总共有50家，所以……
心算心算心算心算心算心算心算心算心算
18/50=36/100=0.36吧！
没错！属于700~800日元的这组，也就是组中值为750日元的拉面馆，其相对次数为0.36，乘以100后，以百分比来表示，就是36%。
噢……又变成数学问题了。

“美味拉面畅销前50”的次数分布表

组 以上	未满	组中值	次数	相对次数
500～600		550	4	0.08
600～700		650	13	0.26
700～800		750	18	0.36
800～900		850	12	0.24
900～1000		950	3	0.06
合计			50	1.00

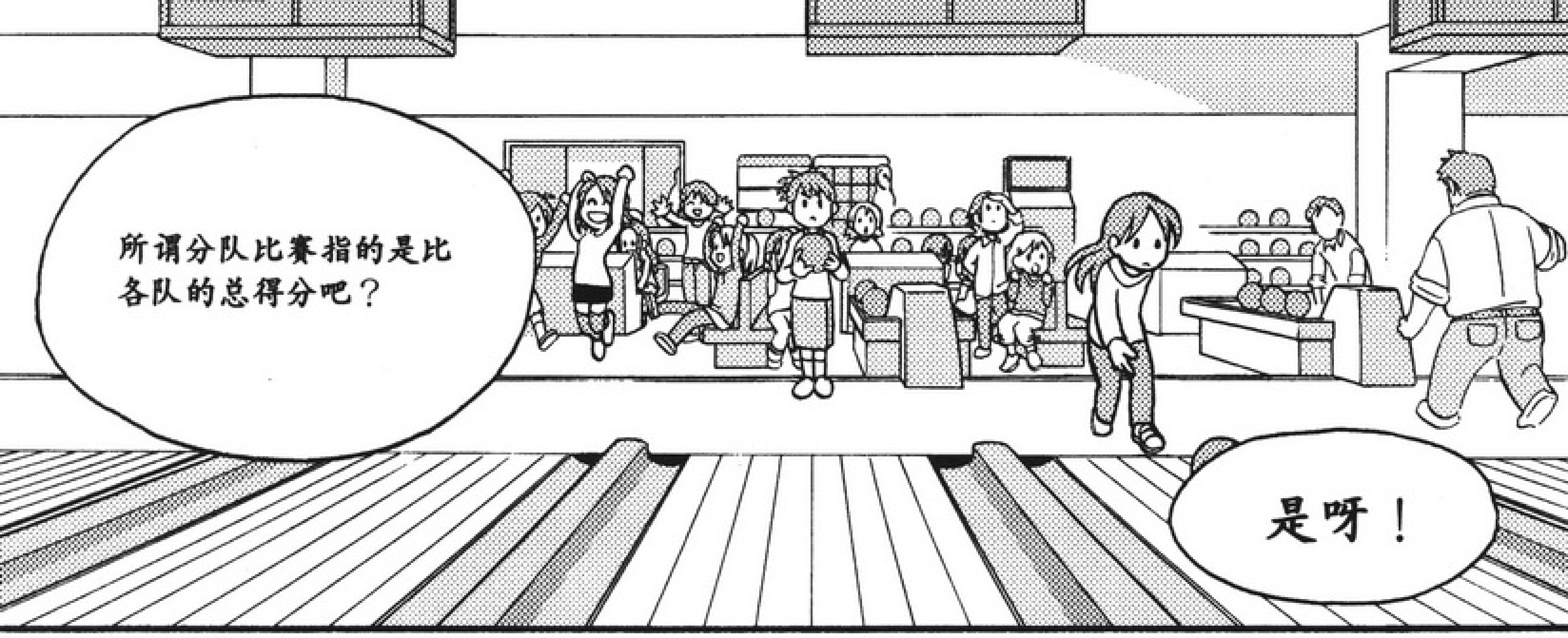
1．次数分布表：Frequency Distribution Talbe。 2．直方图：Histogram。



1．变量：Variable。　2．组距：Class Width。


✿2. 平均数✿
前一阵子我们全班的女生一起去打保龄球（bowling）了！
休息中
挖洞（boring）呀……
有高中女生会做这种事吗?
山本老师,你到底几岁呀?
我糊涂了。
全班的女生，那人不是很多呀！
A
B
C
是呀，总共有18人，所以每6人一组，分成3组做对抗赛呦！
你看！这是当时的得分表！

保龄球大赛的结果
A队 得分 / B队 得分 / C队 得分
琉衣琉衣 86 / 汤米 84 / 小忍 229
小润 73 / 小发 71 / 有纪 77
由美 124 / 小花 103 / 小瞳 59
小静 111 / 芽衣 3 / 理沙子 95
桃子 90 / 加奈 90 / 麻衣 70
小枫 38 / 麻美 89 / 小梢 88
哦！
这可以作为上课的题材喔！
这边的"琉衣琉衣"指的是你吗？
对呀！
琉衣琉衣得了86分呢！
乍看之下，琉衣的得分似乎跟全队平均分差不多呢！
哼——
所谓的平均分就是各队中每个人的大概得分，你懂吗？
我知道呀！
就是全队总分平均后的分数，对吧？
如果琉衣得到比平均数更高的分数的话，要请我吃蛋糕呦！♡
豪爽
那么就赶快来计算平均数吧！



A队

$$\frac{86+73+124+111+90+38}{6}=\frac{522}{6}=87$$

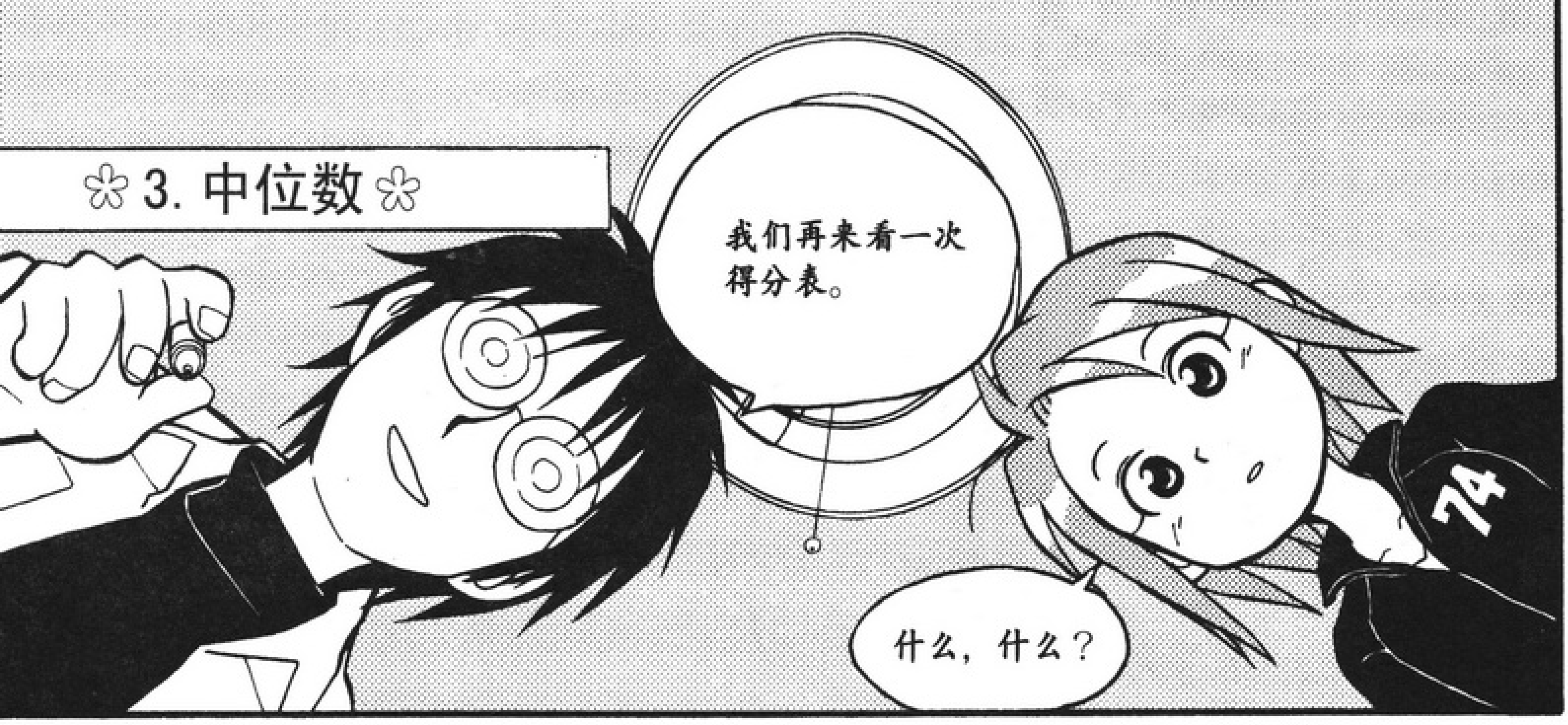
B队

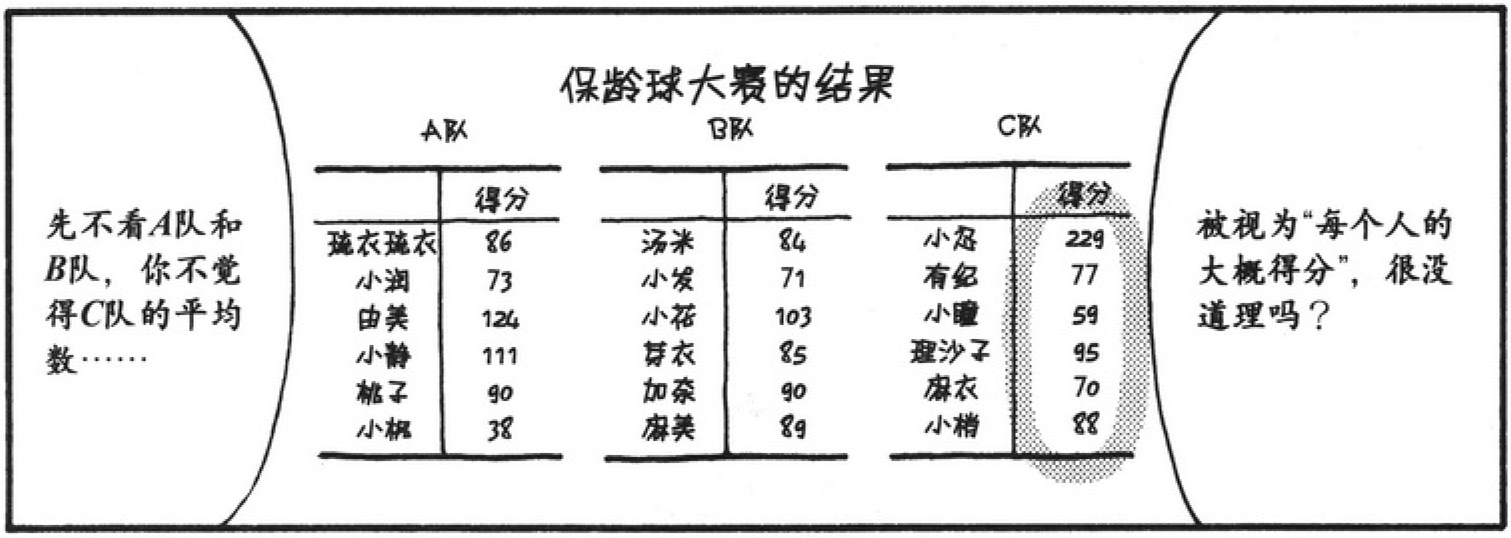
$$\frac{84+71+103+85+90+89}{6}=\frac{522}{6}=87$$

C队

$$\frac{229+77+59+95+70+88}{6}=\frac{618}{6}=103$$

1. 算术平均数：Arithmetic Mean。 2. 几何平均数：Geometric Mean。 3. 调和平均数：Harmonic Mean。





保龄球大赛的结果

A队

	得分
瑞衣瑞衣	86
小润	73
由美	124
小静	111
桃子	90
小枫	38

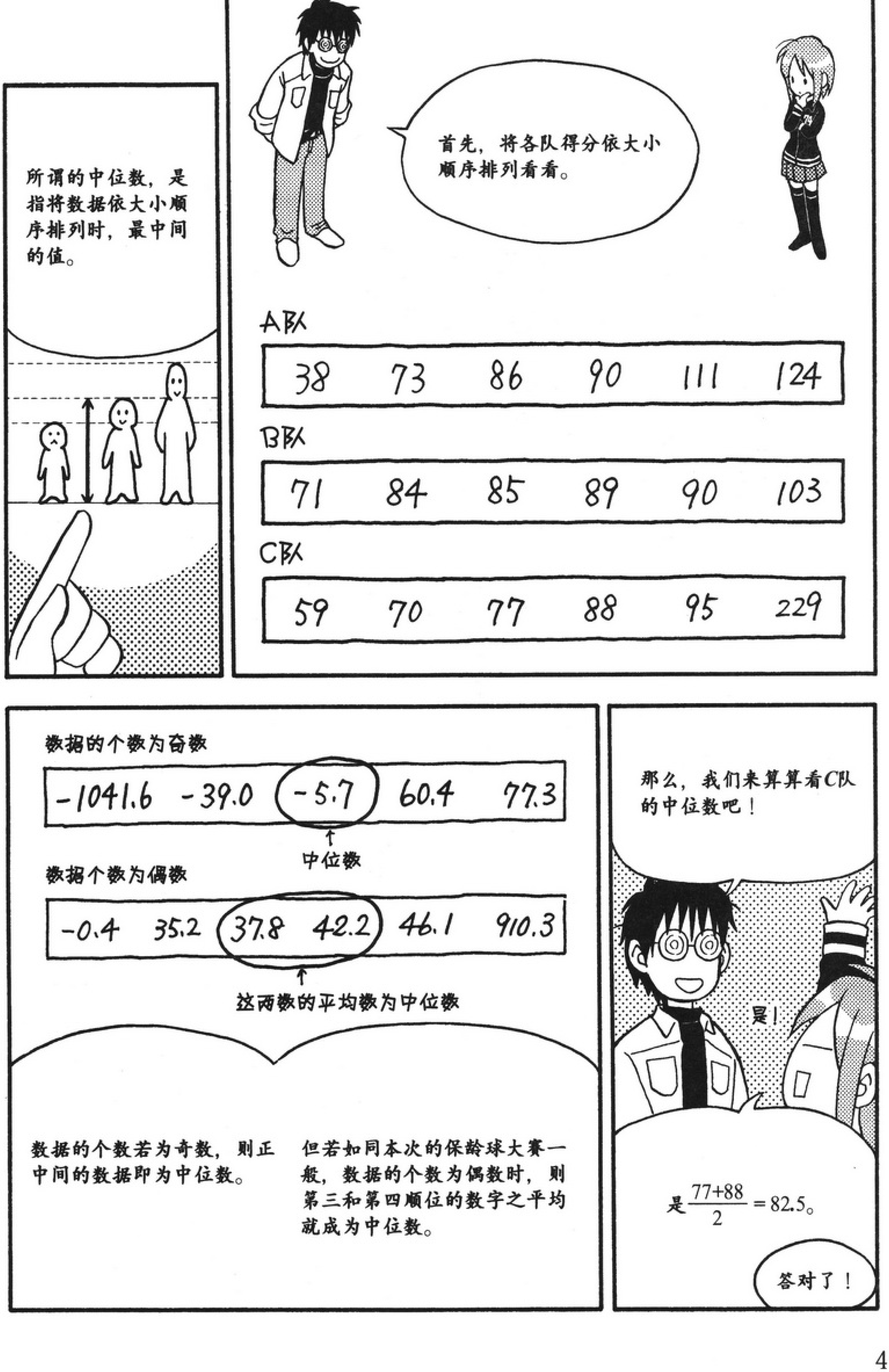
B队

	得分
汤米	84
小发	71
小花	103
芽衣	85
加奈	90
麻美	89

C队

	得分
小忍	229
有纪	77
小瞳	59
理沙子	95
麻衣	70
小梢	88

1．中位数：Median。

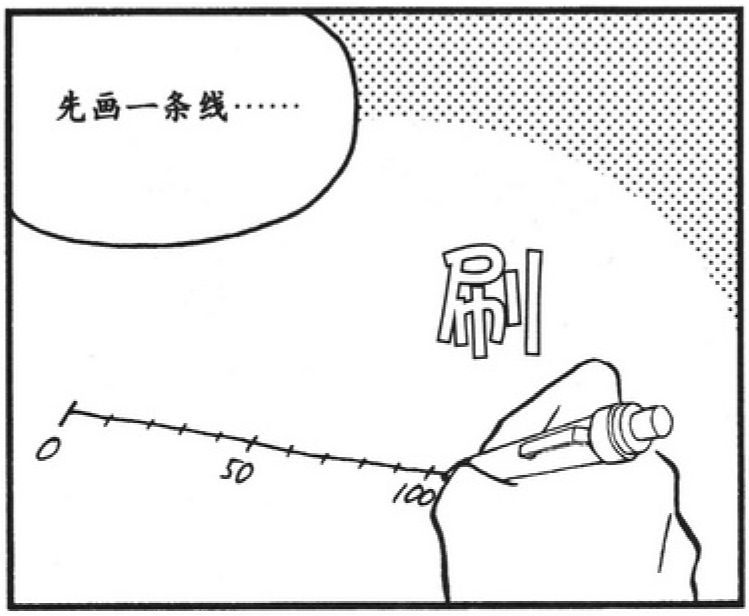

所谓的中位数，是指将数据依大小顺序排列时，最中间的值。
首先，将各队得分依大小顺序排列看看。
A队
38 73 86 90 111 124
B队
71 84 85 89 90 103
C队
59 70 77 88 95 229
数据的个数为奇数
-1041.6 -39.0 -5.7 60.4 77.3
中位数
数据个数为偶数
-0.4 35.2 37.8 42.2 46.1 910.3
这两数的平均数为中位数
数据的个数若为奇数，则正中间的数据即为中位数。
但若如同本次的保龄球大赛一般，数据的个数为偶数时，则第三和第四顺位的数字之平均就成为中位数。
那么，我们来算算看C队的中位数吧！
是！
是$\frac{77+88}{2}=82.5$。
答对了！

再来介绍一个和平均数有关的小知识……
又是小知识啊……
大智慧是由小知识聚集而成的，琉衣，你有存钱的习惯吗？
呵呵呵
当然有啰！虽然只有4位数。
那么，经常在报纸或电视新闻中出现的“日本‘平均’储蓄额”的数值，你没有对此数值之高感到惊讶吗？
OXBANK
OXBANK
证券
证券
银行
银行
当然有呀！原来除了我之外，其他人居然这么有钱呀……

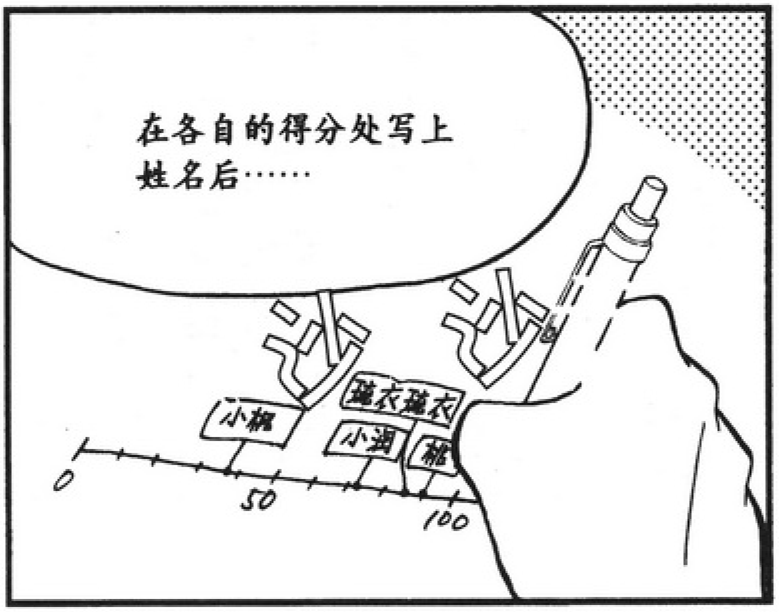

那个数字是被少数的超级大富翁抬高的。
因此，即使自己的储蓄额比“平均”值低相当多，也不必因此感到担忧。
在这种情况下，也许求中位数较能符合一般民众的平均储蓄额。
完全没在听啊…
做白日梦——
好吧！那就和比中位数高出许多的富翁结婚吧！
晕倒

✿4. 标准差✿

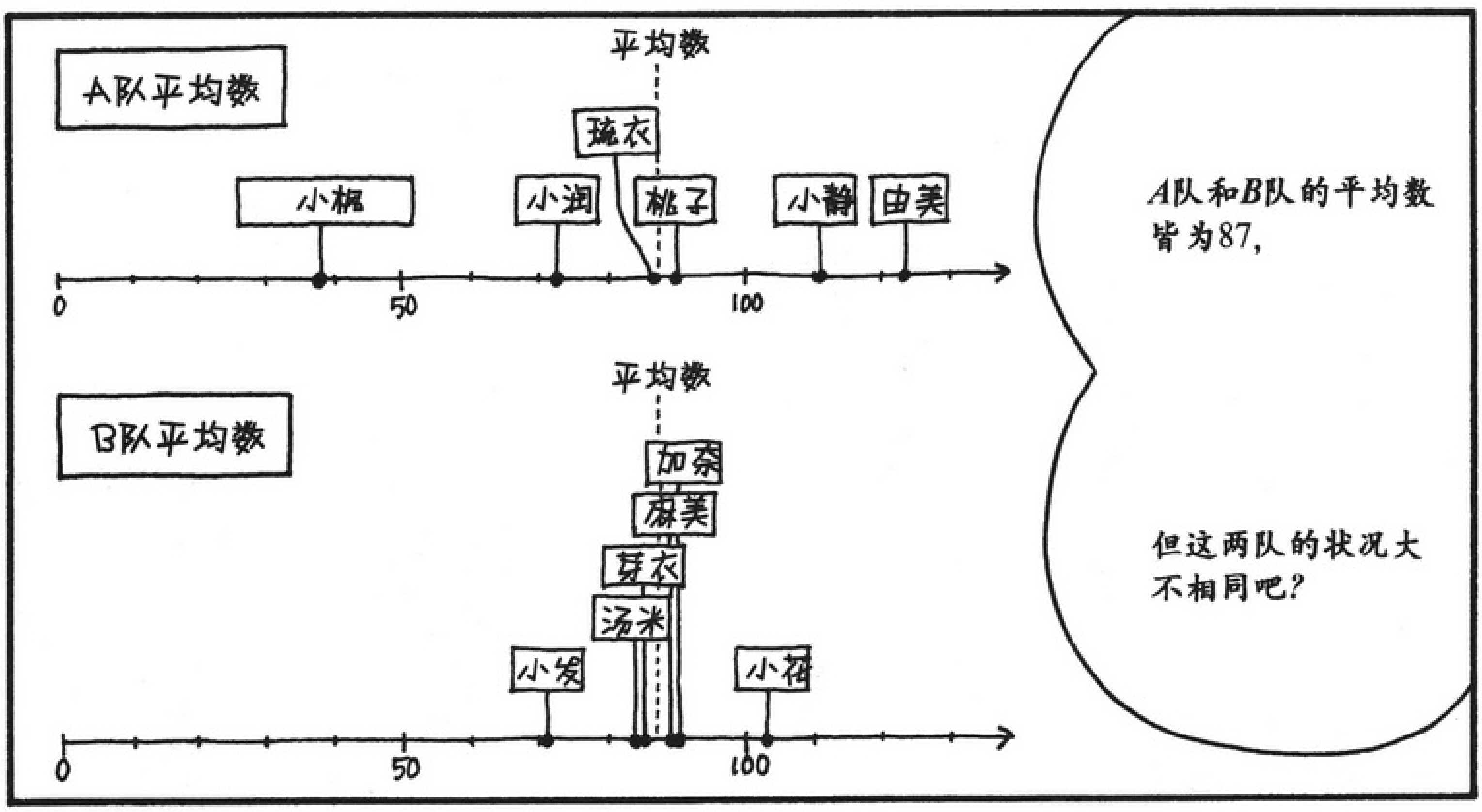
那么，我们来仔细看看，
A队和B队的得分吧！
好啊！



先画一条线……
刷
0
50
100



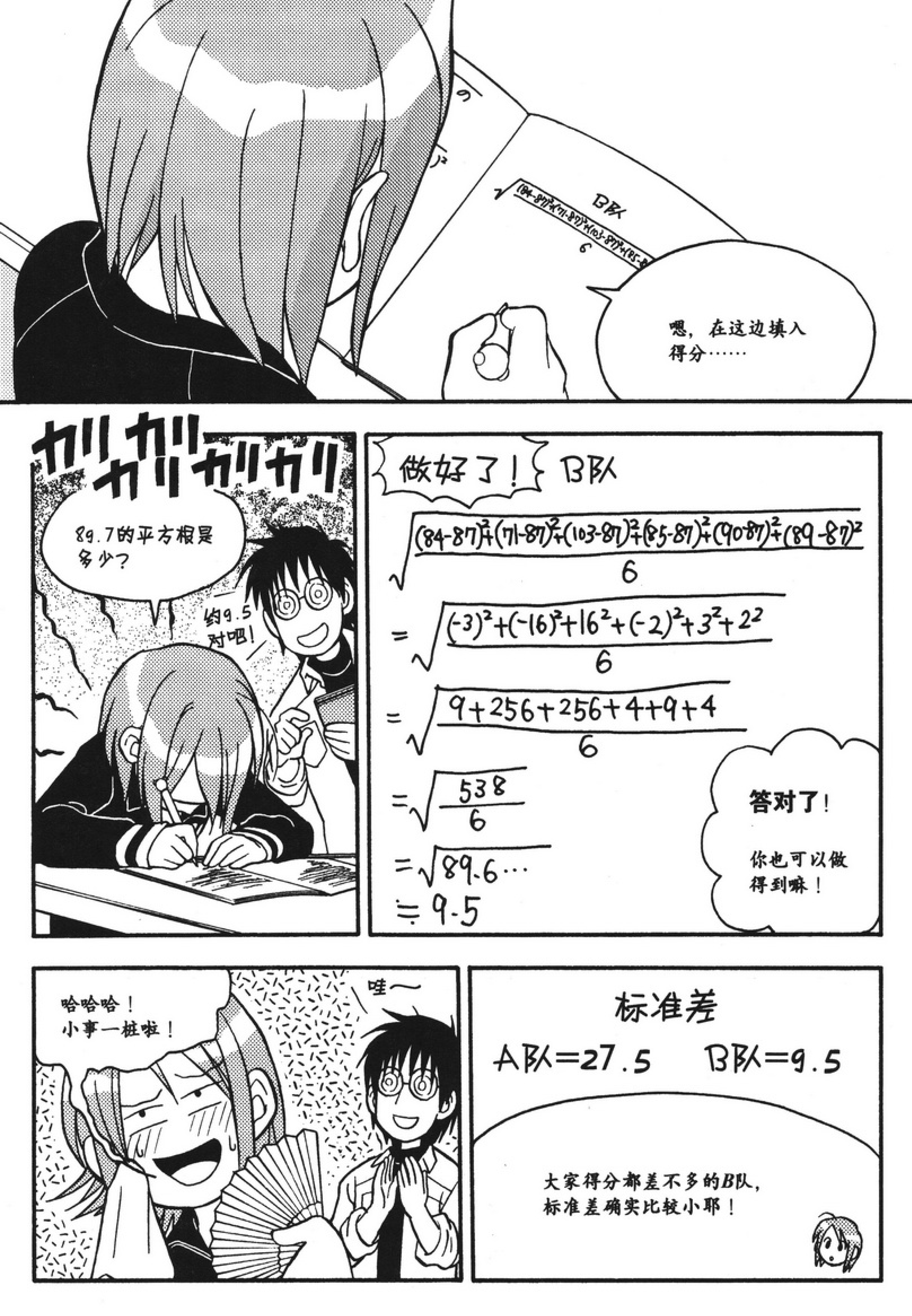
在各自的得分处写上姓名后……
沙
沙
小枫
琉衣琉衣
小润
0
50
100



A队平均数
平均数
琉衣
小枫
小润
桃子
小静
由美
0
50
100
B队平均数
平均数
加奈
麻美
芽衣
汤米
小发
小花
0
50
100
A队和B队的平均数皆为87，
但这两队的状况大不相同吧？

1．标准差：Standard Deviation。


答对了！
具体的计算方法就是这样。
咦？
怎么又变成数学啦？
$\sqrt{\dfrac{(\text{每一数据}-\text{平均数})^2\text{的总和}}{\text{数据的个数}}}$
只要在这里填入具体的数字就可以了！
来，我们一起算算看吧！
好，好吧……
首先是A队。
A队
$\sqrt{\dfrac{(86-87)^2+(73-87)^2+(124-87)^2+(111-87)^2+(90-87)^2+(38-87)^2}{6}}$
$=\sqrt{\dfrac{(-1)^2+(-14)^2+37^2+24^2+3^2+(-49)^2}{6}}$
$=\sqrt{\dfrac{1+196+1369+576+9+2401}{6}}$
$=\sqrt{\dfrac{4552}{6}}$
$=\sqrt{758.6\cdots}$
≈ 27.5
动手算过之后，我好像懂了呢！
那么B队就由你来试试看吧！


B队
嗯，在这边填入得分……
カリカリカリカリカリカリ
89.7的平方根是多少？
约9.5对吧！
做好了！
B队
$\sqrt{\frac{(84-87)^2+(71-87)^2+(103-87)^2+(85-87)^2+(90-87)^2+(89-87)^2}{6}}$
$=\sqrt{\frac{(-3)^2+(-16)^2+16^2+(-2)^2+3^2+2^2}{6}}$
$=\sqrt{\frac{9+256+256+4+9+4}{6}}$
$=\sqrt{\frac{538}{6}}$
$=\sqrt{89.6\cdots}$
$\fallingdotseq 9.5$
答对了！
你也可以做得到嘛！
哈哈哈！小事一桩啦！
哇～
标准差
A队=27.5 B队=9.5
大家得分都差不多的B队，标准差确实比较小耶！

标准差的算式是，
$$\sqrt{\frac{(每一数据-平均数)^2的总和}{数据的个数}}$$
但也有人认为应当是
$$\sqrt{\frac{(每一数据-平均数)^2的总和}{数据的个数-1}}$$
哦！
把数据的个数减掉一个！
是呀！
总体
样本
简单来说，
求总体的标准差就用前面的公式。
求样本的标准差就用后面的公式。
总体是真正想调查的对象的集合，
而样本是从总体中被选出来的人所形成的集合。


没错。
如果可以像琉衣的保龄球队一样，获得整个集合的数据就好了。
OK
但一般而言，这是很困难的，
所以大部分都是使用后面的公式。
嗯……
是这样子呀——
那么，今天的课程就到此结束。
好的！
谢谢老师！

✿5. 次数分布表的组距✿

至此，也许有些人仍然无法完全理解“1.次数分布表和直方图”，我们就再做一些详细的说明吧。

下表同第38页的曾使用过的表。

◆表2.1“美味拉面畅销前50”的次数分布表

组 以上 未满	组中值	次数	相对次数
500~600	550	4	0.08
600~700	650	13	0.26
700~800	750	18	0.36
800~900	850	12	0.24
900~1000	950	3	0.06
合计		50	1.00

如各位所见，上表中的组距是100。之所以先择100，并没有什么数学上的规定，而是全由山本老师主观决定的。没错，组距该设多少，完全依照分析者本身的判断。

“以主观设定的组距而做成的次数分布表并没有说服力，无法在他人面前公开，难道就没有按数学原理制定组距的方法吗？”也许有人会产生这样的疑问。事实上，方法是有的。步骤如下页所述。既然已经有数据了，就让我们来看看，如果以表2.1来试算会产生什么样的结果。

步骤 1

“组”的个数即组数可以使用史特吉斯公式进行计算，即：

$$1+\frac{\log_{10}\text{数据的个数}}{\log_{10}{}^{2}}$$

求出。

$$1+\frac{\log_{10}50}{\log_{10}2}=1+5.6438\cdots=6.6438\cdots\approx 7$$

步骤 2

组距以

$$\frac{(\text{数据的最大值})-(\text{数据的最小值})}{\text{用史特吉斯公式求出的组数}}$$

求出。

$$\frac{980-500}{7}=\frac{480}{7}=68.5714\cdots\approx 69$$

以步骤2求出的组距为基础，做出如下的次数分布表。

◆表2.2 “美味拉面畅销前50”的次数分布表（“组距”以公式求出）

组 以上 未满	组中值	次数	相对次数
500~569	534.5	2	0.04
569~638	603.5	5	0.10
638~707	672.5	15	0.30
707~776	741.5	6	0.12
776~845	810.5	10	0.20
845~914	879.5	10	0.20
914~983	948.5	2	0.14
合 计		50	1.00

结果如何？各位不觉得这样反而做出了一张比表2.1还令人无法理解的表格吗？也就是说，难道各位不会抱着“为何以69元为组距呢？”的疑问吗？然后，即使你努力地说明：“这是使用史特吉斯公式求出的……”你不觉得还是会被质问：“谁知道史特吉斯公式是什么呀！到底为什么要采用这么难以解释的组距呢？”

总而言之，也许有人会质疑以主观设定组距的合理性。但另一方面，我们从上表可以清楚得知，即使用数学方法设定组距，却时常还是会产生不尽理想的结果。因此，这个方法是否恰当，须重新思考。但是，我个人觉得原先的次数分布表就是用来掌握数据整体的“气氛”，因此，以分析者可接受的组距来处理即可。

✿6. 推断统计学和描述统计学✿

在序章中，有这样一段解说："所谓的统计学，即为从样本的信息推测总体状况的学问。"其实这段解说并不恰当。

统计学可分为推断统计学和描述统计学两类。序章所解说的为前者。那么，后者的描述统计学到底是什么呢？也就是借由整理资料，尽可能简单明了地显示出整体状况为目的的统计学。即，将对象集合视为一个总体的统计学。

描述统计学的解说可能由于过于抽象而让人难以理解。让我再举个例子说明。刚才山本求出了琉衣队得分的平均数和标准差。他求出此两者的目的，并非为了推测总体的状况。以琉衣队为样本的总体，究竟是怎样的总体呢？简而言之，山本之所以求出平均数和标准差，仅仅是为了简洁地表示琉衣队的状况。这样的统计学即为描述统计学。

例题和解答

例题

下表为高中女子100米短跑的成绩表。

	100米短跑（秒）
A同学	16.3
B同学	22.4
C同学	18.5
D同学	18.7
E同学	20.1

（1）请求出平均数。
（2）请求出中位数。
（3）请求出标准差。

解答

（1）平均数是 $\dfrac{16.3+22.4+18.5+18.7+20.1}{5}=\dfrac{96}{5}=19.2$

（2）中位数是18.7

16.3	18.5	18.7	20.1	22.4

（3）标准差是

$$\sqrt{\frac{(16.3-19.2)^2+(22.4-19.2)^2+(18.5-19.2)^2+(18.7-19.2)^2+(20.1-19.2)^2}{5}}$$

$$=\sqrt{\frac{(-2.9)^2+3.2^2+(-0.7)^2+(-0.5)^2+0.9^2}{5}}$$

$$=\sqrt{\frac{8.41+10.24+0.49+0.25+0.81}{5}}$$

$$=\sqrt{\frac{20.2}{5}}$$

$$=\sqrt{4.04}$$

$$\approx 2.01$$

总整理

- 利用“直觉”掌握整体数据的“氛围”的方法有：次数分布表及直方图。
- 设定次数分布表的组距可采用史特吉斯公式。
- 用数学原理掌握全体资料“氛围”的方法有，算术平均数、中位数和标准差。
- 当存在过大或过小的数据时，中位数较平均数更能正确掌握数据状态。
- 标准差为表示数据“离散程度”的指标。

◆第 3 章◆

掌握数据整体的状态

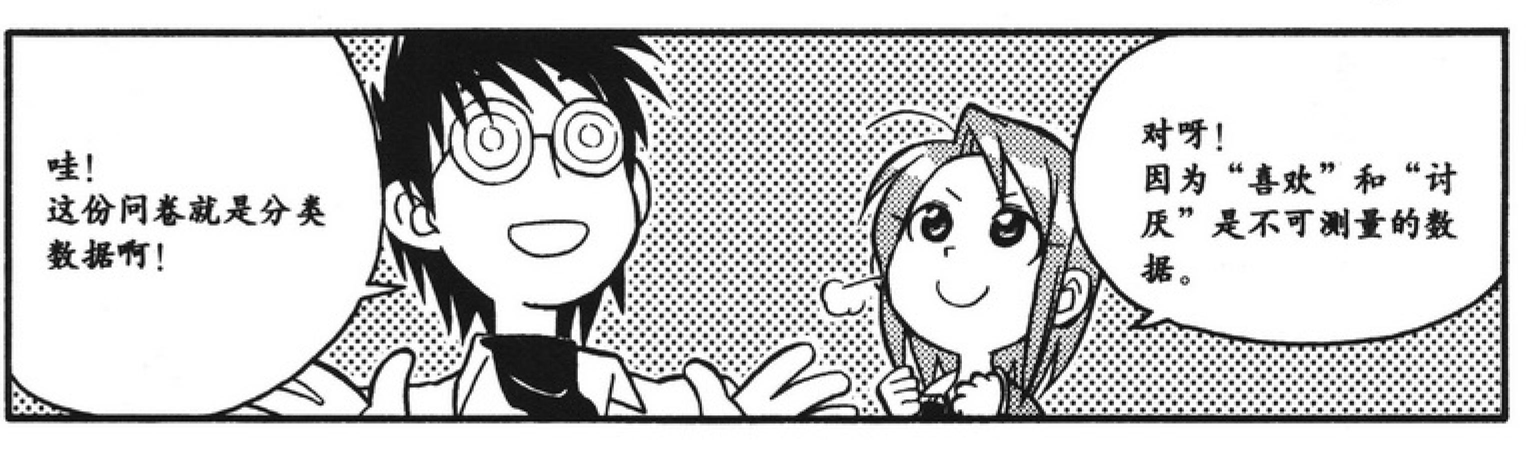
（分类数据篇）

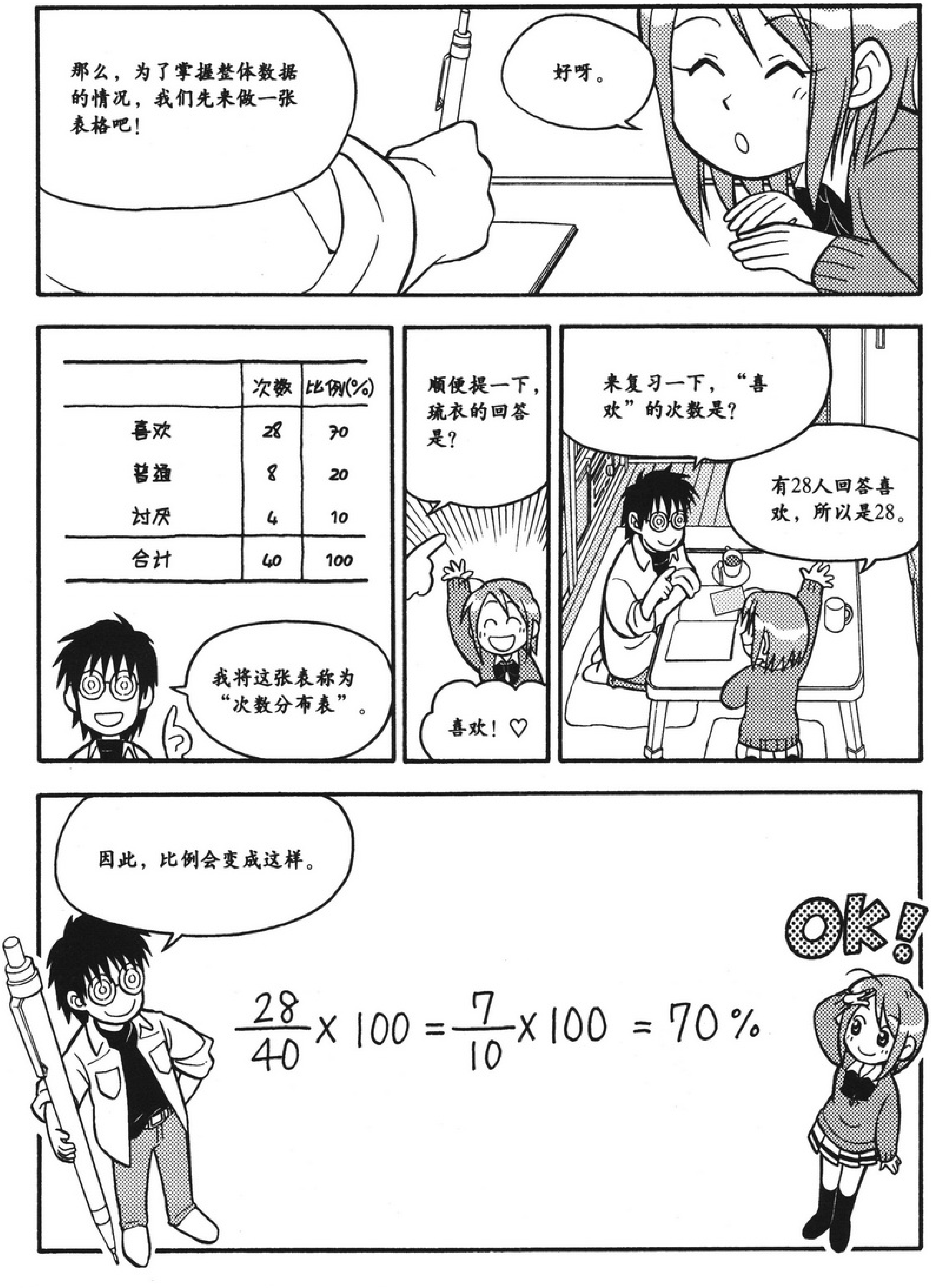
✿1.次数分布表✿

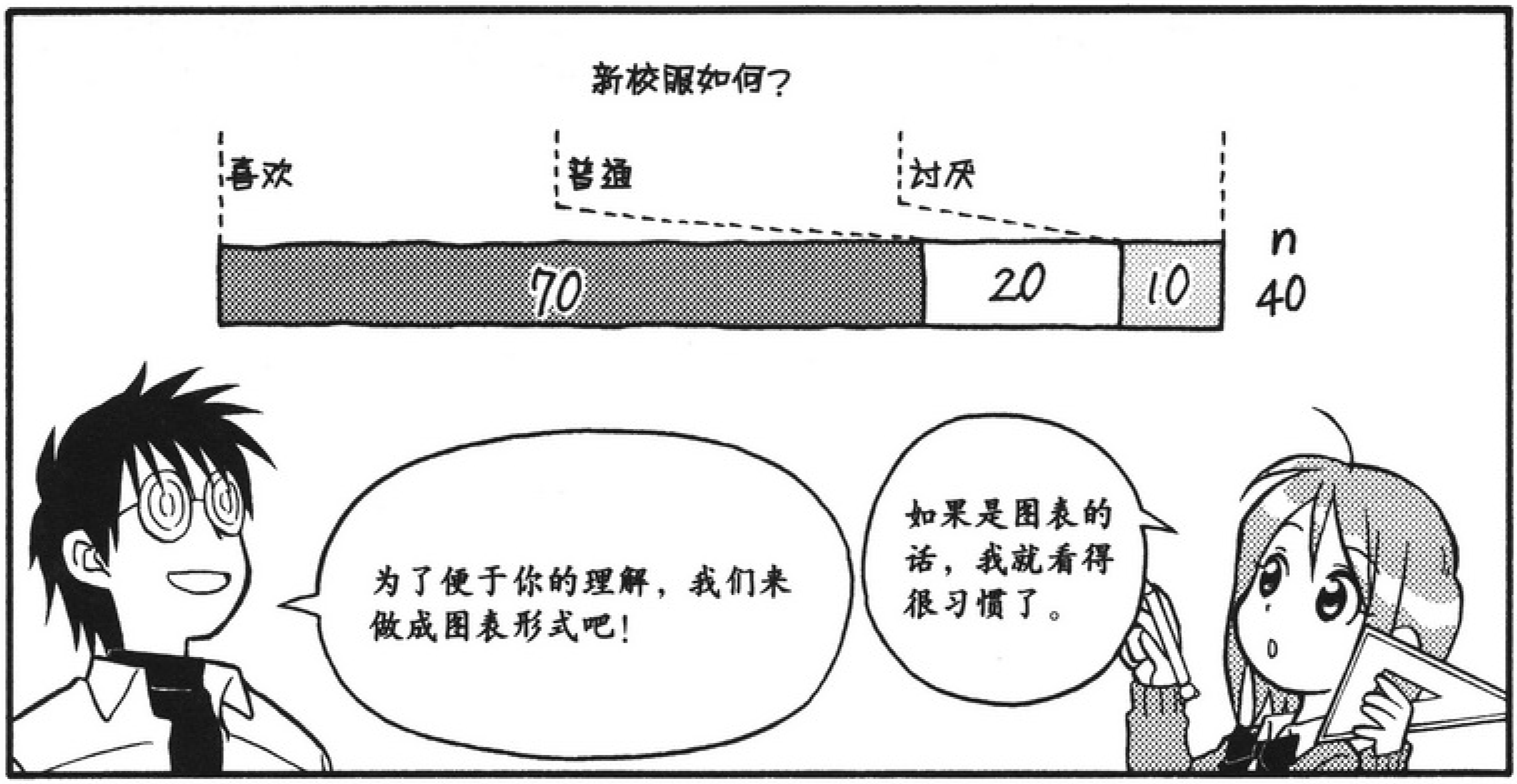
你还记得分类数据是“不可测量”的数据吗？
嗯！似乎还记得。
カチャ
什么？
今天穿校服耶！
啊！这个？
再过不久，
我就不能穿这件校服了……
什么？要毕业了吗？明明才二年级
我们学校要换新校服了！

新校服问卷调查

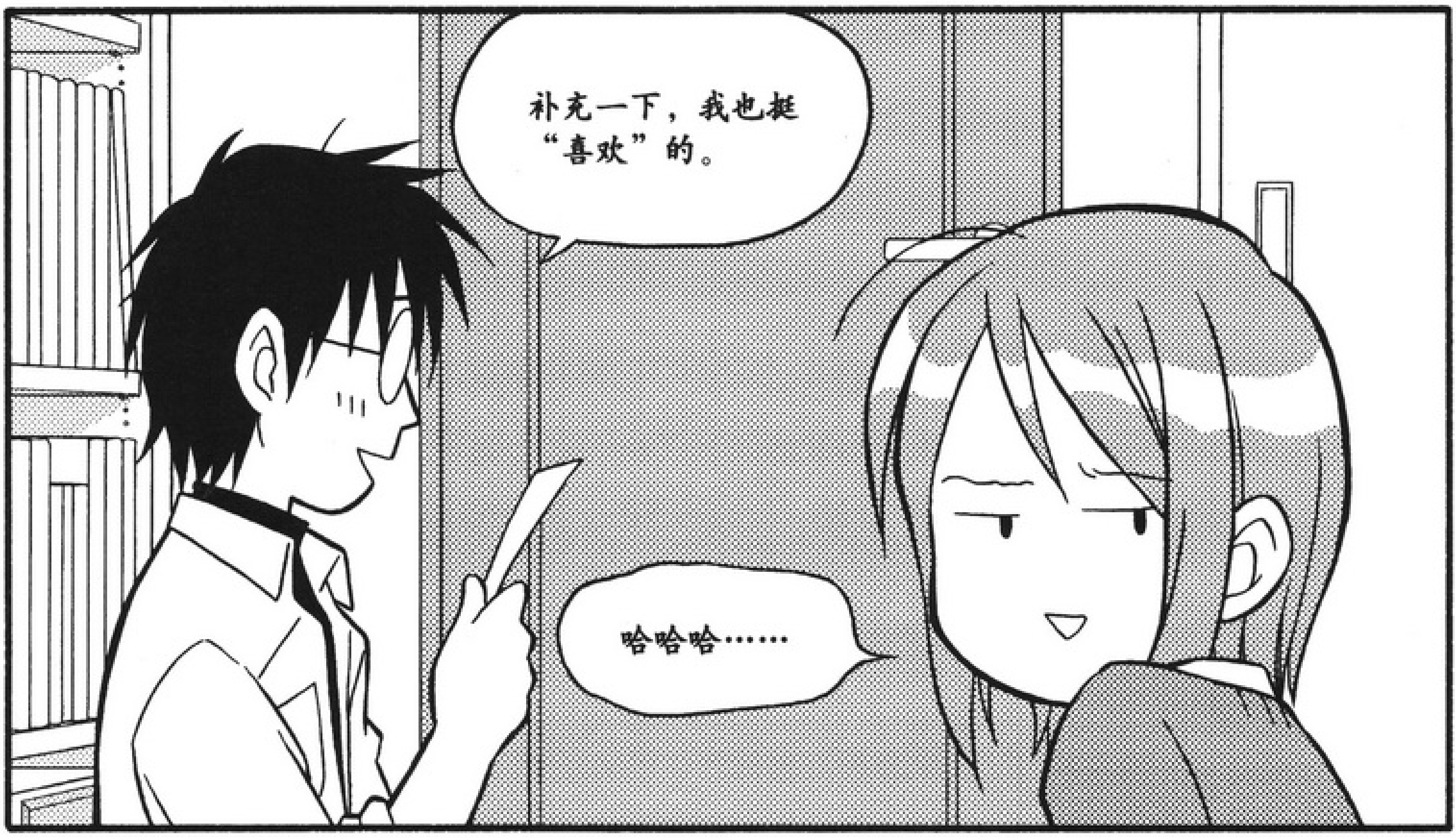
	新校服		新校服		新校服
1	喜欢	16	普通	31	普通
2	普通	17	喜欢	32	普通
3	喜欢	18	喜欢	33	喜欢
4	普通	19	喜欢	34	讨厌
5	讨厌	20	喜欢	35	喜欢
6	喜欢	21	喜欢	36	喜欢
7	喜欢	22	喜欢	37	喜欢
8	喜欢	23	讨厌	38	喜欢
9	喜欢	24	普通	39	普通
10	喜欢	25	喜欢	40	喜欢
11	喜欢	26	喜欢		
12	喜欢	27	讨厌		
13	普通	28	喜欢		
14	喜欢	29	喜欢		
15	喜欢	30	喜欢		




那么，为了掌握整体数据的情况，我们先来做一张表格吧！
好呀。
次数 比例(%)
喜欢 28 70
普通 8 20
讨厌 4 10
合计 40 100
我将这张表称为“次数分布表”。
顺便提一下，琉衣的回答是？
喜欢！♡
来复习一下，“喜欢”的次数是？
有28人回答喜欢，所以是28。
因此，比例会变成这样。
$\frac{28}{40}\times 100=\frac{7}{10}\times 100 = 70\%$
OK!


新校服如何？
喜欢
普通
讨厌
70
20
10
n
40
为了便于你的理解，我们来做成图表形式吧！
如果是图表的话，我就看得很习惯了。


列出图表后，回答“喜欢”的人超过半数，因此这款校服的设计似乎还蛮讨人喜欢的。
当然！因为真的很可爱嘛！


补充一下，我也挺“喜欢”的。
哈哈哈……

例题和解答

例题

某家报社对有意执掌下届政权的△△党，做了份问卷调查表。结果如下表所示。

	相较于○○党，△△党……
回答者1	不值得期待
回答者2	不值得期待
回答者3	不值得期待
回答者4	没意见
回答者5	值得期待
回答者6	不值得期待
回答者7	值得期待
回答者8	没意见
回答者9	不值得期待
回答者10	不值得期待

请将此问卷调查表结果做成“次数分布表”。

解答

“次数分布表”如下所示。

	次数	比例(%)
值得期待	2	20
没意思	2	20
不值得期待	6	60
合计	10	100

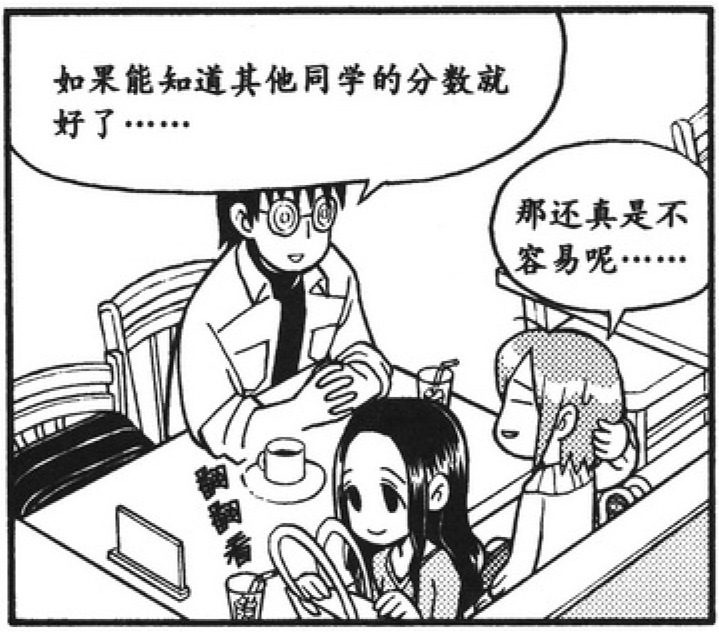
总整理

• 做成“次数分布表”为掌握数据整体状态的方法之一。

第 4 章
标准计分和离差

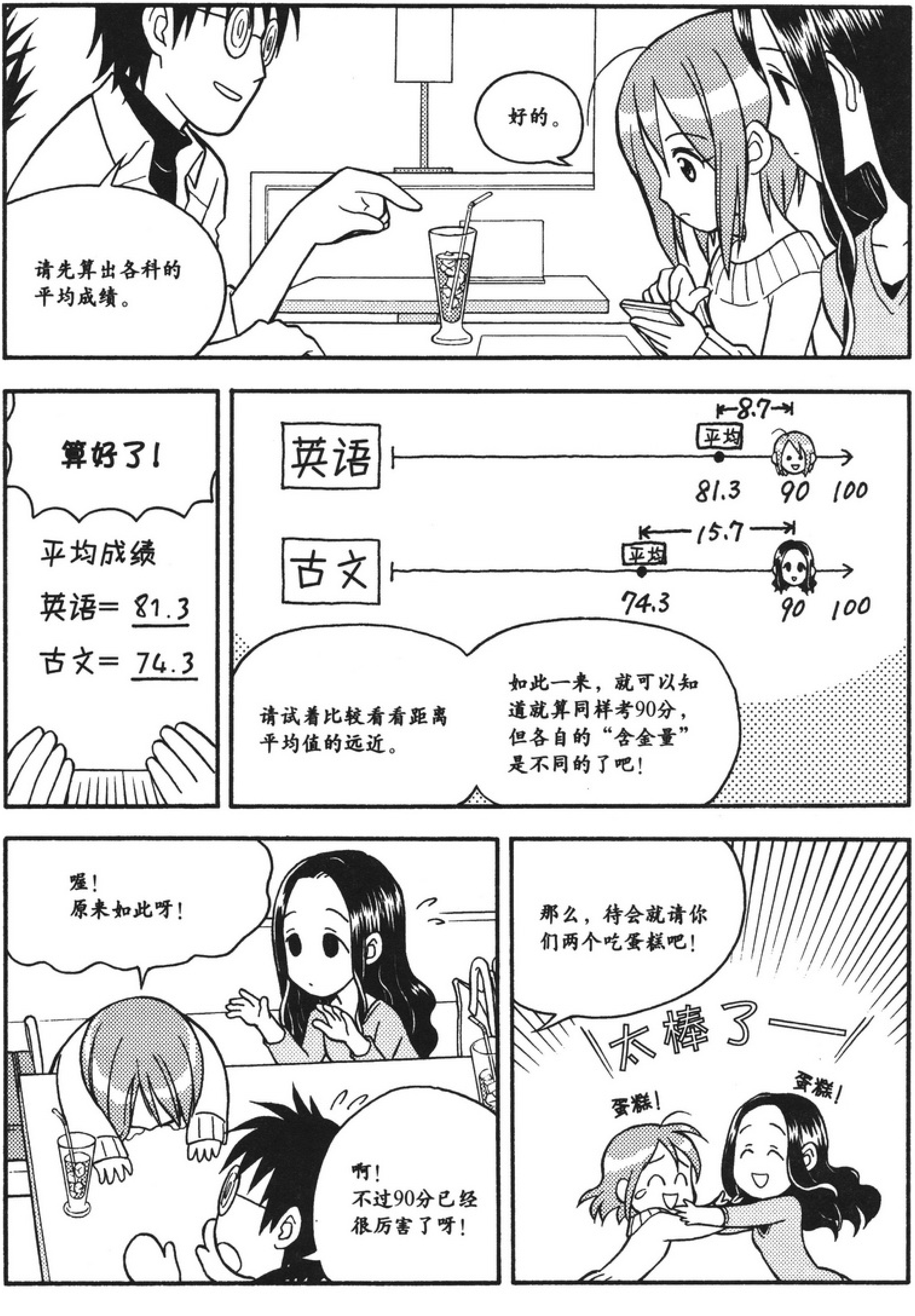
1. 标准化和标准计分

1．离差：Deviation Score。



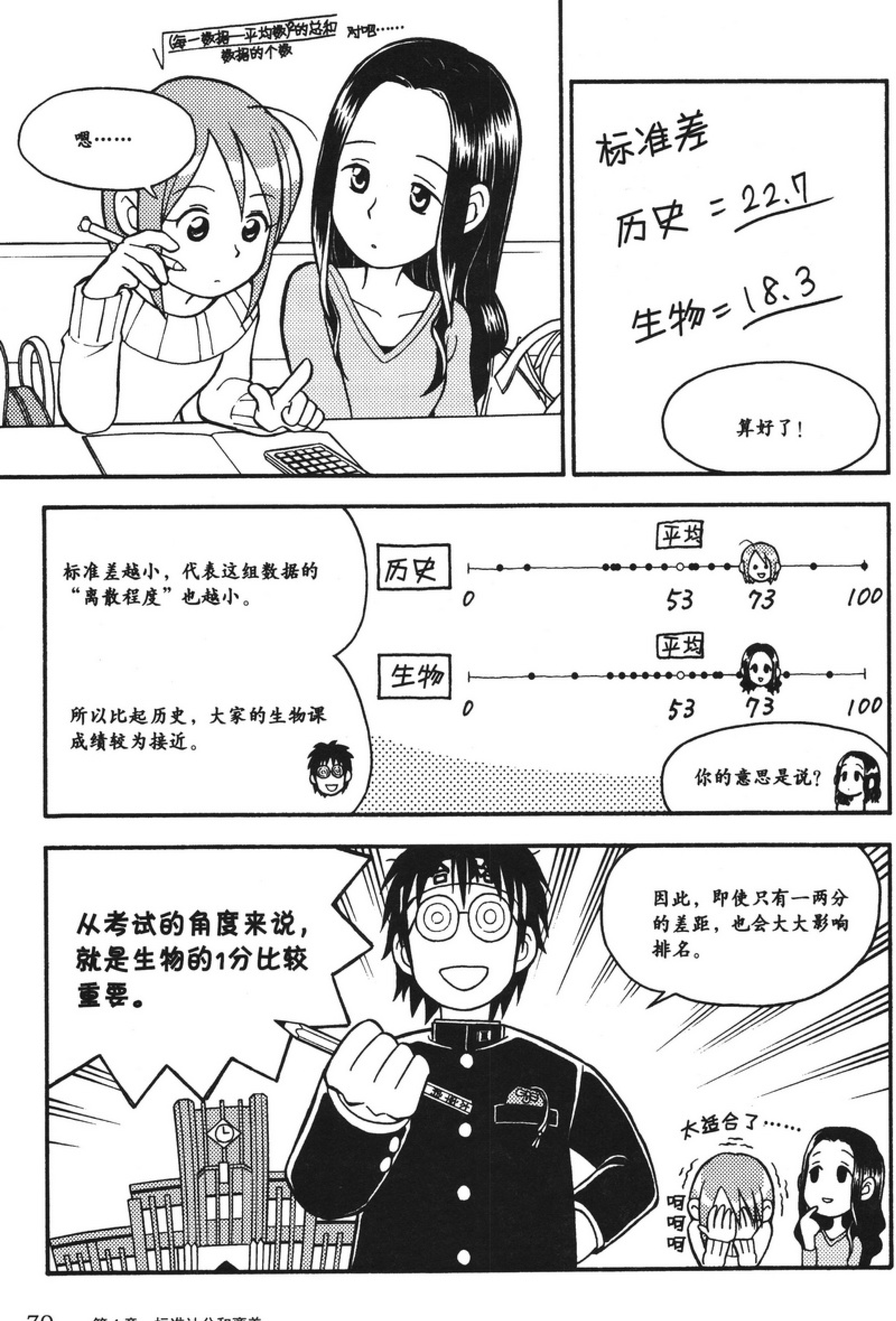
考试成绩（100分满分）

	英语	古文		英语	古文
琉衣	90	71	H	67	85
由美	81	90	I	87	93
A	73	79	J	78	89
B	97	70	K	85	78
C	85	67	L	96	74
D	60	66	M	77	65
E	74	60	N	100	78
F	64	83	O	92	53
G	72	57	P	86	80


请先算出各科的平均成绩。
好的。
算好了!
平均成绩
英语= 81.3
古文= 74.3
英语
8.7
平均
81.3
90
100
古文
15.7
平均
74.3
90
100
请试着比较看看距离平均值的远近。
如此一来，就可以知道就算同样考90分，但各自的“含金量”是不同的了吧!
喔!
原来如此呀!
啊!
不过90分已经很厉害了呀!
那么，待会就请你们两个吃蛋糕吧!
太棒了——
蛋糕!
蛋糕!

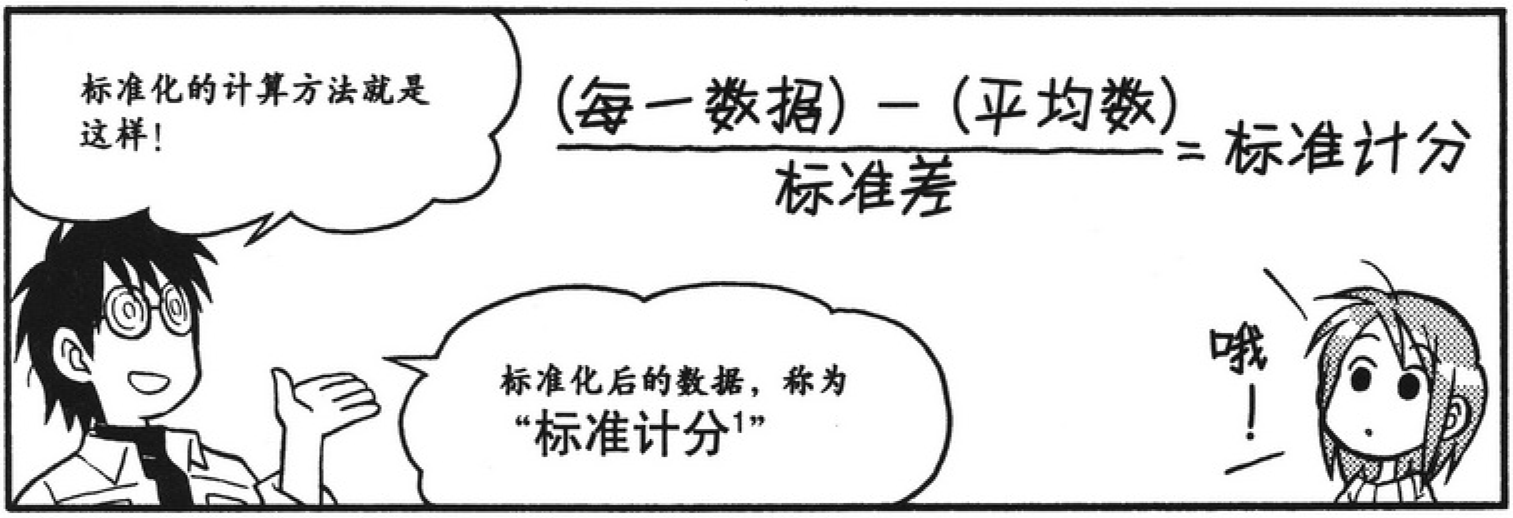
	历史	生物
琉衣	73	59
由美	61	73
A	14	47
B	41	38
C	49	63
D	87	56
E	69	15
F	65	53
G	36	80
H	7	50
I	53	41
J	100	62
K	57	44
L	45	26
M	56	91
N	34	35
O	37	53
P	70	68
平均数	53	53

1．标准差：Standard deviation。



$\sqrt{\frac{(每一数据-平均数)^2的总和}{数据的个数}}$ 对吧……
嗯……
标准差
历史 = 22.7
生物 = 18.3
算好了！
标准差越小，代表这组数据的“离散程度”也越小。
平均
历史
0
53
73
100
平均
生物
0
53
73
100
所以比起历史，大家的生物课成绩较为接近。
你的意思是说？
从考试的角度来说，就是生物的1分比较重要。
因此，即使只有一两分的差距，也会大大影响排名。
太适合了……
呵呵呵

1．标准化：Standardization。



历史和生物的考试成绩及其标准计分

	历史	生物	历史的标准计分	生物的标准计分
琉衣	73	59	0.88	0.33
由美	61	73	0.35	1.09
A	14	47	−1.71	−0.33
B	41	38	−0.53	−0.82
C	49	63	−0.18	0.55
D	87	56	1.49	0.16
E	69	15	0.70	−2.08
F	65	53	0.53	0
G	36	80	−0.75	1.48
H	7	50	−2.02	−0.16
I	53	41	0	−0.66
J	100	62	2.07	0.49
K	57	44	0.18	−0.49
L	45	26	−0.35	−1.48
M	56	91	0.13	2.08
N	34	35	−0.84	−0.98
O	37	53	−0.70	0
P	70	68	0.75	0.82
平均	53	53	0	0
标准差	22.7	18.3	1	1

琉衣的历史标准计分 $\frac{73-53}{22.7}=\frac{20}{22.7}=0.88$

由美的生物标准计分 $\frac{73-53}{18.3}=\frac{20}{18.3}=1.09$

1．标准计分：Standard Score。


✿2. 标准计分的特征✿
那么，这些数字代表什么？
0.88和1.09
标准化后，求出标准计分具有某些特征。
①无论作为变量的满分为几分，其标准计分的平均数势必为0，而其标准差势必为1.
满分为100分的考试和满分为200分的考试也可以比较呦！
②无论作为变量的单位是什么，其标准计分的平均数势必为0，而其标准差势必为1.
安打率和全垒打数等即使单位不同也可以进行比较。
由于标准计分中，
0.88 < 1.09
（历史） （生物）
因此，哪一个73分较有价值，我想应该可以很明显地看出来了吧！
胜负立判呀！


✿3. 离　差✿
而且，离差就是应用标准计分所得的数值哦！
哦—
它的公式就像这样。
离差＝标准计分×10＋50
真的耶！含有标准计分在里面—
来算算看，你们考试分数的离差吧！
琉衣(历史)
0.88×10＋50＝8.8＋50＝58.8
由美(生物)
1.09×10＋50＝10.9＋50＝60.9
对对！就是这个结果—
特征请看！
标准计分
①无论作为变量的满分为几分，其标准计分的平均数势必为0，而其标准差势必为1。
②无论作为变量的单位是什么，其标准计分的平均数势必为0，而其标准差势必为1。
离　差
①无论作为变量的满分为几分，其离差的平均数势必为50，而其标准差势必为10。
②无论作为变量的单位是什么，其离差的平均数势必为50，而其标准差势必为10。


原来如此——
对于考试成绩来说，就算只是一分之差也影响深远，所以才会使用离差这个指标吧！
那么，今天的课程就到此为止吧！
太棒了——
蛋糕——
我呢，觉得这个和这个都不错呢！
"水果塔"
"草莓千层派"
60日元
80日元
え??
え??
我想尝尝这个……
蘑菇的童话风蛋糕
150日元
山本老师，我们开吃啦！♡
冒冷汗
冒冷汗
我带的钱够吗？

✿4. 关于离差的解释✿

在此，有必要加强离差的解说。

离差如同74页的解说，是以下述算式求得的：

$$离差=标准计分\times10+50=\frac{(每一数据)-(平均数)}{标准差}\times10+50$$

那么，琉衣的班上，如同61页中说明的，全班共有40人。琉衣班上的“女生”，如40页所示，共有18人。所以69页的离差实例，并非以全班同学为对象，而仅以女生为对象。若以全班同学为对象，平均数和标准差的值就会和仅以女生为对象时迥然不同，琉衣和由美的离差值也势必有差异产生。实际上，若以全班同学为对象的情况下，琉衣的离差值较高。全班的测验结果如表4.1所示。请各位务必试着算算离差。我先将答案说出来，琉衣的历史成绩离差值为59.1，而由美的生物成绩离差值为56.7。

另外，假设在2年1班及2年2班也举办了相同的测验。2年1班只求出自己班上的平均数和标准差，再以此为基础，求出离差值。2班也只求出自己班上的平均数和标准差，再以此为基础，求出离差。结果，1班的A同学和B同学实力相当。然而，由于求出A同学和B同学的离差值时，所采用的平均数和标准差并不一致，既然两班的平均数和标准差并不相同，那么两人的离差值并无可比性。

我再举个例子，A同学于4月间参加了某补习班的模拟测验，其考试成绩的离差值为54。而在暑期补习班中努力用功的A同学，为了想确认实力提升了多少，9月间又参加了另一个补习班所举办的模拟测验，其考试成绩的离差值为62。从两次离差值来看，乍看之下，A同学的实力似乎有所提升。然而，由于4月和9月之模拟测验分数的举办者不同，因此考生不相同。再加上，从4月与9月的考试结果，在欲求出离差之际，所使用的平均数与标准差一定不同，因此无法就两者得出的离差做比较。

各位觉得如何呢？关于离差的解释，相当有深度啊！

◆表4.1历史和生物的测验结果（琉衣的全班同学）

全体女同学		历史	生物
	琉衣	73	59
	由美	61	73
	A	14	47
	B	41	38
	C	49	63
	D	87	56
	E	69	15
	F	65	53
	G	36	80
	H	7	50
	I	53	41
	J	100	62
	K	57	44
	L	45	26
	M	56	91
	N	34	35
	O	37	53
	P	70	68

全体男同学		历史	生物
	a	54	2
	b	93	7
	c	91	98
	d	37	85
	e	44	100
	f	16	29
	g	12	57
	h	44	37
	i	4	95
	j	17	39
	k	66	70
	l	53	14
	m	14	97
	n	73	39
	o	6	75
	p	22	80
	q	69	77
	r	95	14
	s	16	24
	t	37	91
	u	14	36
	v	88	76
全班同学成绩的平均数		48.0	54.9
全班同学成绩的标准差		27.5	26.9

例题

下表为高中女子100米的短跑成绩。

	100米短跑（秒）
A同学	16.3
B同学	22.4
C同学	18.5
D同学	18.7
E同学	20.1
平均	19.2
标准差	2.01

（1）请确认“100米短跑成绩的标准计分”之平均数是否为0。

（2）请确认“100米短跑成绩的标准计分”之标准差是否为1。

解答

(1)"100米短跑的标准成绩"之平均数

$$=\frac{\left(\frac{16.3-19.2}{2.01}\right)+\left(\frac{22.4-19.2}{2.01}\right)+\left(\frac{18.5-19.2}{2.01}\right)+\left(\frac{18.7-19.2}{2.01}\right)+\left(\frac{20.1-19.2}{2.01}\right)}{5}$$

$$=\frac{\left\{\frac{(16.3-19.2)+(22.4-19.2)+(18.5-19.2)+(18.7-19.2)+(20.1-19.2)}{2.01}\right\}}{5}$$

整理分子

$$=\frac{\left\{\frac{16.3+22.4+18.5+18.7+20.1-19.2-19.2-19.2-19.2-19.2}{2.01}\right\}}{5}$$

将分子分为每笔数据和-19.2

$$=\frac{\left\{\frac{96-19.2\times 5}{2.01}\right\}}{5}$$

$$=\frac{\left\{\frac{96-96}{2.01}\right\}}{5}$$

$$=\frac{0}{5}$$

$$=0$$

(2)"100米短跑的标准计分"之标准差

$$=\sqrt{\frac{\left(\frac{16.3-19.2}{2.01}-0\right)^2+\left(\frac{22.4-19.2}{2.01}-0\right)^2+\left(\frac{18.5-19.2}{2.01}-0\right)^2+\left(\frac{18.7-19.2}{2.01}-0\right)^2+\left(\frac{20.1-19.2}{2.01}-0\right)^2}{5}}$$

$$=\sqrt{\frac{\left(\frac{16.3-19.2}{2.01}\right)^2+\left(\frac{22.4-19.2}{2.01}\right)^2+\left(\frac{18.5-19.2}{2.01}\right)^2+\left(\frac{18.7-19.2}{2.01}\right)^2+\left(\frac{20.1-19.2}{2.01}\right)^2}{5}}$$

$$=\sqrt{\frac{\left\{\frac{(16.3-19.2)^2+(22.4-19.2)^2+(18.5-19.2)^2+(18.7-19.2)^2+(20.1-19.2)^2}{2.01^2}\right\}}{5}}$$

整理分子

$$=\sqrt{\frac{1}{2.01^2}\times\frac{(16.3-19.2)^2+(22.4-19.2)^2+(18.5-19.2)^2+(18.7-19.2)^2+(20.1-19.2)^2}{5}}$$

整理分子

$$=\frac{1}{2.01}\times\sqrt{\frac{(16.3-19.2)^2+(22.4-19.2)^2+(18.5-19.2)^2+(18.7-19.2)^2+(20.1-19.2)^2}{5}}$$

$$=\frac{1}{\text{“100米短跑”的标准差}}\times\text{“100米短跑”的标准差}$$

详见78页的表

$$=1$$

总整理

- 标准化即为，以距离平均数的远近程度及数据的“离散程度”为基础，将数据的价值转换为易于探讨的数值。
- 若执行标准化，则可以比较
  - 满分不同的变量
  - 单位不同的变量
- 标准化后的数据称为标准计分。
- 求离差值必须应用到标准计分。

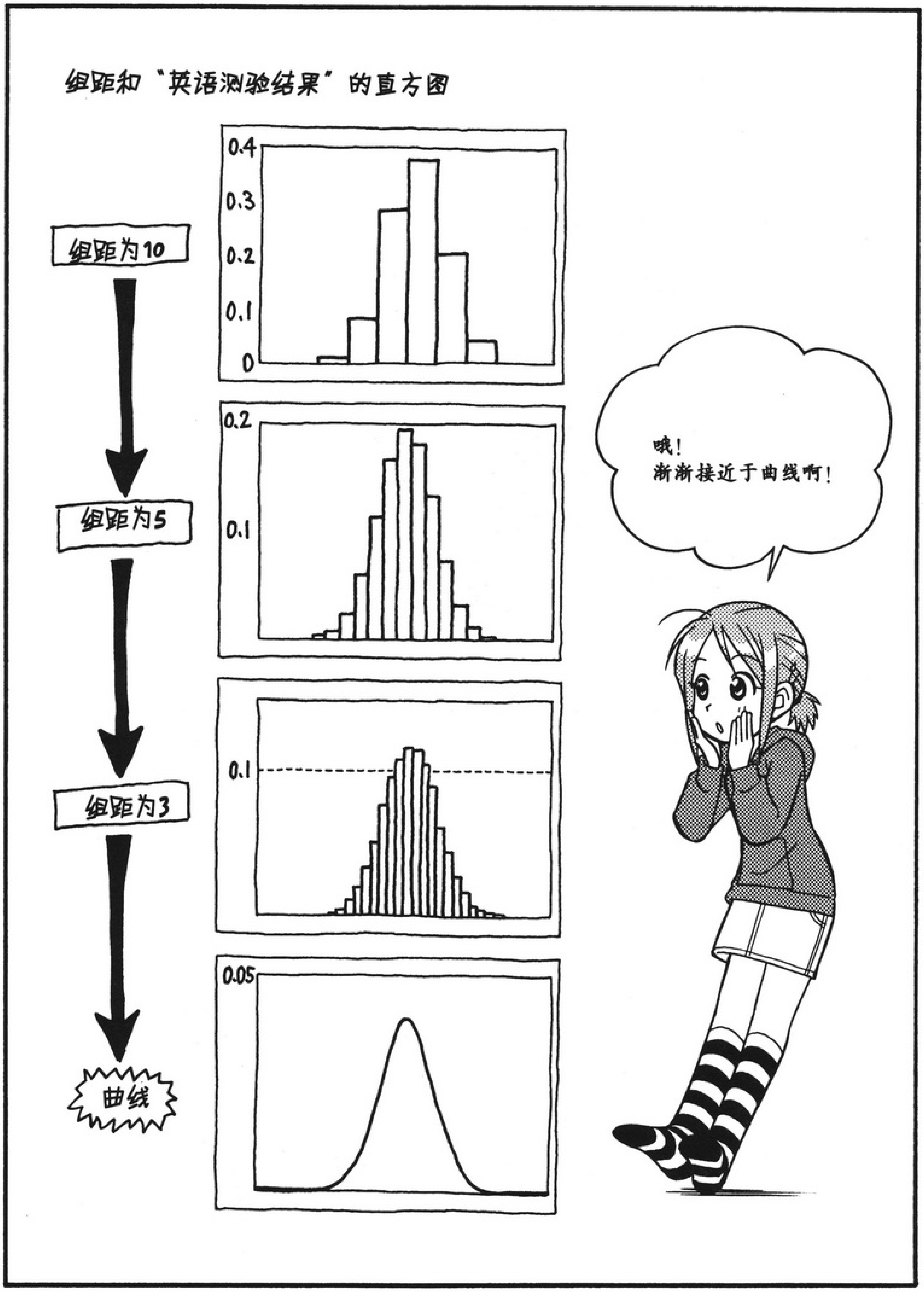
◆第5章◆

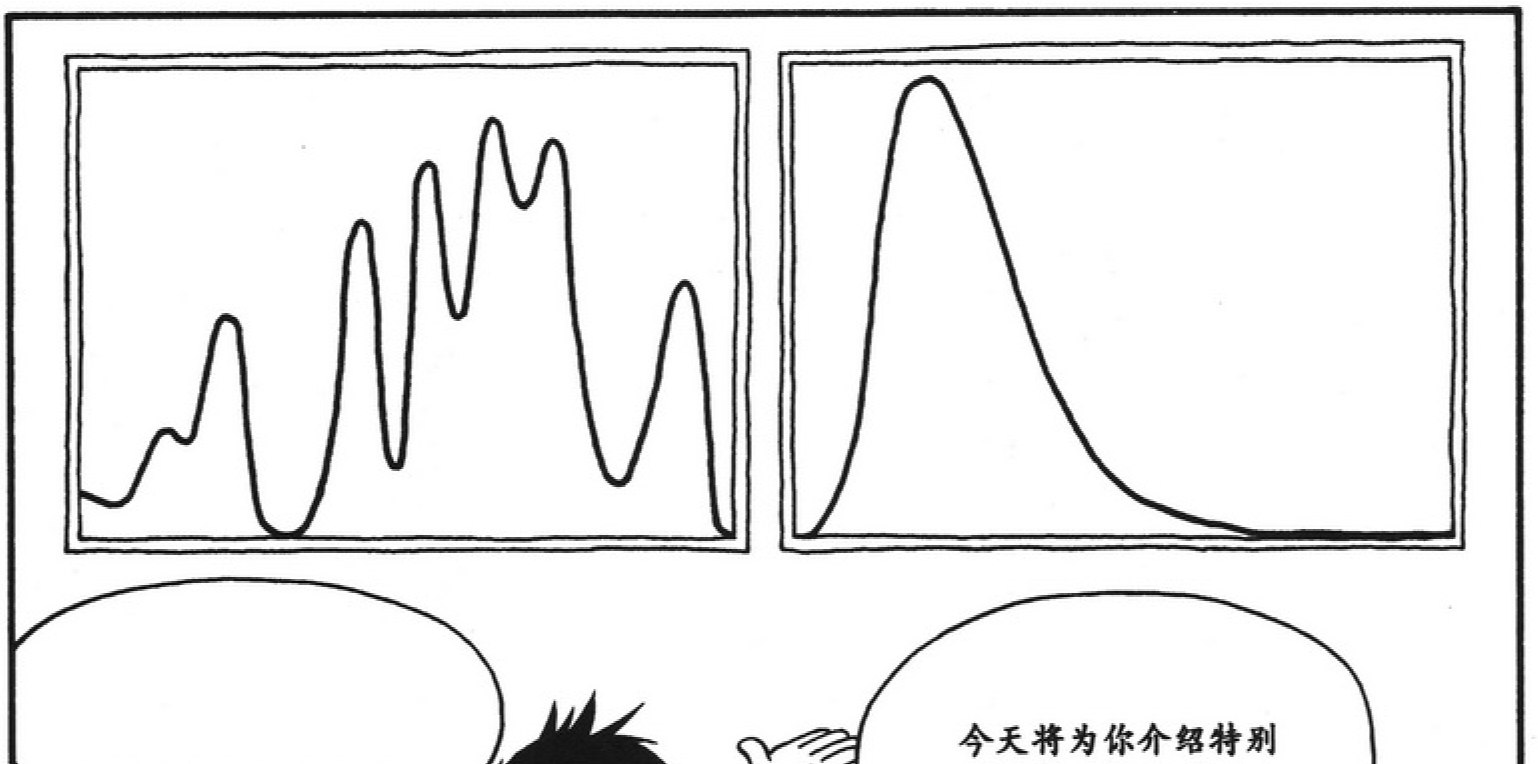
求 机 率

1. 机率密度函数

统计学中有时会提到“某某机率”小于0.05——
终于要开始进入机率的课程了。
今天就来谈谈求“某某机率”所须具备的知识吧！
じーっ
山本老师不是条件超好的吗？
哪里啊……我还是喜欢五十岚先生……
琉衣？
啊！
抱歉！
那机率是指会出现在天气预报中的那种机率吗？
是呀。
今天的上课内容会有点抽象。
抽象！？
琉衣
但是，从现在起所学的知识，在统计学中会常常出现，一定要认真听讲啊！
好，好吧……


那么，
假设A县的全体高三学生，
A县的全体高三学生英语测验结果
英语测验结果
学生1 42
学生2 91
⋮ ⋮
学生10421 50
平均数 53
标准差 10
参加了某补习班的英语测验，结果如下。
今天倒是做了万全准备嘛。
呵呵呵
ペラ
将刚才的表格做成直方图，就会变成这样。
英语测验结果的直方图（组距为10）
0.4
0.3
0.2
0.1
0
5 15 25 35 45 55 65 75 85 95
哦！
果然做成直方图后比较容易理解啊！
因为视觉效果比较好！
若将这份直方图的组距缩小，结果会变得如何呢？
严厉
咦？

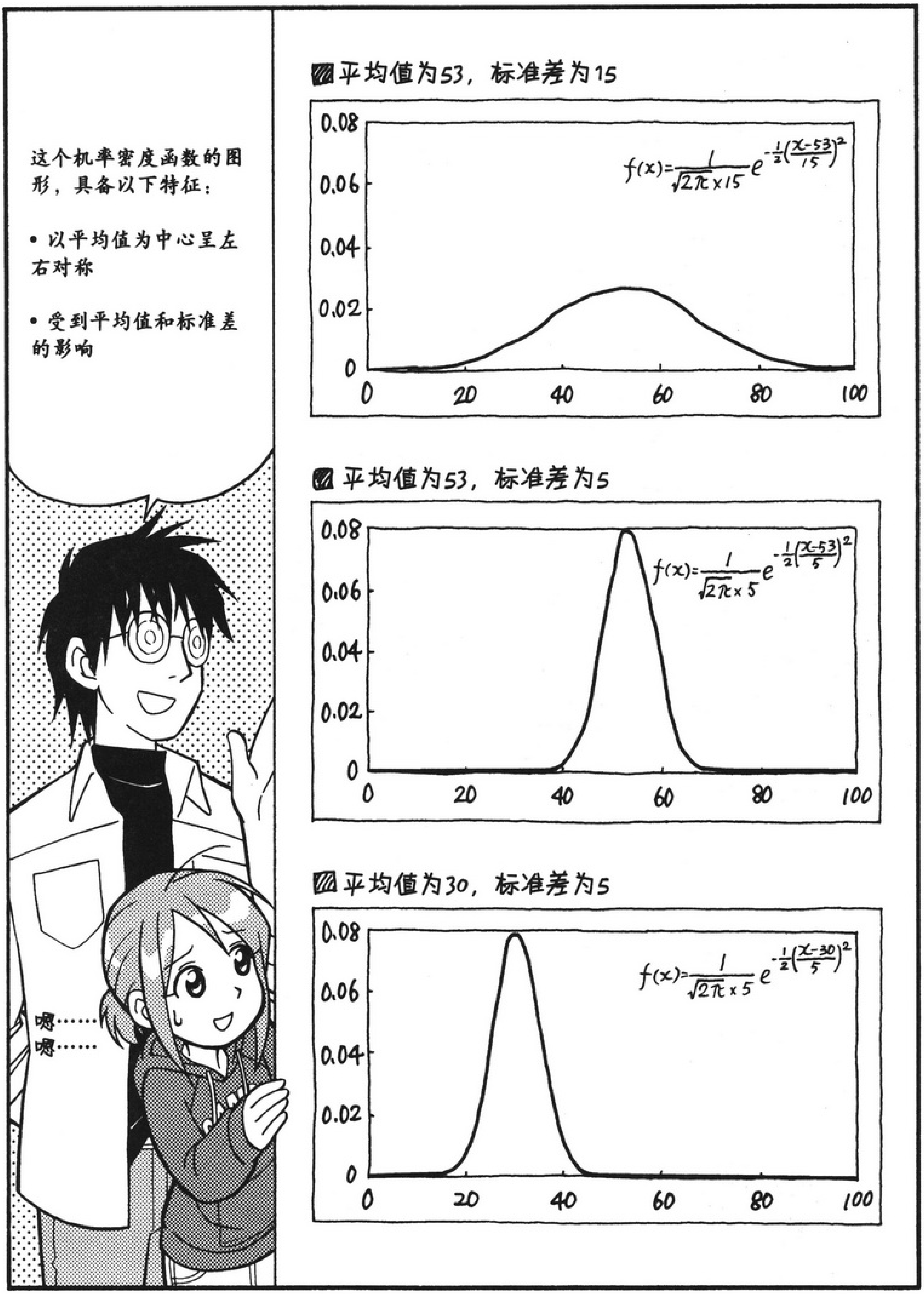

组距和"英语测验结果"的直方图
组距为10
0.4
0.3
0.2
0.1
0
组距为5
0.2
0.1
组距为3
0.1
曲线
0.05
哦!
渐渐接近于曲线啊!

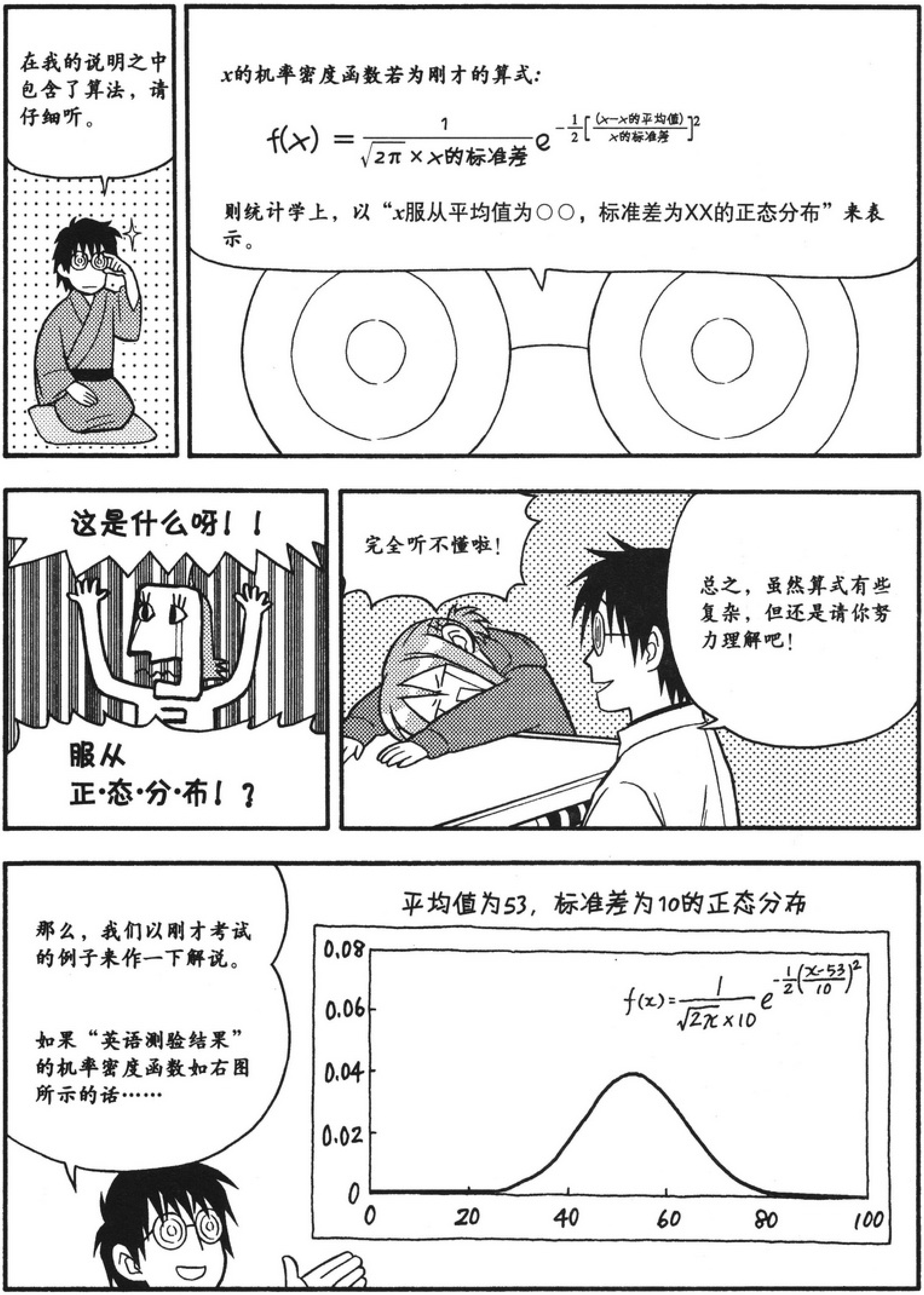


1．机率密度函数：Probability Density Function，简称“pdf”。

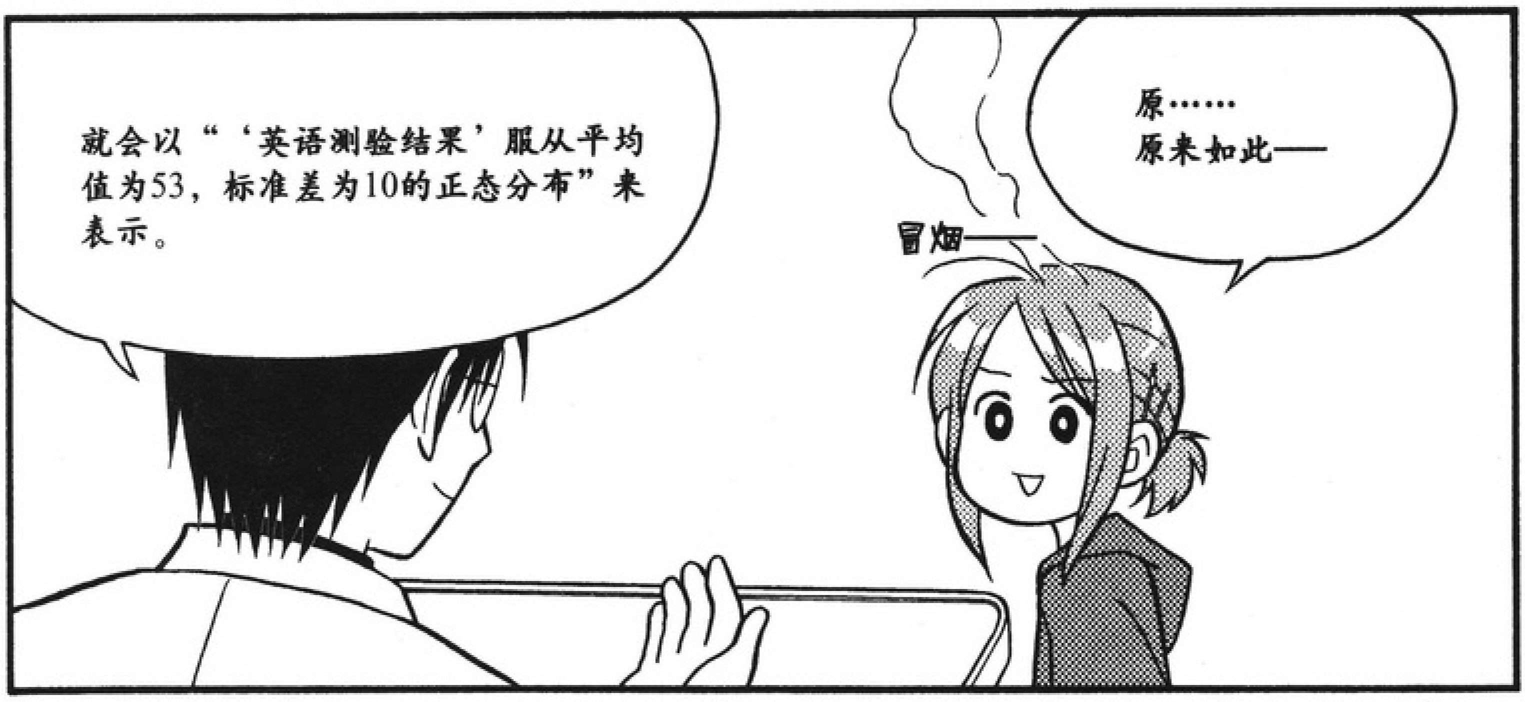
2. 正态分布


$f(x)=\frac{1}{\sqrt{2\pi}\times x\text{的标准差}}e^{-\frac{1}{2}\left[\frac{(x-x\text{的平均值})}{x\text{的标准差}}\right]^2}$
先来看看这个。
这是什么鬼东西呀！？
啊哈哈哈
这是在统计学上经常出现的机率密度函数呦！
这里的“e”是什么啊？
“e”被称为“自然对数的底”，其值大约为2.7182……
底
把它想成跟“π”类似的数就好了！
嗯！
这样或许更容易理解……

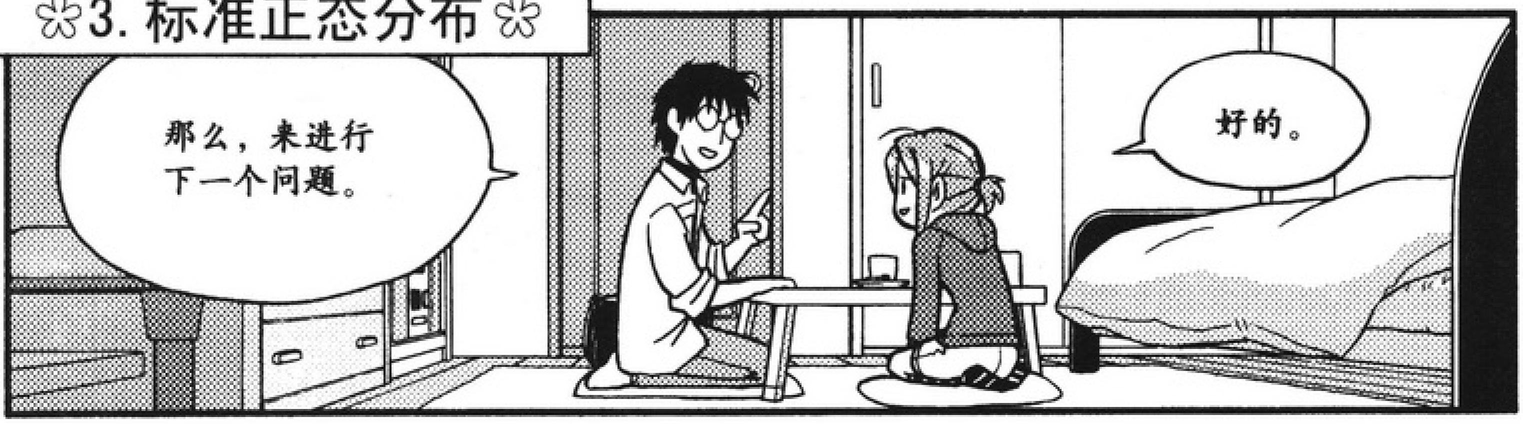

这个机率密度函数的图形，具备以下特征：
• 以平均值为中心呈左右对称
• 受到平均值和标准差的影响
嗯……嗯……
平均值为53，标准差为15
$f(x)=\frac{1}{\sqrt{2\pi}\times 15}e^{-\frac{1}{2}\left(\frac{x-53}{15}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
平均值为53，标准差为5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-53}{5}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100
平均值为30，标准差为5
$f(x)=\frac{1}{\sqrt{2\pi}\times 5}e^{-\frac{1}{2}\left(\frac{x-30}{5}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100

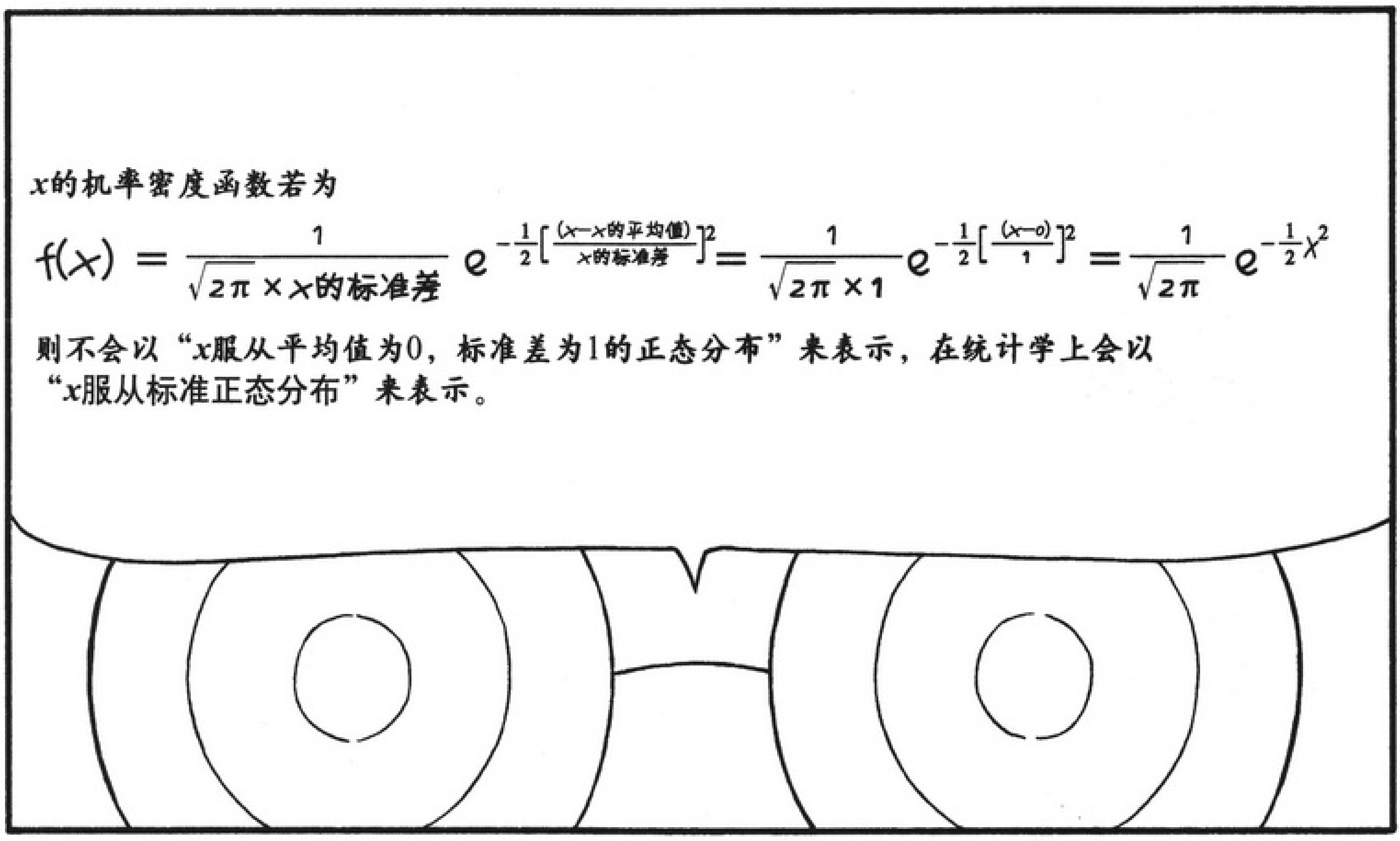


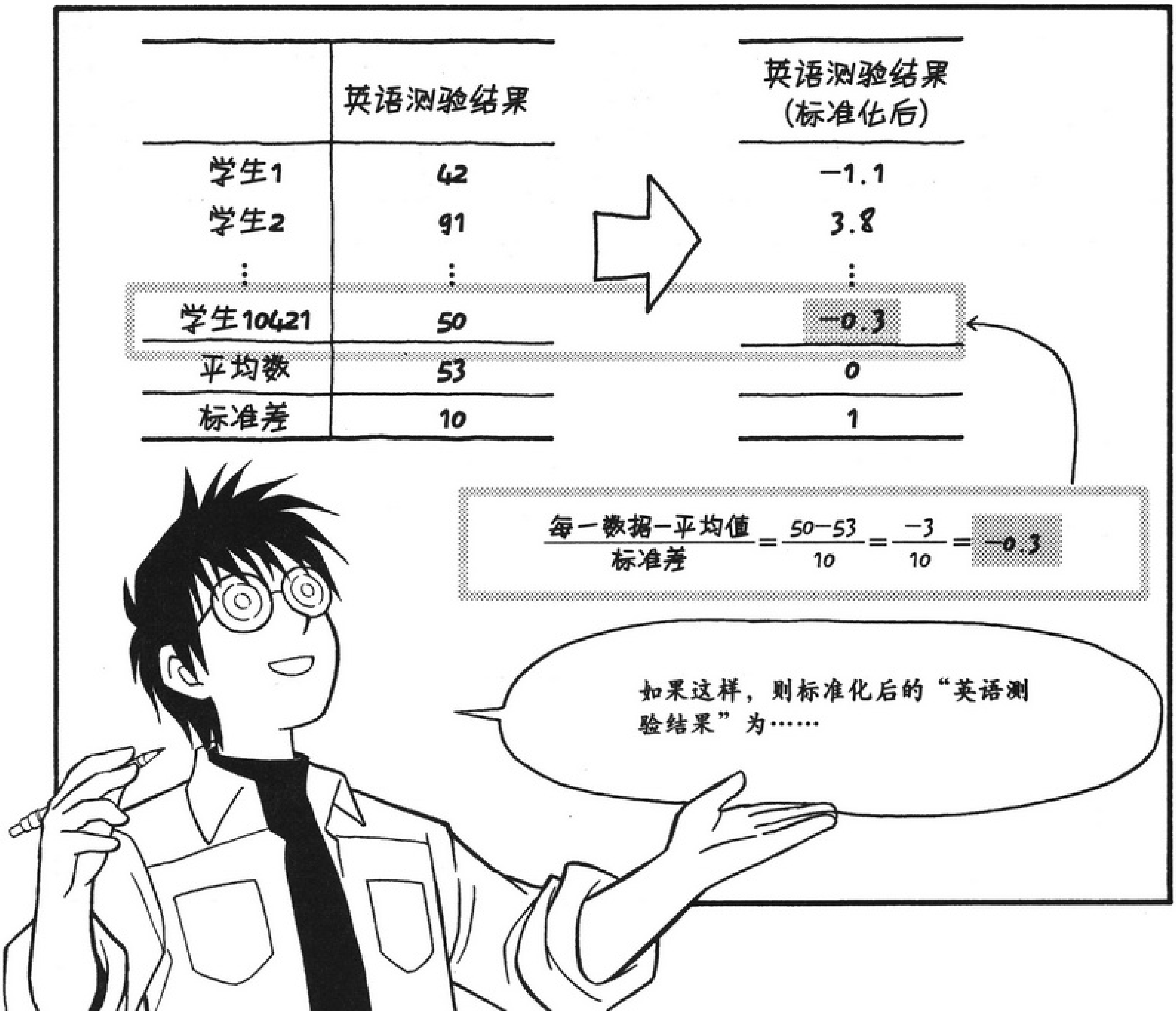
在我的说明之中包含了算法，请仔细听。
x的机率密度函数若为刚才的算式：
$f(x) = \frac{1}{\sqrt{2\pi} \times x\text{的标准差}} e^{-\frac{1}{2}\left[\frac{(x - x\text{的平均值})}{x\text{的标准差}}\right]^2}$
则统计学上，以“x服从平均值为○○，标准差为XX的正态分布”来表示。
这是什么呀！！
服从
正·态·分·布！？
完全听不懂啦！
总之，虽然算式有些复杂，但还是请你努力理解吧！
那么，我们以刚才考试的例子来作一下解说。
如果“英语测验结果”的机率密度函数如右图所示的话……
平均值为53，标准差为10的正态分布
$f(x) = \frac{1}{\sqrt{2\pi} \times 10} e^{-\frac{1}{2}\left(\frac{x-53}{10}\right)^2}$
0.08
0.06
0.04
0.02
0
0
20
40
60
80
100


就会以“‘英语测验结果’服从平均值为53，标准差为10的正态分布”来表示。
原……
原来如此——
冒烟——

✿3. 标准正态分布✿

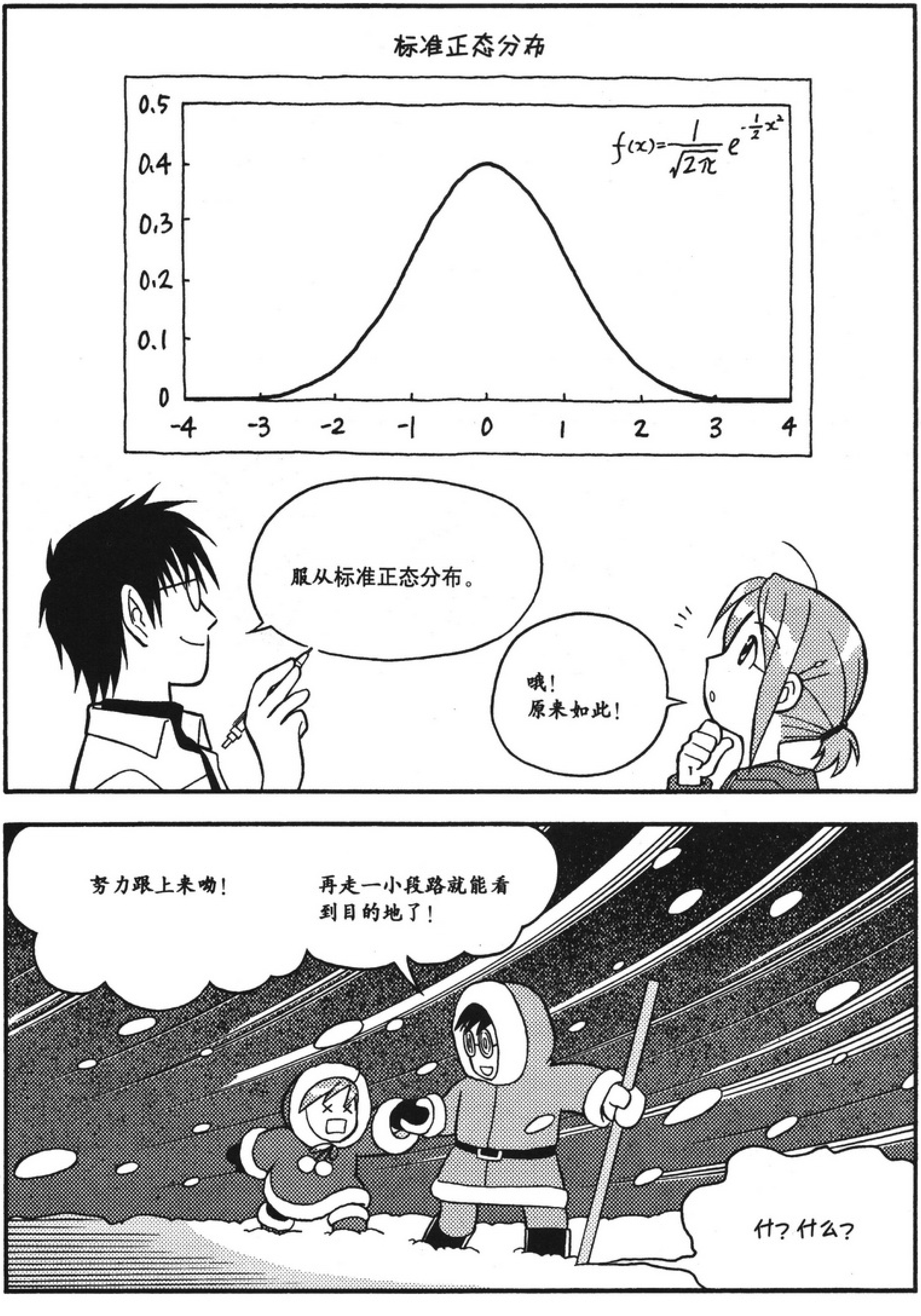

那么，来进行下一个问题。
好的。

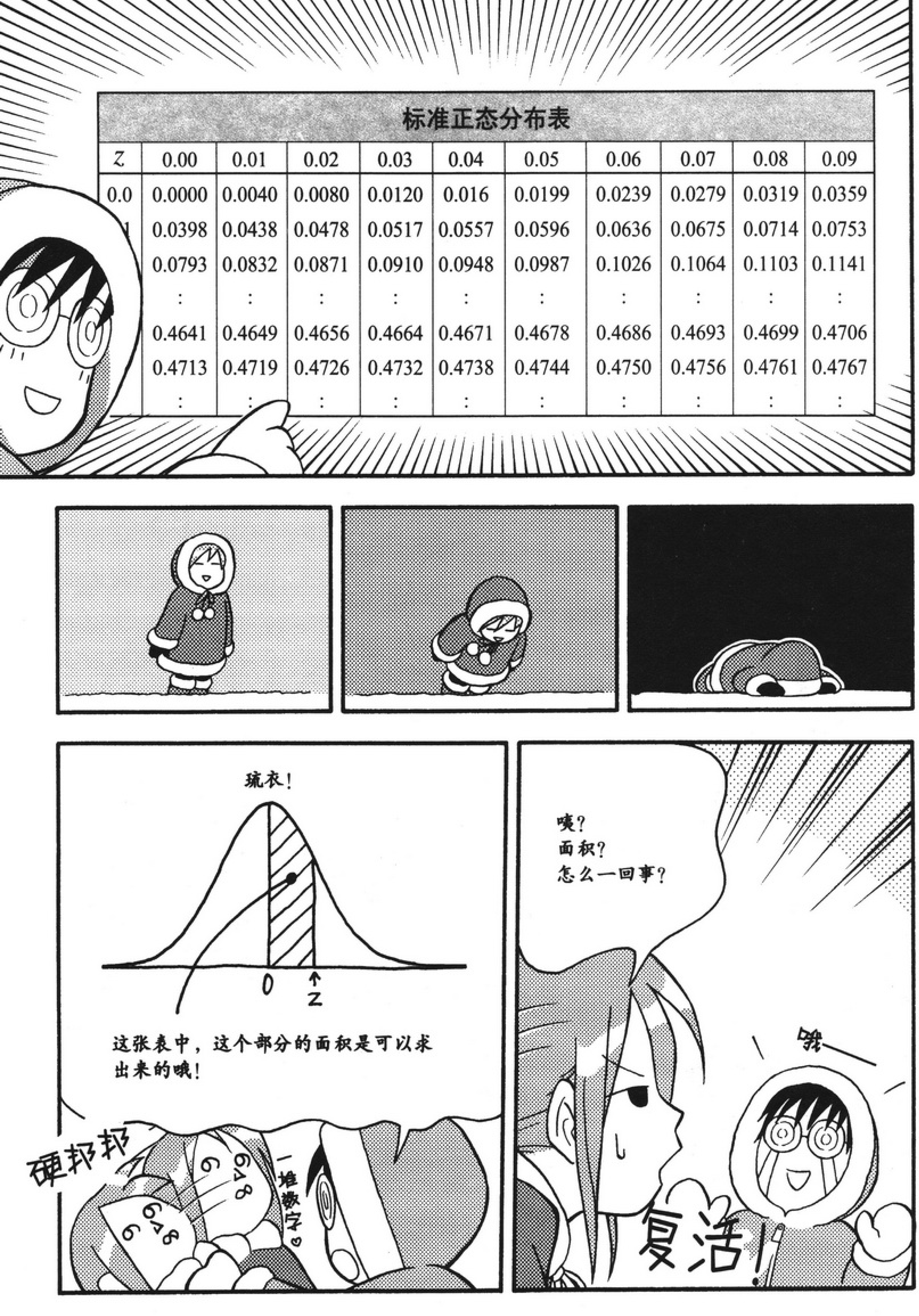

x的机率密度函数若为
$$f(x)=\frac{1}{\sqrt{2\pi}\times x\text{的标准差}}e^{-\frac{1}{2}\left[\frac{(x-x\text{的平均值})}{x\text{的标准差}}\right]^2}=\frac{1}{\sqrt{2\pi}\times 1}e^{-\frac{1}{2}\left[\frac{(x-0)}{1}\right]^2}=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$
则不会以“x服从平均值为0，标准差为1的正态分布”来表示，在统计学上会以“x服从标准正态分布”来表示。



	英语测验结果	英语测验结果（标准化后）
学生1	42	−1.1
学生2	91	3.8
⋮	⋮	⋮
学生10421	50	−0.3
平均数	53	0
标准差	10	1

$$\frac{\text{每一数据}-\text{平均值}}{\text{标准差}}=\frac{50-53}{10}=\frac{-3}{10}=-0.3$$


标准正态分布
0.5
0.4
0.3
0.2
0.1
0
-4 -3 -2 -1 0 1 2 3 4
$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$
服从标准正态分布。
哦！
原来如此！
努力跟上来呦！
再走一小段路就能看到目的地了！
什？什么？



标准正态分布表										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2	0.00	0.01	0.02	0.03	0.04	[illegible]	0.06	0.07	0.08	[illegible]
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	[illegible]
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	[illegible]	[illegible]
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	[illegible]	0.1141
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮


那么，现在开始要说明的内容就是
今天的“主菜”。请用心听讲。
哦——

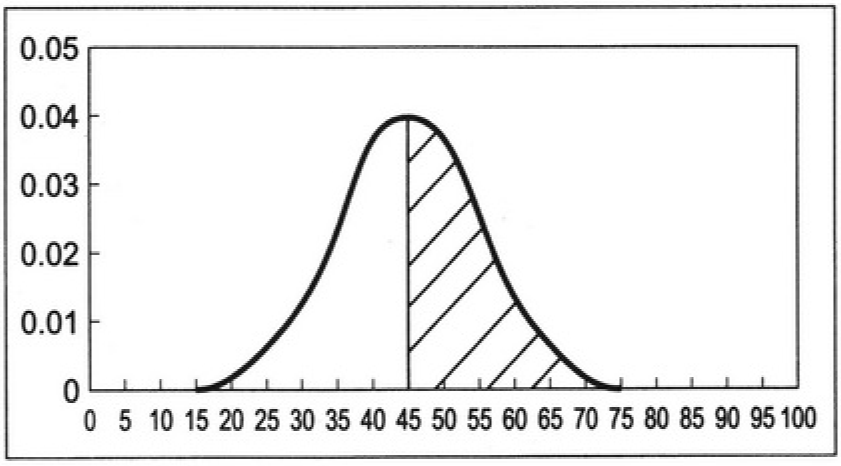

标准正态分布的图形和横轴围成的面积，
与其比例及机率是相同的。


举两个例子给你看，你努力尝试着去理解吧！

例 1

B县的全体高一学生参加某补习班的数学测验。计算分数后，得知“数学测验结果”可视为服从平均值为45，标准差为10的正态分布。那么，请思考看看。下列5种表示方法同义。

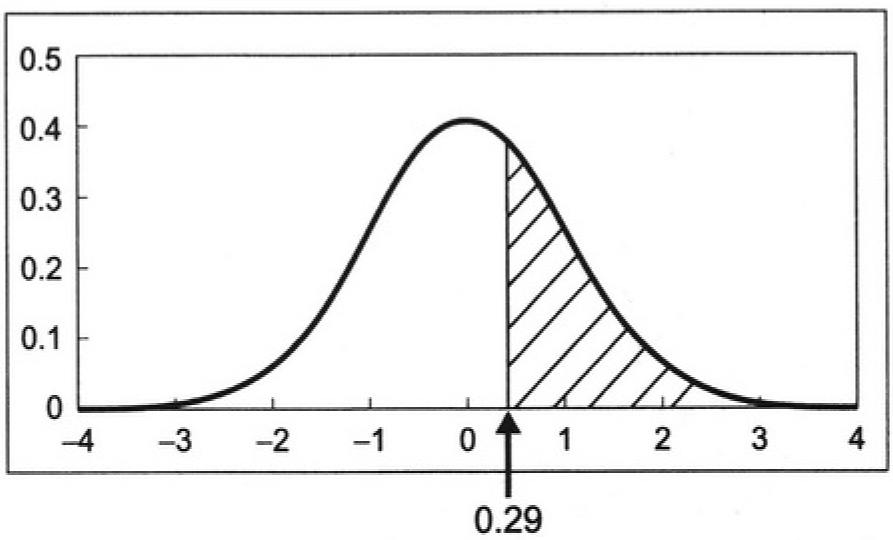
① 平均值为45，标准差为10的正态分布表中，下图斜线部分的面积为0.5。



② 得分在45分以上的考生比例，占全体考生总数的0.5（=50%）。

③ 从全体考生中，随机抽出一人，其得分在45分以上的机率为0.5（=50%）。

④ 在“数学测验结果”标准化后的标准正态分布中，0以上的考生比例，占全体考生总数的0.5（=50%）。



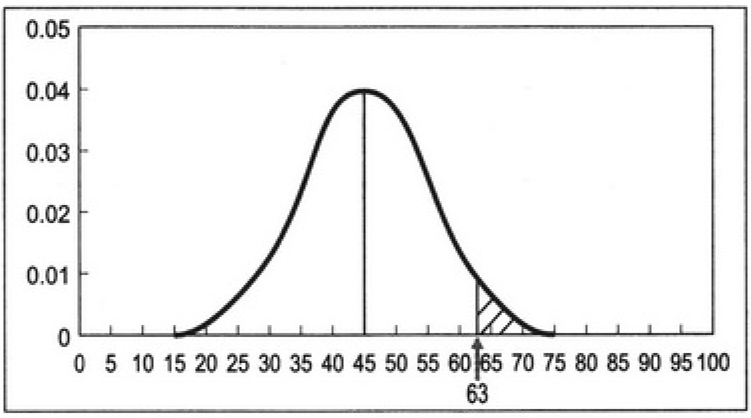
⑤ 从全体考生中，随机抽出一人。在“数学测验结果”标准化后的标准正态分布中，他的标准计分为0以上的机率为0.5（=50%）。


因为平均值是45分，所以可以画出以45分处为高峰，左右对称的图形。
没错！
其中得分在45分以上的，正好占图形的右半边，因此是50%。
是的！
这种程度的内容我当然可以理解啦。
呼……这下安心了。那么例Ⅱ为例Ⅰ的应用题哦！

例Ⅱ

B县的全体高一学生参加某补习班的数学测验。计算分数后，得知“数学测验结果”可视为服从平均值为45，标准差为10的正态分布。那么，请思考看看。下列提示的5点均为同义。并请先阅读第④点。

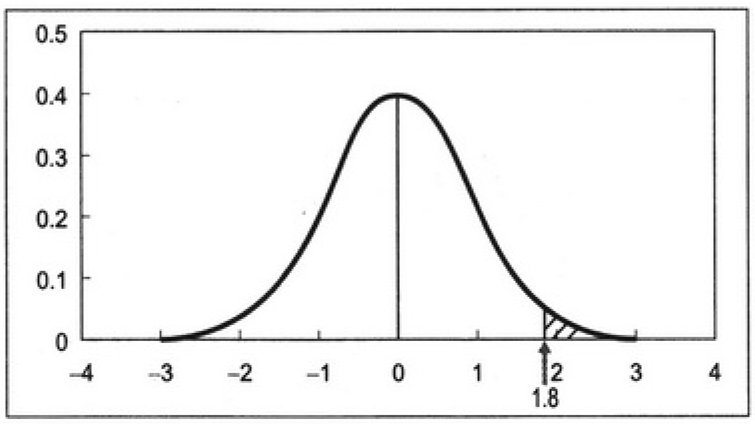
①平均值为45，标准差为10的正态分布中，下图斜线部分的面积为0.5−0.4641=0.0359。



②得分在63分以上的考生，占全体考生的0.5−0.4641=0.0359（=3.59%）。

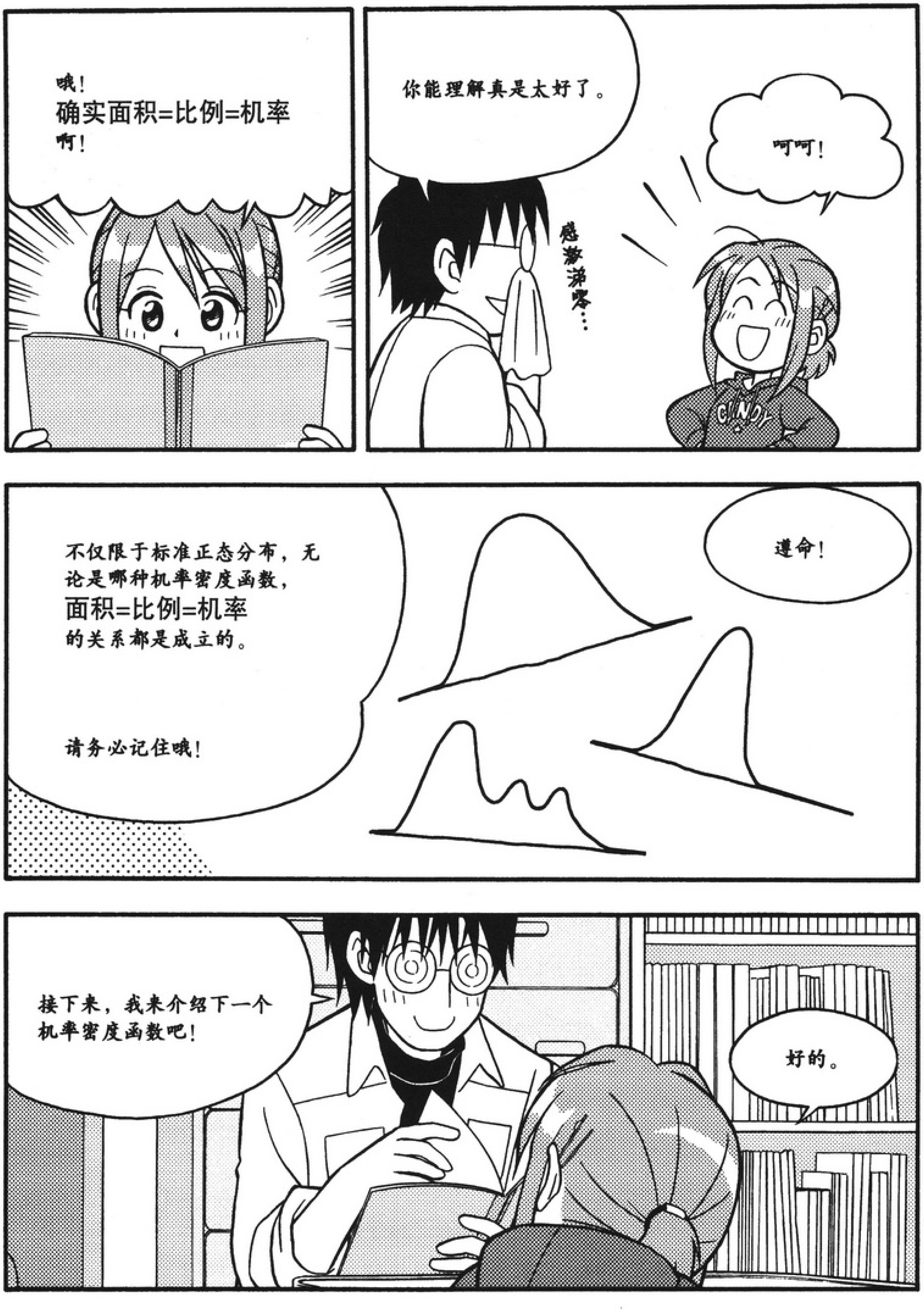
③从全体考生之中，随机抽出一人，其得分在63分以上的机率为0.5−0.4641=0.0359（=3.59%）。

④在“数学测验结果”标准化后的标准正态分布中，

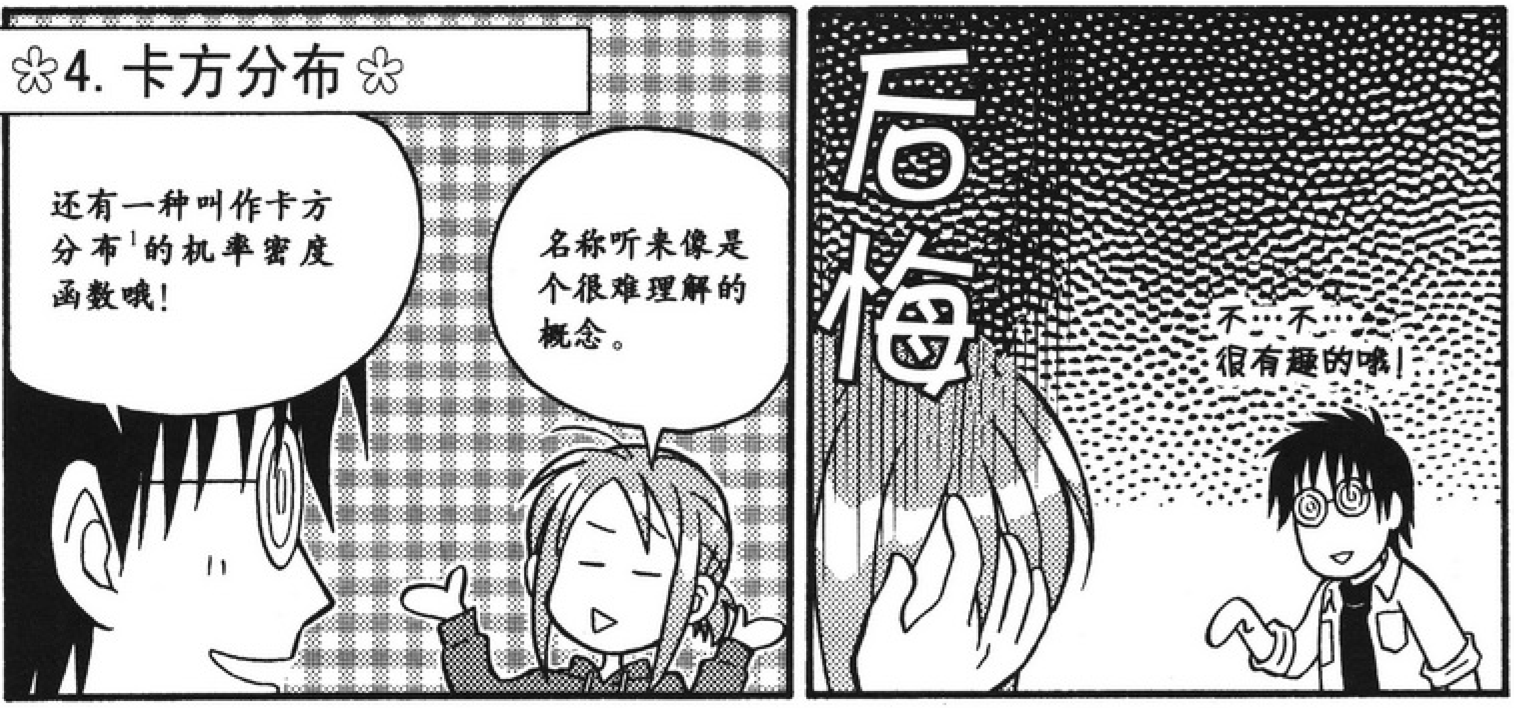


$1.8=\frac{18}{10}=\frac{63\text{-}45}{10}=\frac{\text{每一数据—平均值}}{\text{标准差}}$ 以上的考生比例，从标准正态分布表可清楚得知，占全体考生的0.5−0.4641=0.0359（=3.59%）。

⑤从全体考生之中，随机抽出一人。在“数学测验结果”标准化后的标准正态分布中，他的标准计分为1.8以上的机率为0.5−0.4641=0.0359（=3.59%）。


哦！
确实面积=比例=机率
啊！
你能理解真是太好了。
感激涕零…
呵呵！
不仅限于标准正态分布，无论是哪种机率密度函数，
面积=比例=机率
的关系都是成立的。
请务必记住哦！
遵命！
接下来，我来介绍下一个机率密度函数吧！
好的。

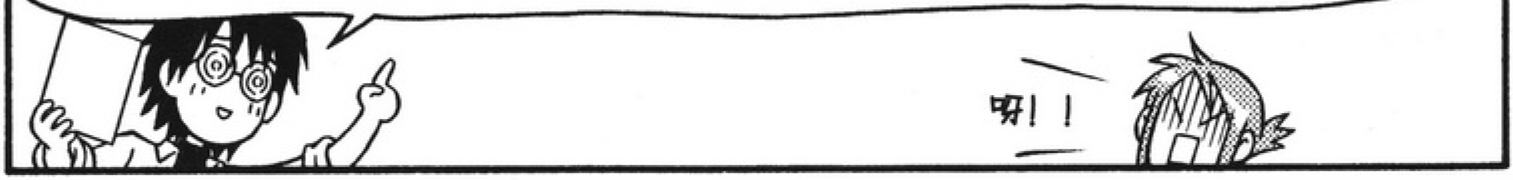
✿4. 卡方分布✿



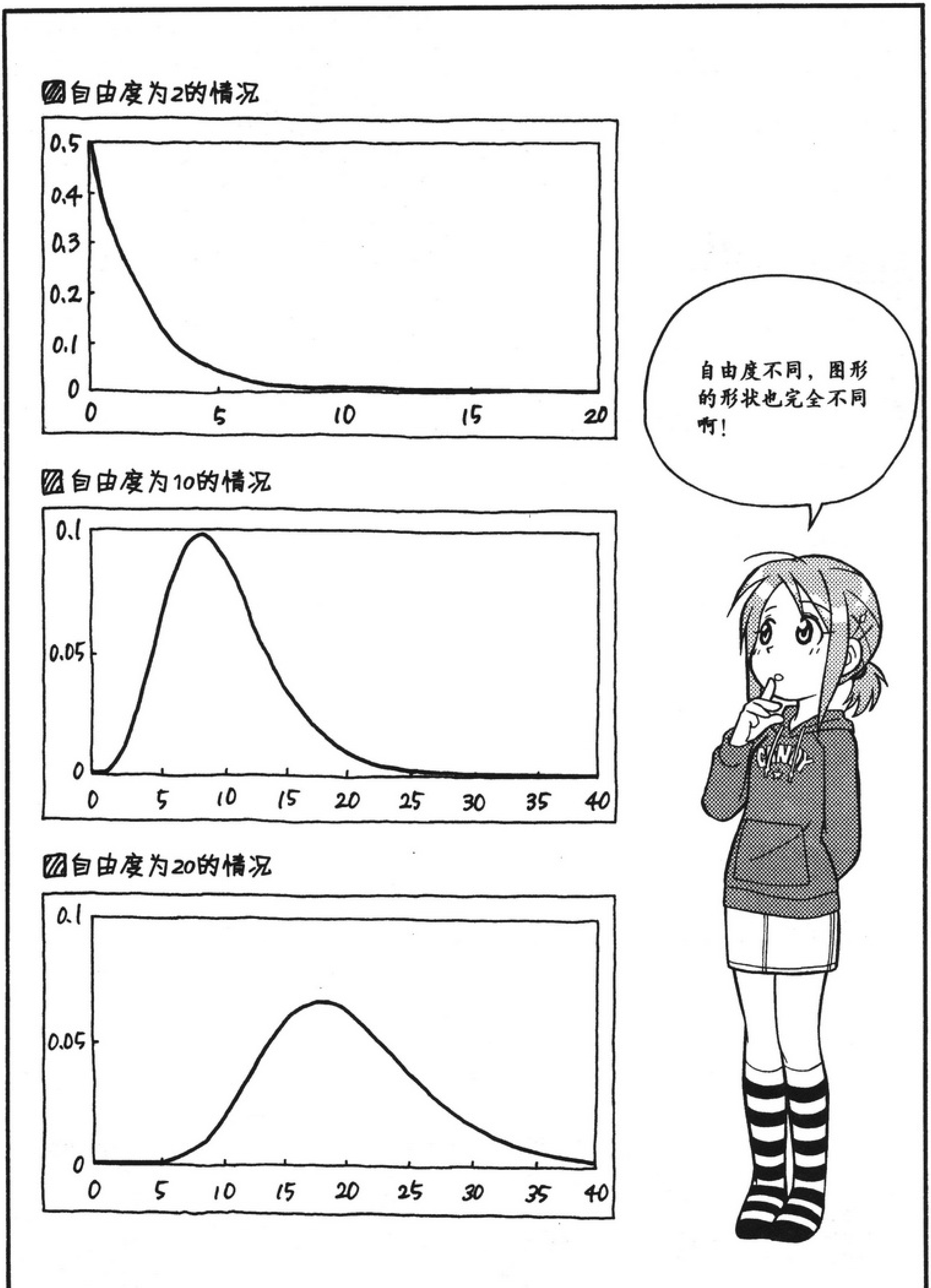
x的机率密度函数若为，

$$f(x)=\begin{cases} x>0\text{时，}\dfrac{1}{2^{\frac{\text{自由度}}{2}}\times\int_0^{\infty}x^{\frac{\text{自由度}}{2}-1}e^{-x}dx}\times x^{\frac{\text{自由度}}{2}-1}\times e^{-\frac{x}{2}} \\ \text{上述以外的情况则为}0 \end{cases}$$

在统计学上，用“x服从自由度[2]为〇〇的卡方分布”来表示。

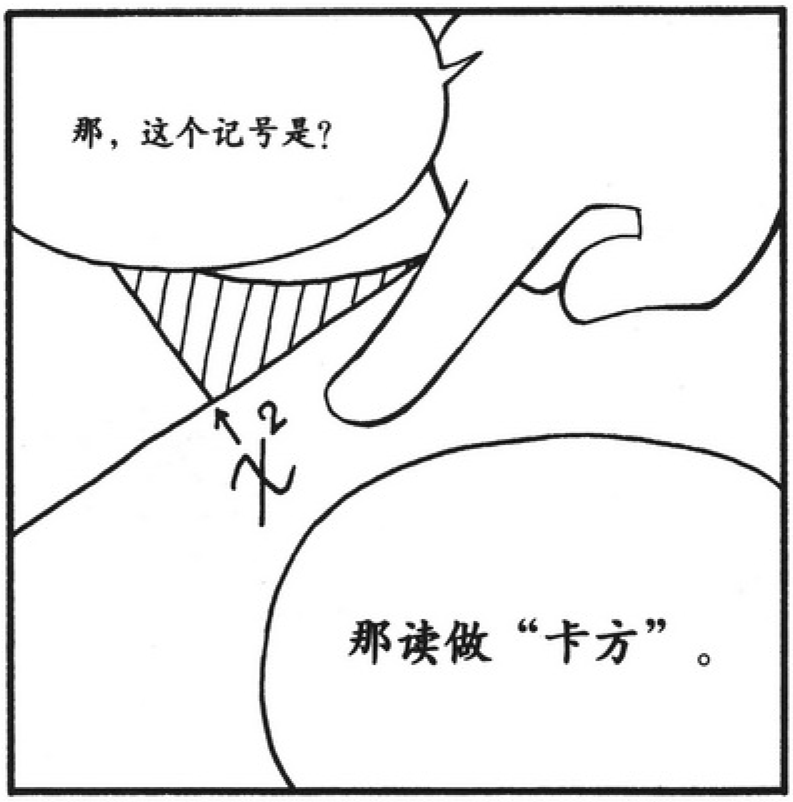


1. 卡方分布：Chisquare Distribution。 2. 自由度：Degree of Freedom。


自由度为2的情况
0.5
0.4
0.3
0.2
0.1
0
0 5 10 15 20
自由度为10的情况
0.1
0.05
0
0 5 10 15 20 25 30 35 40
自由度为20的情况
0.1
0.05
0
0 5 10 15 20 25 30 35 40
自由度不同，图形的形状也完全不同啊！

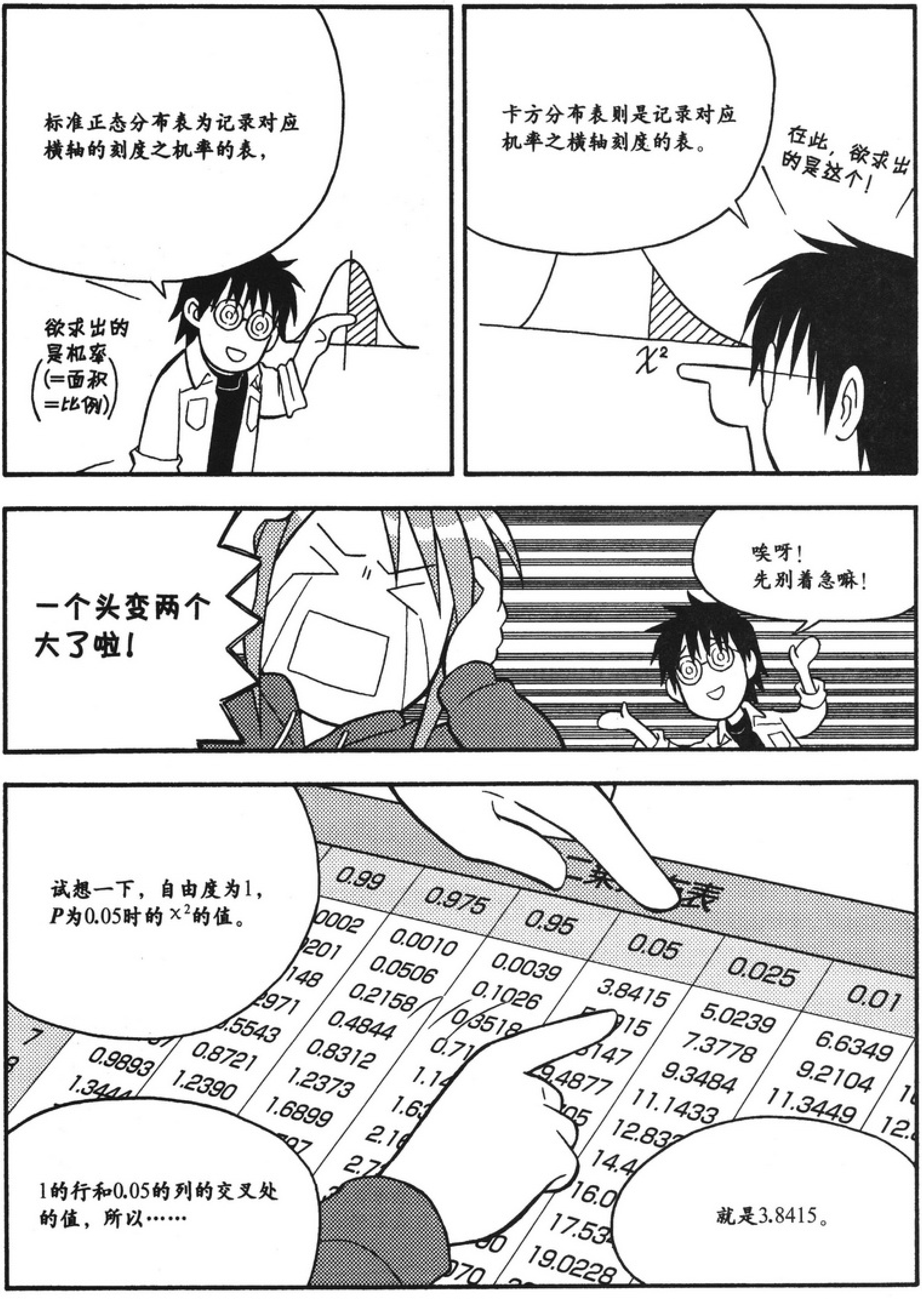

那“自由度”到底是什么东西呀！？
啊！对了！
1次函数f(x)=ax+b之中，a称为什么呢？
啊……冷不防这么一问……
b
f(x)=ax+b
嗯……是“斜率”吗？
没错！随着a值的改变，图形的倾斜度也会跟着改变吧！
对呀！
“自由度”和“斜率”一样，都是会影响图形形状的数值。
所以只要自由度的值有所改变，图形的形状也会随之改变。
原来如此。这就是自由度呀？
是的。就只是这样而已。


就像标准正态分布有标准正态分布表一样，卡方分布也有卡方分布表。
所谓的卡方分布表，
就是记录了对应这个部分的机率（=面积=比例）P的横轴刻度χ^2之值的表。
P
0
χ^2

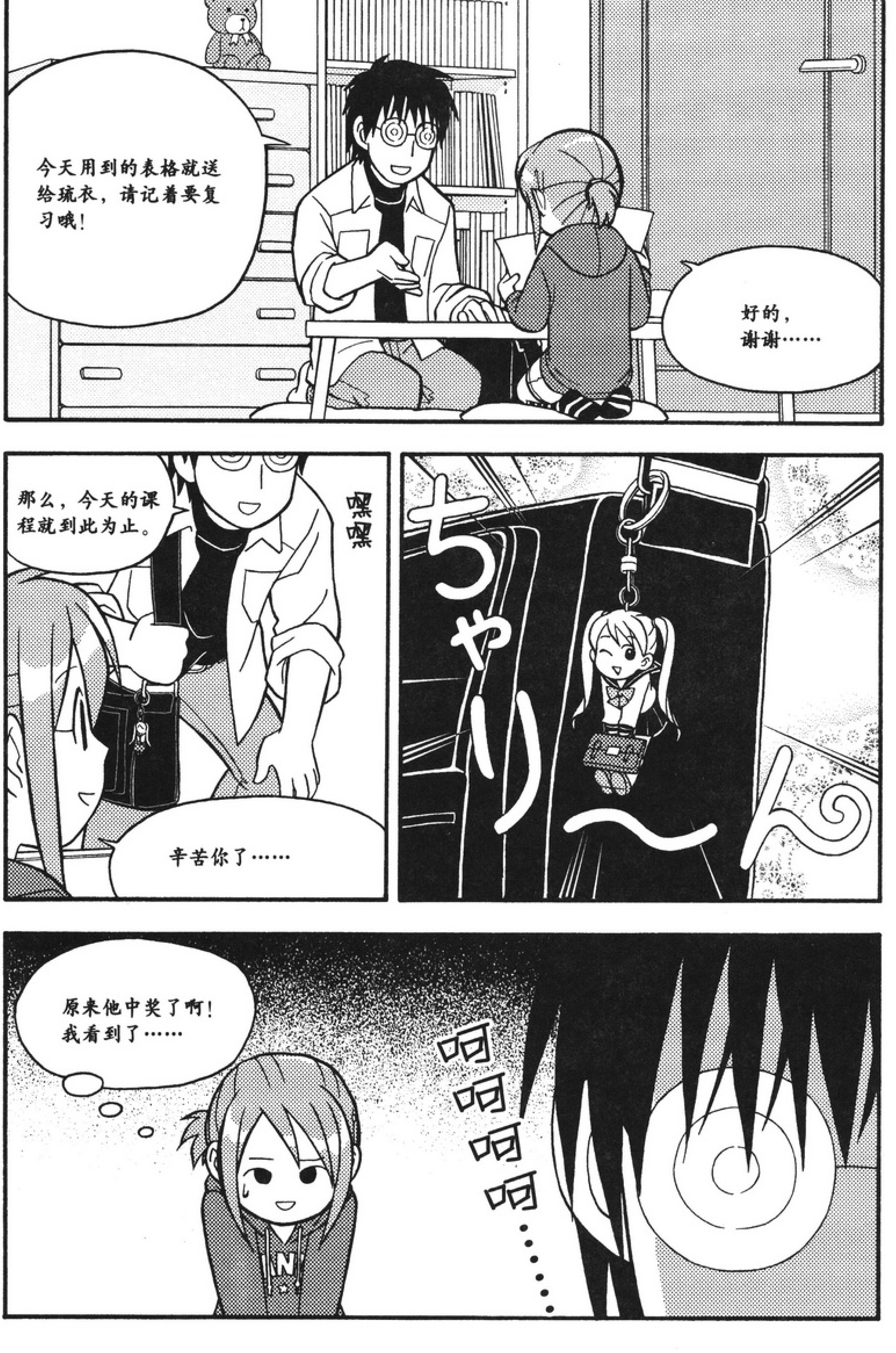


卡方分布表

自由度 \ P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



标准正态分布表为记录对应横轴的刻度之机率的表，
欲求出的是机率（=面积=比例）
卡方分布表则是记录对应机率之横轴刻度的表。
在此，欲求出的是这个！
χ^2
一个头变两个大了啦！
唉呀！先别着急嘛！
试想一下，自由度为1，P为0.05时的χ^2的值。
0.99 0.975 0.95 0.05 0.025 0.01
1的行和0.05的列的交叉处的值，所以……
就是3.8415。


今天用到的表格就送给琉衣，请记着要复习哦！
好的，谢谢……
那么，今天的课程就到此为止。
嘿嘿
辛苦你了……
ちゃり〜ん
原来他中奖了啊！我看到了……
呵呵呵呵……

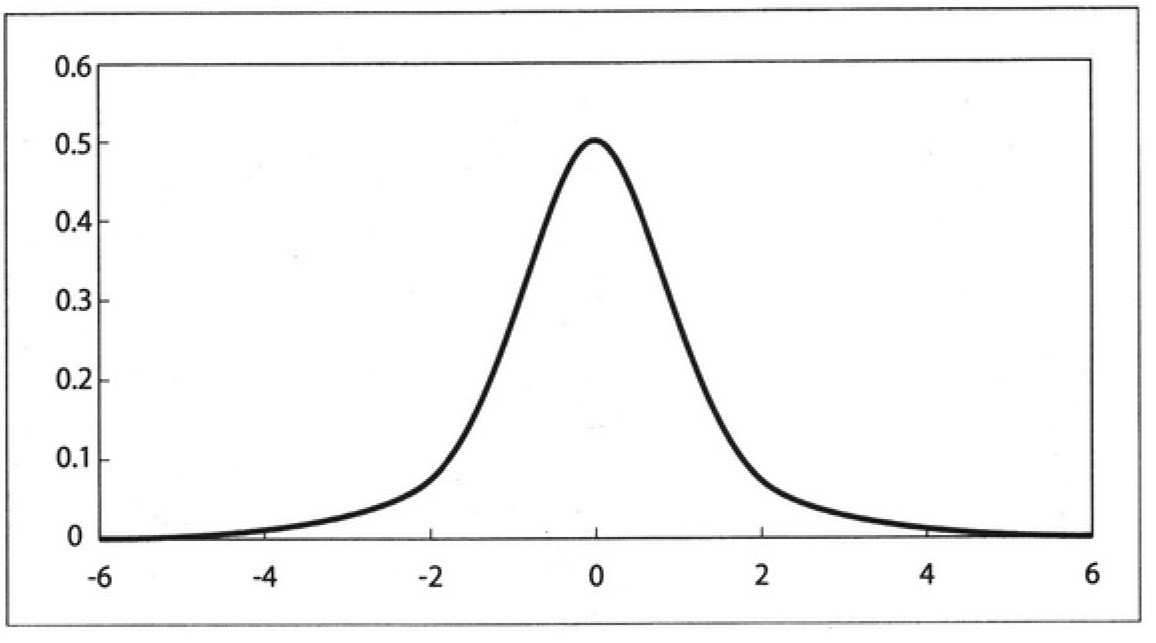
✿5. t分布✿

统计学上，以下的机率密度函数经常出现。

$$f(x)=\frac{\int_0^{\infty} x^{\frac{\text{自由度}+1}{2}-1}e^{-x}\,dx}{\sqrt{\text{自由度}\times\pi}\times\int_0^{\infty} x^{\frac{\text{自由度}}{2}-1}e^{-x}\,dx}\times\left(1+\frac{x^2}{\text{自由度}}\right)^{-\frac{\text{自由度}+1}{2}}$$

x的机率密度函数若如上述，在统计学上则以“x服从自由度为××的t分配”来表示。

■自由度为5的情况



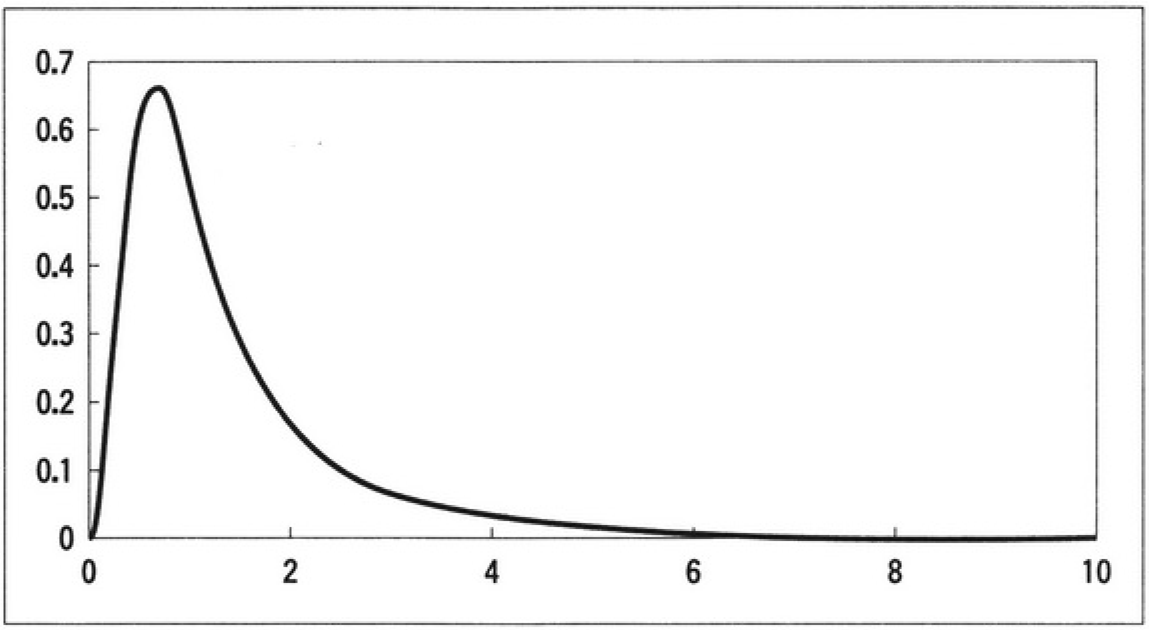
✿6. F分布✿

统计学上，以下的机率密度函数也是经常出现的主题。

$$f(x)=\begin{cases} x>0\text{时} & \dfrac{\left(\int_0^{\infty} x^{\frac{\text{第1自由度}+\text{第2自由度}}{2}-1}e^{-x}dx\right)\times(\text{第1自由度})^{\frac{\text{第1自由度}}{2}}\times(\text{第2自由度})^{\frac{\text{第2自由度}}{2}}}{\left(\int_0^{\infty} x^{\frac{\text{第1自由度}}{2}-1}e^{-x}dx\right)\times\left(\int_0^{\infty} x^{\frac{\text{第2自由度}}{2}-1}e^{-x}dx\right)}\times\dfrac{x^{\frac{\text{第1自由度}}{2}-1}}{(\text{第1自由度}\times x+\text{第2自由度})^{\frac{\text{第1自由度}+\text{第2自由度}}{2}}} \\ \text{上述以外的情况为0} & \end{cases}$$

x的机率密度函数若如上述所示，在统计学上则以“x服从自由度为○○，第2自由度为××的F分布”来表示。

■第1自由度为10，第2自由度为5的情况



❀ 7. “××分布”和EXCEL ❀

如果不使用标准正态分布表及卡方分布表来计算机率及横轴的刻度，在电脑尚未普及时（约是20世纪90年代初期），这些计算对个人而言是相当浩大的工程。因此，这些分布表实在是相当重要的“宝物”。然而，现今已经不太使用分布表了。因为使用EXCEL的函数计算功能，便可轻松地求出分布表中的值，不仅如此，比起分布表，EXCEL还可以求出更多种类的值。

我将与“××分配”相关的函数总整理如下表。

◆表5.1　与“××分配”相关的函数

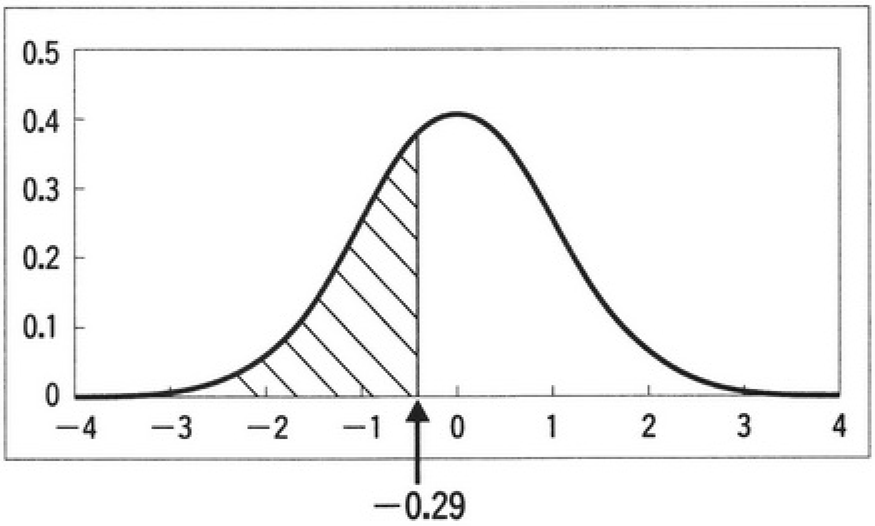
分布	函数	函数的特征
正态分布[1]	NOPMDIST	可计算对应横轴刻度的机率
正态分布	NORMINV	可计算对应机率的横轴刻度
标准正态分布	NOPMDIST	可计算对应横轴刻度的机率
标准正态分布	NORMSINV	可计算对应机率的横轴刻度
卡方分布	CHIDIST	可计算对应横轴刻度的机率
卡方分布	CHIINV	可计算对应机率的横轴刻度
t分布	TDIST	可计算对应横轴刻度的机率
t分布	TINV	可计算对应机率的横轴刻度
F分布	FDIST	可计算对应横轴刻度的机率
F分布	FINY	可计算对应机率的横轴刻度

1. 正态分布：由于正态分布的机率密度函数受到平均值和标准差的影响。因此即使想做出“正态分布表”也是不可能的。然而，利用EXCEL来求出与“正态分布表”相当的值却非常便利。

例题和解答

例题

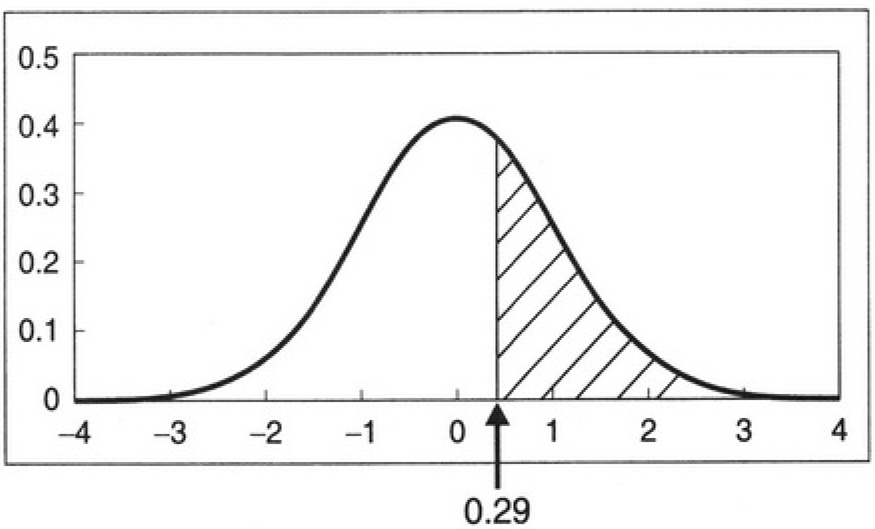
（1）请利用93页的标准正态分布表表示出下图斜线部分的机率。



（2）请利用103页的卡方分布表求出自由度为2，P为0.05时的χ^2的值。

解答

（1）必须求出的机率，和下图斜线部分的机率相同。



欲求出z=0.29=0.2+0.09的情况下的机率，可从标准正态分配表得知，是0.1141。因此，必须求出的机率为0.5-0.1141=0.3859。

（2）根据卡方分布表，必须求出χ^2的值，其值为5.9915。

总整理

- 代表性的机率密度函数，可举出与下列对应者：
  - ■正态分布
  - ■标准正态分布
  - ■卡方分布
  - ■*t*分布
  - ■*F*分布
- 机率密度函数的图形和横轴所围成的面积为1。
- 机率密度函数的图形和横轴所围成的面积，可视为与比例及机率相同。
- 若利用“××分布表”或Excel的函数，则可求出，
  - ■对应横轴刻度的机率
  - ■对应机率的横轴刻度

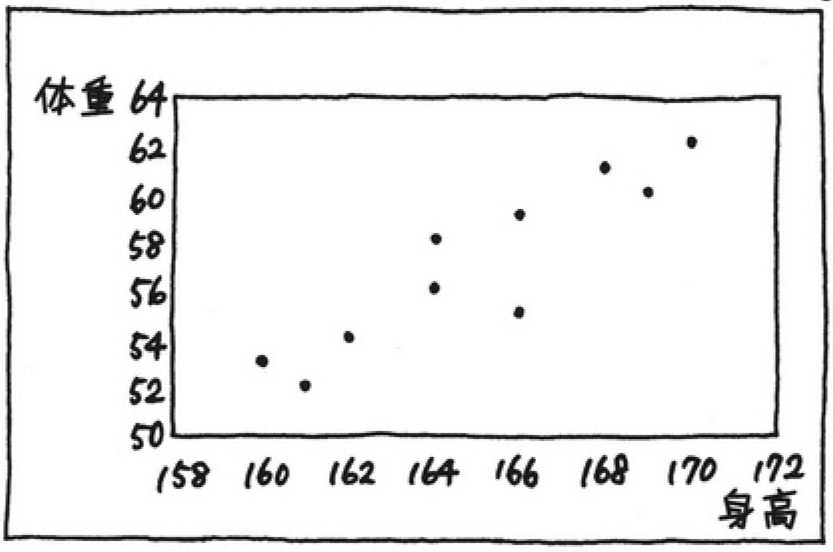
◆第 6 章◆

双变量的相关分析


叭哒叭哒
原来如此——
偶尔在外面上课
也不错啊！
叭哒
叭哒
叭哒
叭哒
你那是什么打扮呀！
穿错鞋子了吧！
啊——
真的呀！
啊……
如果能和五十
岚先生在一起
的话该多好
啊……
五十岚先生……
五十岚先生啊？
糟糕……心目
中的形象越来
越模糊
不
安
今天就来谈谈双
变量吧！

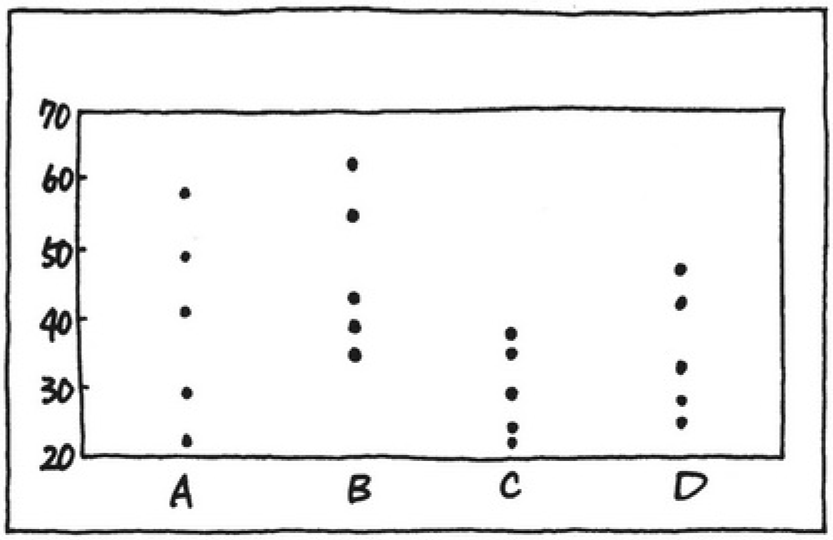

都是这个人的形象太强烈了哦!
你在听吗?
啊……
我在听呀!
例如,"是否身高越高,体重就越重",或是"是否因年龄不同,喜欢的啤酒品牌也不同",以及……
"是否居住地不同,支持的政党也会不同",等等。
啊,谢谢。
"吵架后会不会肚子饿"等等。
的确有这首歌啊!
那么,
拿出来了。
啪

"身高"和"体重"的散布图(Scatter Diagram)



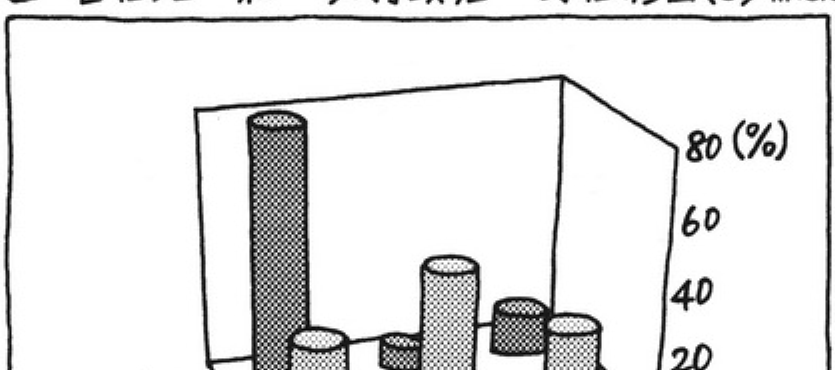
数值和数值

"喜欢的啤酒品牌"和"年龄"的散布图

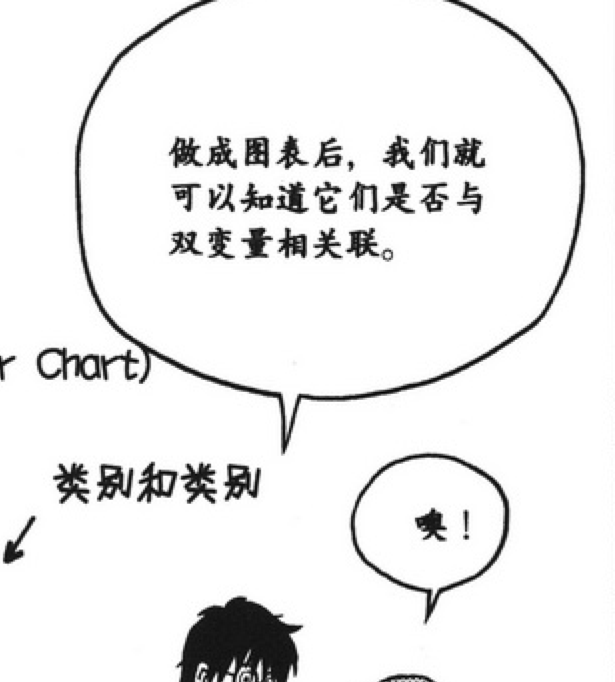


数值和类别

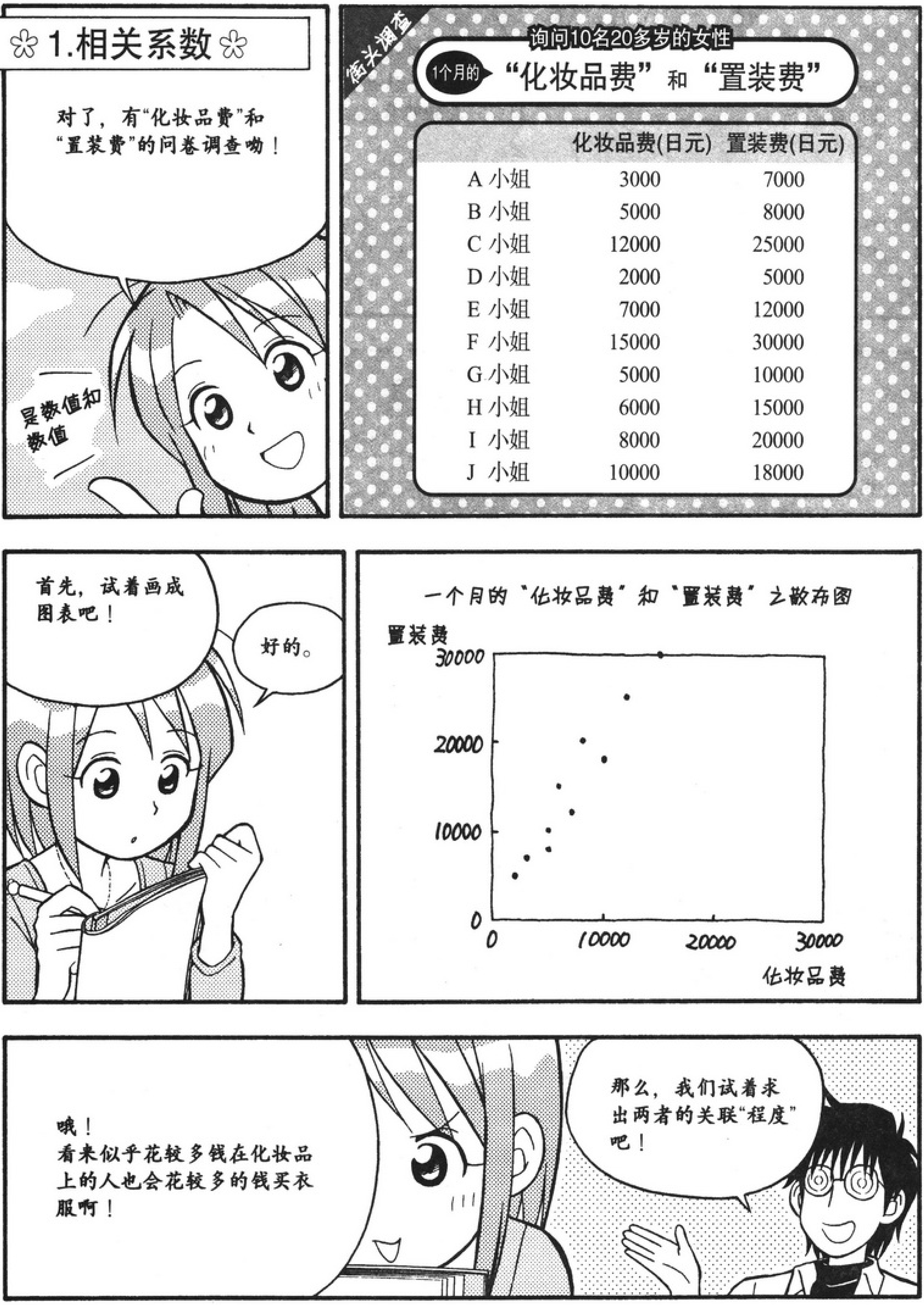
"居住地"和"支持政党"的柱形图(Cylinder Chart)



类别和类别



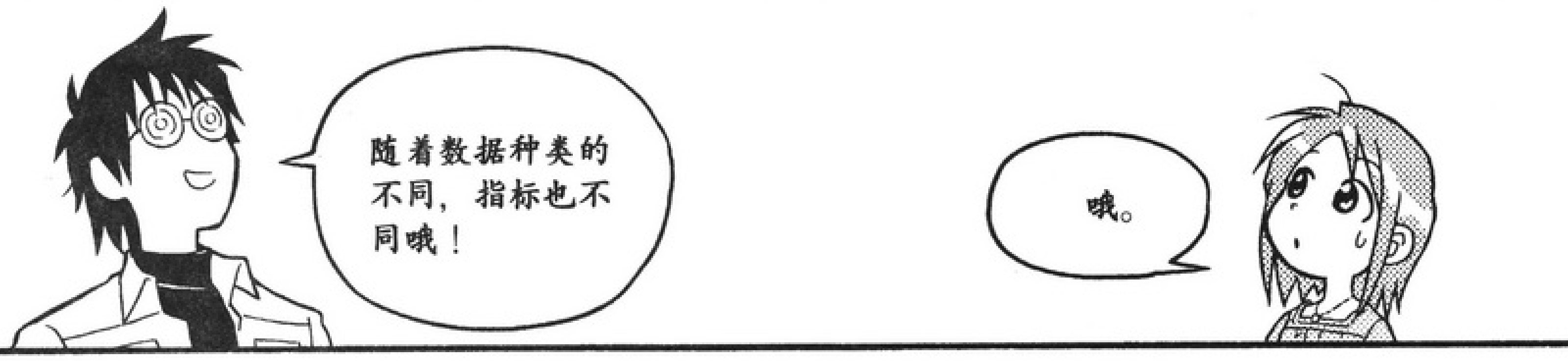

但是！
关联性究竟有
多高呢？
很可惜！
我们无法确切得知两者
关联的程度。
那该怎么办呢？
？
通常我会结合图形，
然后求出用来表示2变
数的关联程度的指标
值。
哦——
那么……
P.girls.
特集！
时尚女孩
卷调查
用在流行装扮的费用？
超人气品牌
效果卓越
像这样的内容也能用
统计学来分析吗？
啊！
这个内容还真适合
呢！

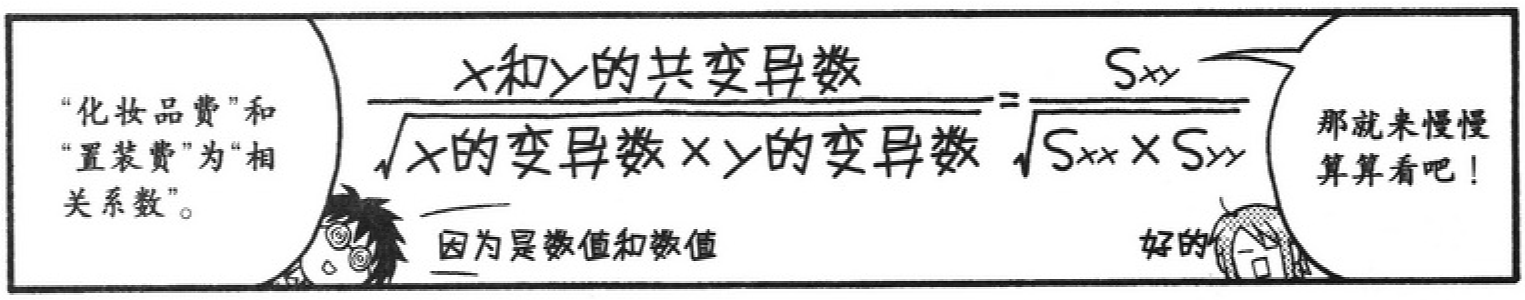

1.相关系数
对了，有“化妆品费”和“置装费”的问卷调查呦！
是数值和数值
街头调查
询问10名20多岁的女性
1个月的“化妆品费”和“置装费”

	化妆品费(日元)	置装费(日元)
A 小姐	3000	7000
B 小姐	5000	8000
C 小姐	12000	25000
D 小姐	2000	5000
E 小姐	7000	12000
F 小姐	15000	30000
G 小姐	5000	10000
H 小姐	6000	15000
I 小姐	8000	20000
J 小姐	10000	18000

首先，试着画成图表吧！
好的。
一个月的“化妆品费”和“置装费”之散布图
置装费
30000
20000
10000
0
0
10000
20000
30000
化妆品费
哦！
看来似乎花较多钱在化妆品上的人也会花较多的钱买衣服啊！
那么，我们试着求出两者的关联“程度”吧！

	指标	值的范围	计算式
数值数据和数值数据	相关系数[1]	−1～1	$\dfrac{x\text{和}y\text{的共变异数}^{[2]}}{\sqrt{x\text{的变异数}^{[3]} \times y\text{的变异数}}} = \dfrac{Sxy}{\sqrt{Sxx \times Syy}}$
数值数据和分类数据	相关比	0～1	$\dfrac{\text{组间变异}}{\text{组内变异} + \text{组间变异}}$ →参见P121“2.相关比”
分类数据和分类数据	克莱姆相关系数	0～1	$\sqrt{\dfrac{\chi_0^2}{\text{数据个数} \times (\min\{\text{交叉资料表的行数，交叉资料表的列数}\}-1)}}$ →参见P127“3.克莱姆相关系数”



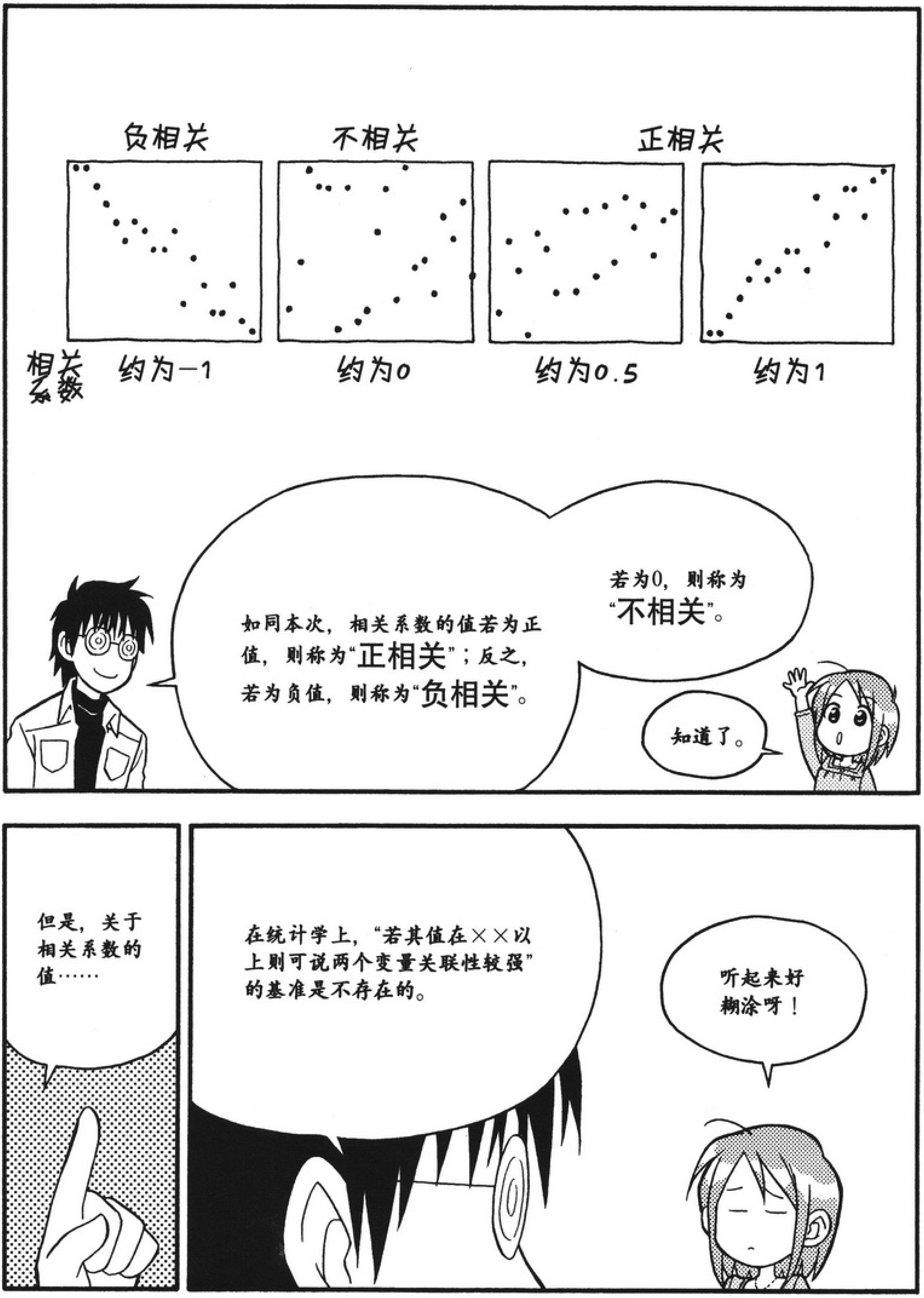


1个月的“化妆品费”和“置装费”之相关系数计算过程

	化装品费	置装费					
	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
A小姐	3000	7000	-4300	-8000	18490000	64000000	34400000
B小姐	5000	8000	-2300	-7000	5290000	49000000	16100000
C小姐	12000	25000	4700	10000	22090000	100000000	47000000
D小姐	2000	5000	-5300	-10000	28090000	100000000	53000000
E小姐	7000	12000	-300	-3000	90000	9000000	900000
F小姐	15000	30000	7700	15000	59290000	225000000	115500000
G小姐	5000	10000	-2300	-5000	5290000	25000000	11500000
H小姐	6000	15000	-1300	0	1690000	0	0
I小姐	8000	20000	700	5000	490000	25000000	3500000
J小姐	10000	18000	2700	3000	7290000	9000000	8100000
合计	73000	150000	0	0	148100000	606000000	290000000
平均数	7300	15000			↓ Sxx	↓ Syy	↓ Sxy
	↓ $\bar{x}$	↓ $\bar{y}$					

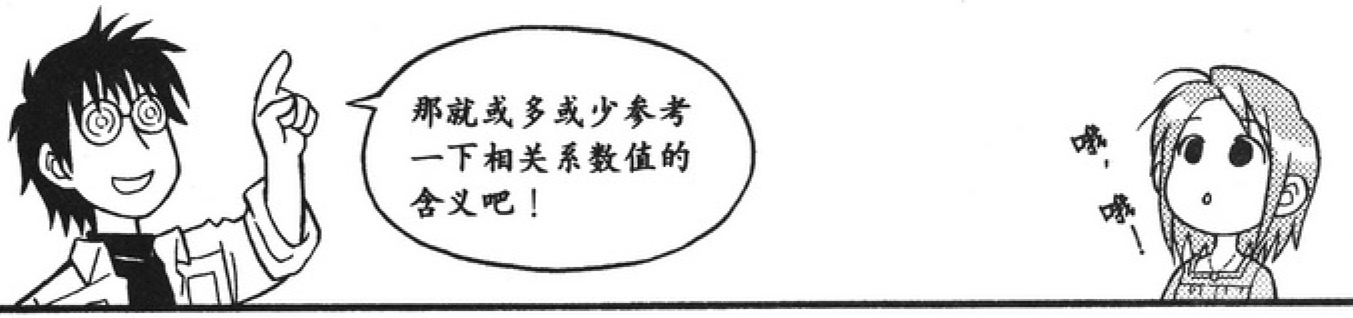
1. 相关系数：Correlation Coefficient。 2. 共变异数：Covariance。 3. 变异数：Variance。


接下来,代入公式吧。
Sxy / √(Sxx×Syy) = 290000000 / √(148100000×606000000) = 0.9680
用电脑就可以马上求出。
相关系数的值是0.9680！
而且，若两个变量的相关性越强，则相关系数就会越接近±1。
如果关联性越弱，相关系数则会越接近0。
-1 ← 0 → +1
嗯。
由于这个结果相当接近1，所以“化妆品费”和“置装费”的关联性非常强。
P-girls.
是啊！
那什么情况下接近-1呢？
“化妆品费”越高，而“置装费”越低的情况。


负相关
不相关
正相关
相关系数
约为−1
约为0
约为0.5
约为1
如同本次，相关系数的值若为正值，则称为“正相关”；反之，若为负值，则称为“负相关”。
若为0，则称为“不相关”。
知道了。
但是，关于相关系数的值……
在统计学上，“若其值在××以上则可说两个变量关联性较强”的基准是不存在的。
听起来好糊涂呀！

相关系数值之意义

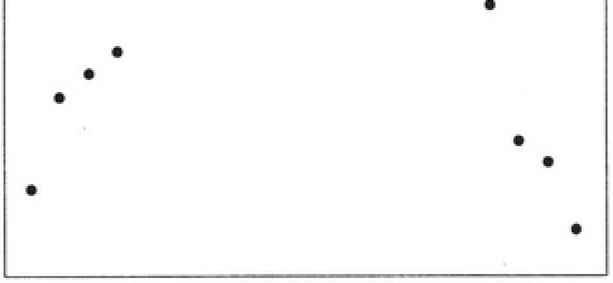
相关系数的绝对值		若细分……	若大体上划分……
1.0 ~ 0.9	⇒	相关性非常强	相关
0.9 ~ 0.7	⇒	相关性有点强	
0.7 ~ 0.5	⇒	相关性有点弱	
未满0.5	⇒	相关性非常弱	不相关



注意要点

之前说过，相关系数为表示数值数据与数值数据的关联性程度的指标。不过，严格说来并非如此。相关系数为清楚表示数值数据与数值数据之间是否具有“直线性”关联的指标。

不适用于相关系数的例子



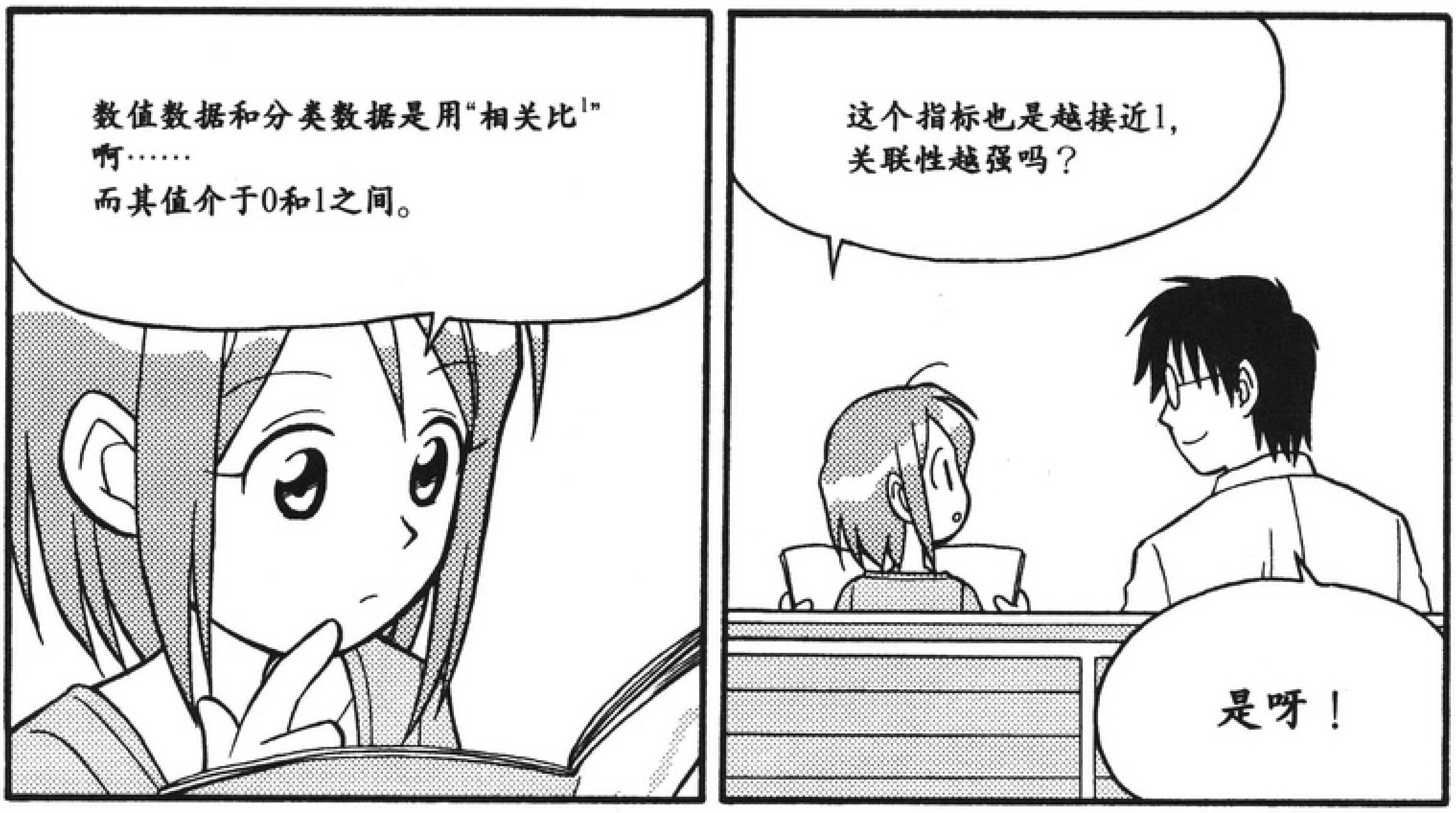
如同左图所示，可看出这两个变量具有明确的相关性。然而，由于其关联性为“曲线”的状态，因此相关系数的值接近于0。

✿2.相关比✿

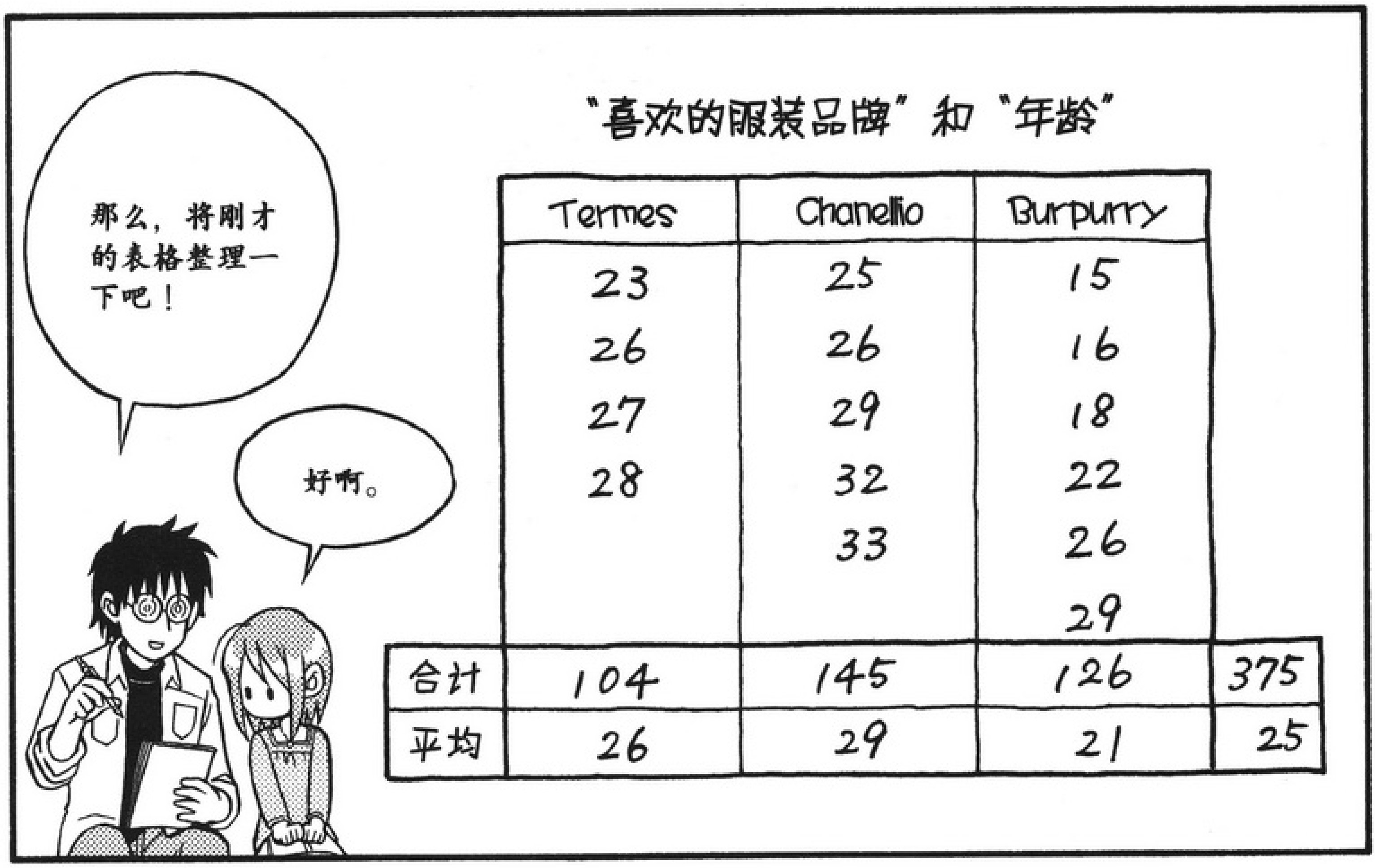
在五本木之丘所做的调查

“年龄”和“喜欢的服装品牌”

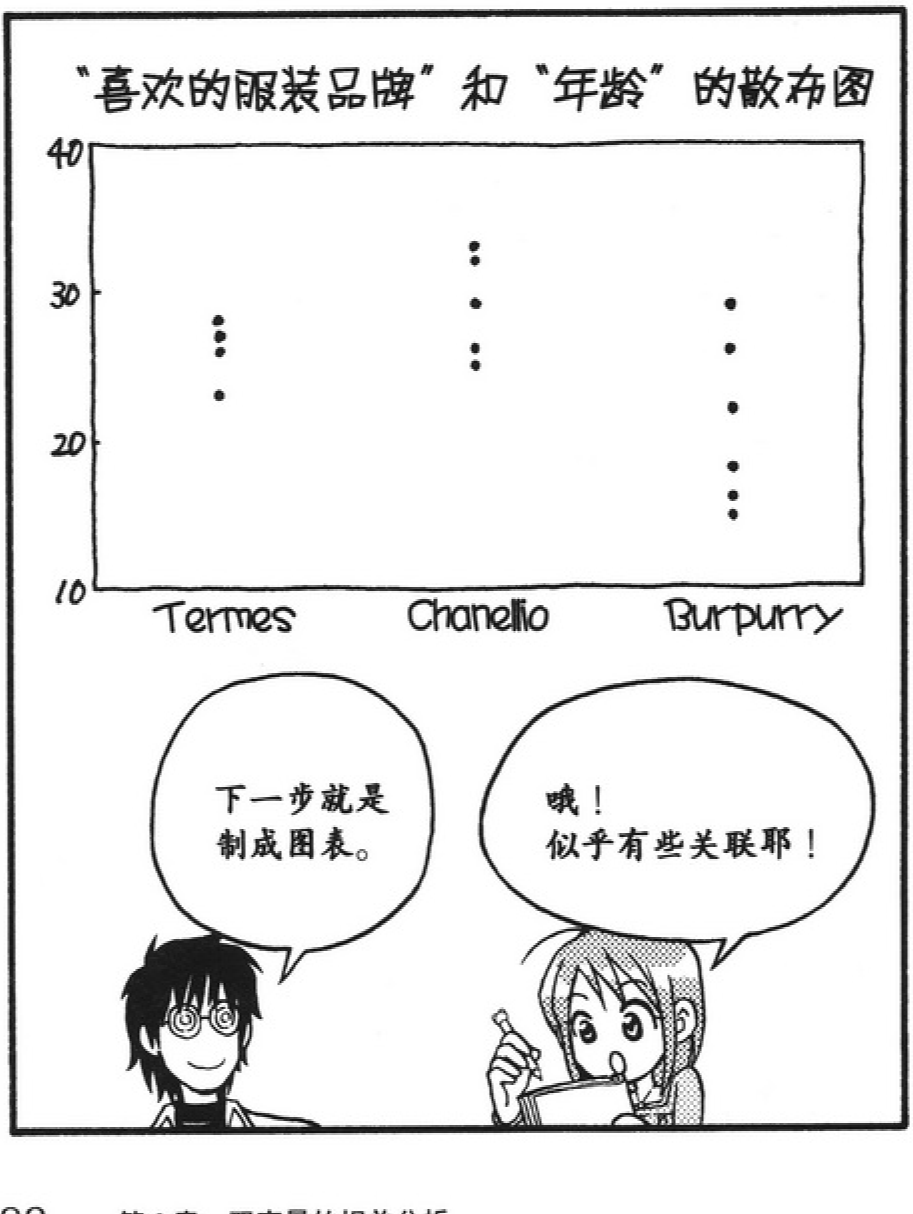
	年龄	品牌
A 小姐	27	Termes
B 小姐	33	Chanellio
C 小姐	16	Burpurry
D 小姐	29	Burpurry
E 小姐	32	Chanellio
F 小姐	23	Termes
G 小姐	25	Chanellio
H 小姐	28	Termes
I 小姐	22	Burpurry
J 小姐	18	Burpurry
K 小姐	26	Chanellio
L 小姐	26	Termes
M 小姐	15	Burpurry
N 小姐	29	Chanellio
O 小姐	26	Burpurry

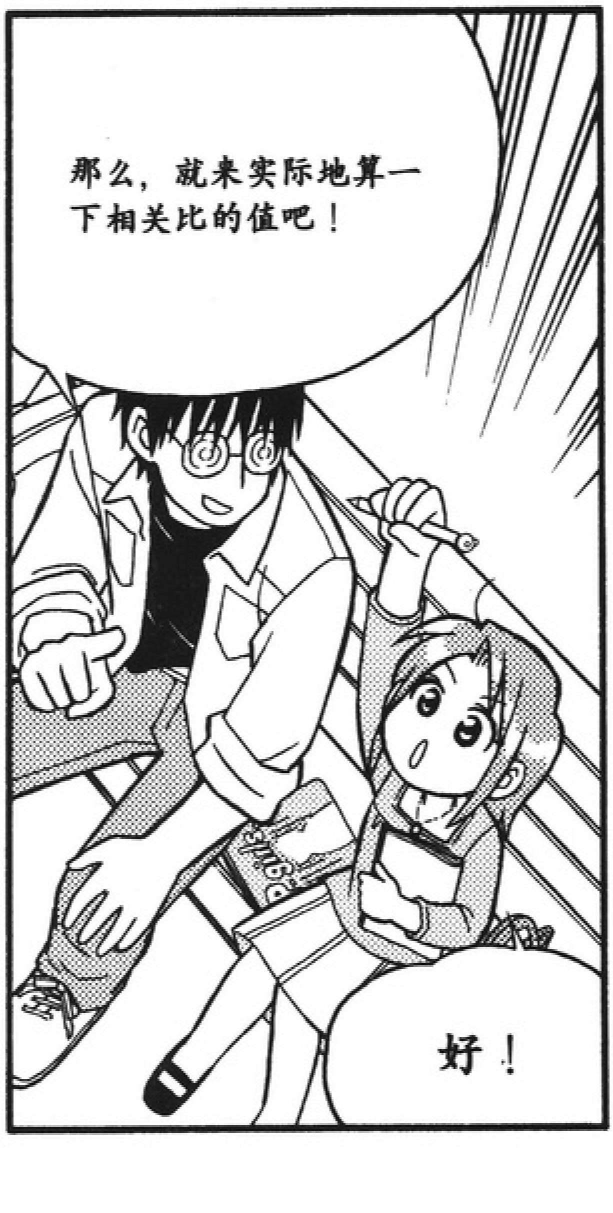


1. 相关比：Correlation Ratio。



	Termes	Chanellio	Burpurry	
	23	25	15	
	26	26	16	
	27	29	18	
	28	32	22	
		33	26	
			29	
合计	104	145	126	375
平均	26	29	21	25





相关比的值，只要依照以下的步骤1到步骤4的计算，就可以求出。

步骤1

进行如下表的计算。

	（Termes－Termes的平均值）^{2}	（Chanellio－Chanellio的平均值）^{2}	（Burpurry－Burpurry的平均值）^{2}
	$(23-26)^2=(-3)^2=9$	$(25-29)^2=(-4)^2=16$	$(15-21)^2=(-6)^2=36$
	$(26-26)^2=0^2=0$	$(26-29)^2=(-3)^2=9$	$(16-21)^2=(-5)^2=25$
	$(27-26)^2=1^2=1$	$(29-29)^2=0^2=0$	$(18-21)^2=(-3)^2=9$
	$(28-26)^2=2^2=4$	$(32-29)^2=3^2=9$	$(22-21)^2=1^2=1$
		$(33-29)^2=4^2=16$	$(26-21)^2=5^2=25$
			$(29-21)^2=8^2=64$
合计	14	50	160
	↓ S_{TT}	↓ S_{CC}	↓ S_{BB}

步骤2

求出组内变异，也就是$S_{TT}+S_{CC}+S_{BB}$

$$\boxed{S_{TT}+S_{CC}+S_{BB}=14+50+160=224}$$

步骤3

组间变异，也就是求出:

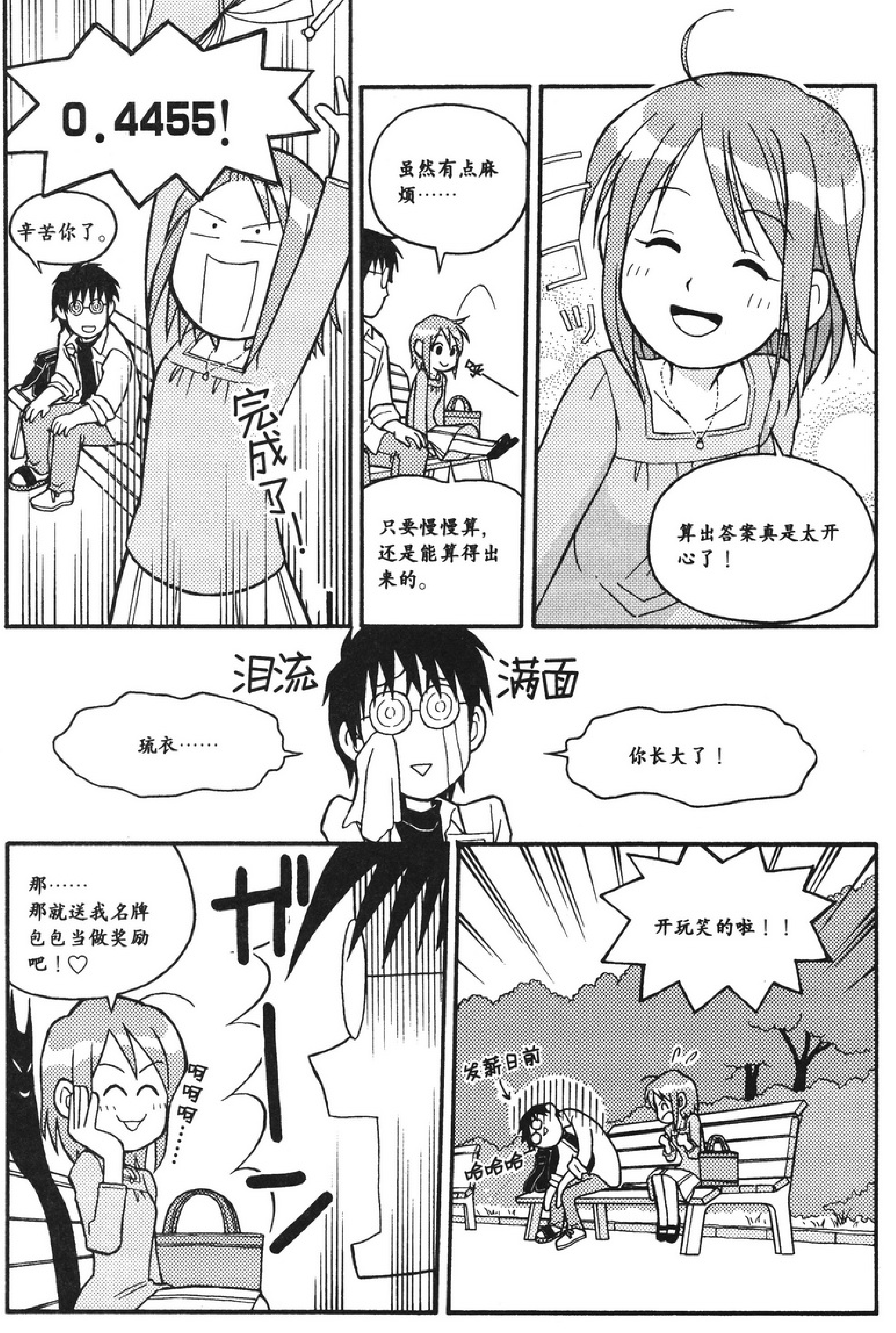
（Termes的数据个数）×（Termes的平均值–整体平均值）^{2}+（Chanellio的数据个数）×（Chanellio的平均值–全体平均值）^{2}+（Burpurry的数据个数）×（Burpurry的平均值）^{2}

$$
\begin{aligned}
&4\times(26-25)^2+5\times(29-25)^2+6\times(21-25)^2\\
&=4\times1+5\times16+6\times16\\
&=4+80+96\\
&=180
\end{aligned}
$$

步骤4

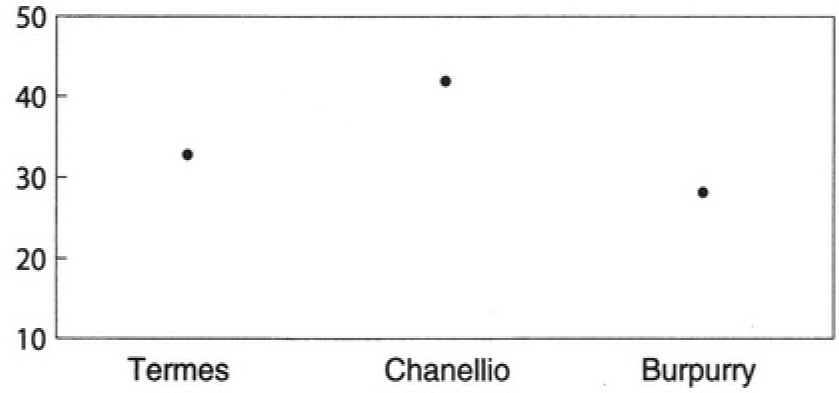
相关比的值，也就是求出 $\dfrac{\text{级间变异}}{\text{级内变异}+\text{级间变异}}$。

$$\frac{180}{224+180}=\frac{180}{404}=0.4455$$


0.4455！
辛苦你了。
完成了！
虽然有点麻烦……
只要慢慢算，还是能算得出来的。
算出答案真是太开心了！
泪流满面
琉衣……
你长大了！
那……那就送我名牌包包当做奖励吧！♡
呵呵呵……
开玩笑的啦！！
哈哈哈

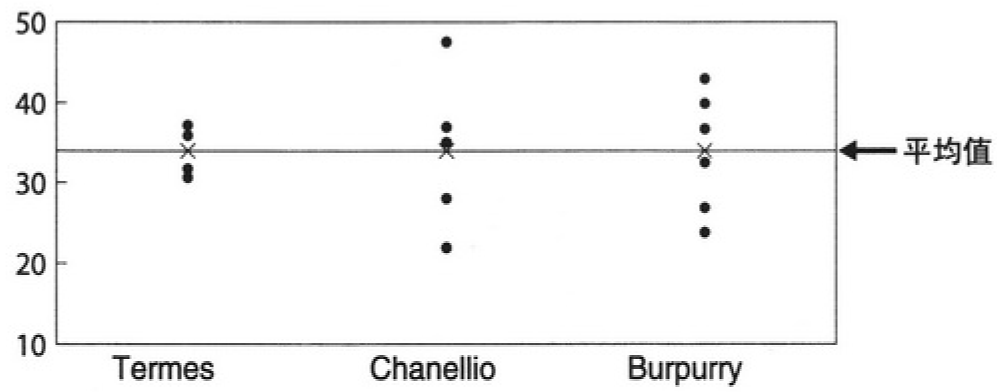
如前所述，相关比的值之范围，介于0和1之间。两个变量的关联性越强，则此值就会越接近1，反之，则会越接近0。详细内容请参照下图。

“喜欢的服装品牌”和“年龄”的散布图（相关比的值为1）。



相关比的值为1 ⇔ 各组所含数据相同 ⇔ 组内变异为0

“喜欢的服装品牌”和“年龄”的散布图（相关比的值为0）。



相关比的值为0 ⇔ 各组的平均值相同 ⇔ 组间变异为0

“相关比的值在XX以上，则可说两变量关联性强”这类标准，在统计学上是不存在的。
请参考下列相关比的值之意义。

相关比的值之意义

相关比的值		若细分……	若大略上划分……
1.0~0.8	⇨	相关性非常强	相关
0.8~0.5	⇨	相关性有点强	相关
0.5~0.25	⇨	相关性有点弱	相关
未满0.25	⇨	相关性非常弱	不相关

那么，由于本次的结果是0.4455，因此意思是“相关性有点弱”。

✿3.克莱姆相关系数✿

“性别”和“希望的表白方式”的交叉资料表

		希望的表白方法			合计
		打电话	发短信	当面	
性别	女性	34	61	53	148
	男性	38	40	74	152
合计		72	101	127	300

希望当面被表白的男性回答者，在152人中占了74人。

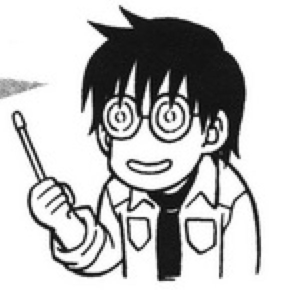
“性别”和“希望的表白方式”的交叉资料表(行%)

		希望的表白方法			合计
		打电话	发短信	当面	
性别	女性	23	41	36	100
	男性	25	26	49	100
合计		24	34	42	100

希望当面表白的男性回答者，在152人中，占了74/152×100=49（%）。

1. 克莱姆相关系数：Cramer’s V。

克莱姆相关系数的值,可用下列步骤1到步骤5的计算方式来求出。



步骤1

准备交叉资料表。此外，粗框内的各个数值，称为观测次数[1]。

		希望的表白方式			合计
		打电话	发短信	当面	
性别	女性	34	61	53	148
	男性	38	40	74	152
合计		72	101	127	300

步骤2

进行下表的计算。此外，粗框内的各个数值，称为期望次数[2]。

		希望的表白方式			合计
		打电话	短信	当面	
性别	女性	$\frac{148\times72}{300}$	$\frac{148\times101}{300}$	$\frac{148\times127}{300}$	148
	男性	$\frac{152\times72}{300}$	$\frac{152\times101}{300}$	$\frac{152\times127}{300}$	152
合计		72	101	127	300

$\dfrac{\text{“男性”的合计}\times\text{“当面”的合计}}{\text{数据个数}}$[2]

1. 观测次数：Observed Frequency。 2. 期望次数：Expected Frequency。

如果“性别”和“希望的表白方式”完全不相关，则打电话：发短信：当面的比值，无论是女性或男性都会根据步骤2的表中的“合计”得出以下比例：

$$72:101:127=\frac{72}{72+101+127}:\frac{101}{72+101+127}:\frac{127}{72+101+127}$$
$$=\frac{72}{300}:\frac{101}{300}:\frac{127}{300}$$

换句话说，表示当“性别”与“希望的表白方式”完全不相关时的“希望当面表白的男性人数”为

$$152\times\frac{127}{300}=\frac{152\times127}{300}$$

步骤3

每笔数值以 $\frac{(\text{观测次数}-\text{期望次数})^2}{\text{期望次数}}$ 来计算。

		希望的表白方式			合计
		打电话	发短信	当面	
性别	女性	$\frac{\left(34-\frac{148\times72}{300}\right)^2}{\frac{148\times72}{300}}$	$\frac{\left(61-\frac{148\times101}{300}\right)^2}{\frac{148\times101}{300}}$	$\frac{\left(53-\frac{148\times127}{300}\right)^2}{\frac{148\times127}{300}}$	148
	男性	$\frac{\left(38-\frac{152\times72}{300}\right)^2}{\frac{152\times72}{300}}$	$\frac{\left(40-\frac{152\times101}{300}\right)^2}{\frac{152\times101}{300}}$	$\frac{\left(74-\frac{152\times127}{300}\right)^2}{\frac{152\times127}{300}}$	152
合 计		72	101	127	300

观测次数和期望次数的差异越大，意即“性别”和“希望的表白方式”之间的关联程度越强，则粗框内的各数值也会越大。

步骤4

求出步骤3的表中粗框内的值之总和，意即皮尔森的卡方统计量之值。此外，皮尔森的卡方统计量，以下用“χ_0^2”表示。

$$
\begin{aligned}
\chi_0^2 = {} & \frac{\left(34-\dfrac{148\times72}{300}\right)^2}{\dfrac{148\times72}{300}}+\frac{\left(61-\dfrac{148\times101}{300}\right)^2}{\dfrac{148\times101}{300}}+\frac{\left(53-\dfrac{148\times127}{300}\right)^2}{\dfrac{148\times127}{300}} \\
& +\frac{\left(38-\dfrac{152\times72}{300}\right)^2}{\dfrac{152\times72}{300}}+\frac{\left(40-\dfrac{152\times101}{300}\right)^2}{\dfrac{152\times101}{300}}+\frac{\left(74-\dfrac{152\times127}{300}\right)^2}{\dfrac{152\times127}{300}} \\
= {} & 8.0091
\end{aligned}
$$

如同步骤3中的说明，观测次数和期望次数的差异越大，意即“性别”和“希望的表白方式”之间的关联程度越强，则皮尔森的卡方统计量 χ_0^2 也会越大。

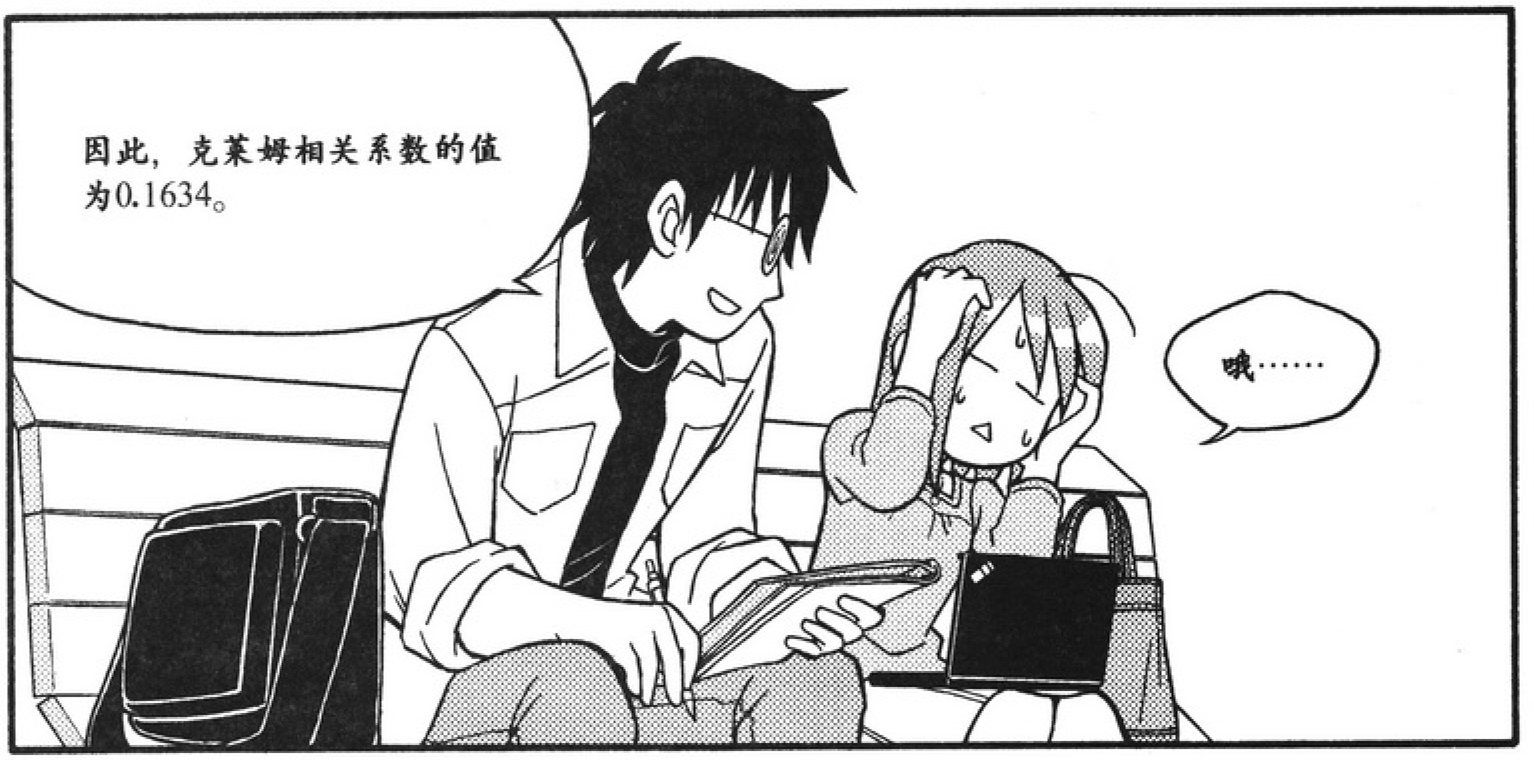
步骤5

求出克莱姆相关系数的值，即

$$\sqrt{\frac{\chi_0^2}{\text{数据个数}\times(\min\{\text{交叉资料表的行数},\text{交叉资料表的列数}\}-1)}}$$

此外，min{a，b}为表示a和b中较小的值之记号。

$$\sqrt{\frac{8.0091}{300\times(\min\{2,3\}-1)}}=\sqrt{\frac{8.0091}{300\times(2-1)}}=\sqrt{\frac{8.0091}{300}}=0.1634$$




一个人根本没办法计算嘛……
一个人也没问题的！计算确实很繁琐，但如果照程序来做，一定可以算出来的。
如果还有不懂的地方，来问我就可以了喔
！
吃惊！
啊？
扭头
怎么了呀？
没事。
揉眼
一不小心，把山本老师看成帅哥了。

先前谈过的，克莱姆相关系数的值介于0和1之间，两个变量的关联性越强，则此值就会越接近1，反之，则会越接近0。详细情形请参照下面的交叉资料表（行%）。

"性别"和"希望的表白方式"的交叉资料表

（克莱姆相关系数为1）

		希望的表白方式			合计
		打电话	发短信	直接见面	
性别	女性	17	83	0	100
	男性	0	0	100	100

克莱姆相关系数的值为1 ⟺ 女性和男性的喜好完全不同

"性别"和"希望的表白方式"的交叉资料表

（克莱姆相关系数为0）

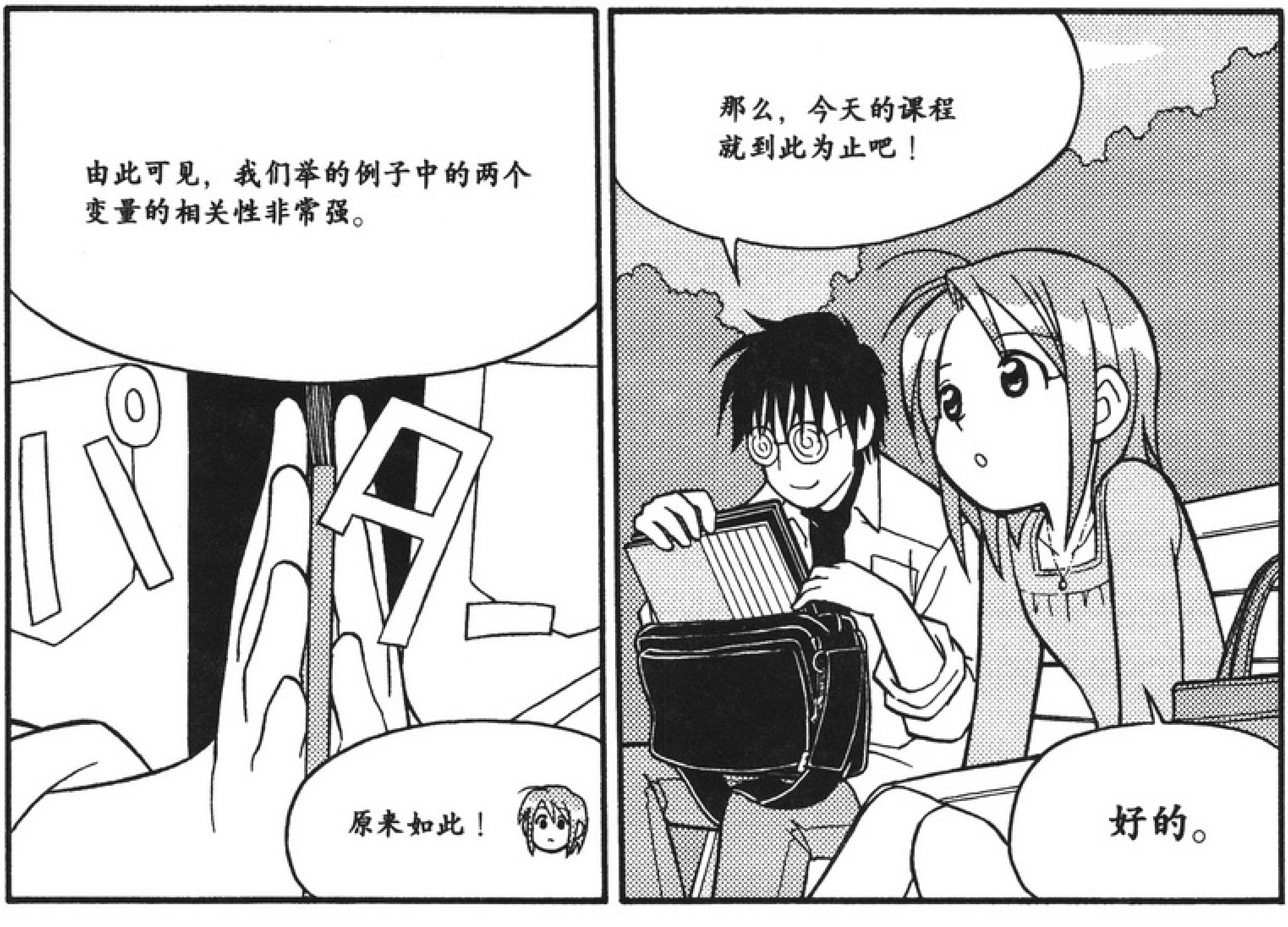
		希望的表白方式			合计
		打电话	发短信	直接见面	
性别	女性	17	48	35	100
	男性	17	48	35	100

克莱姆相关系数的值为0 ⟺ 女性和男性的喜好完全相同

“若克莱姆相关系数的值在XX以上，则可说两个变量的关联性较强”，在统计学上并不存在这个基准。请参考下面为克莱姆相关系数的值之意义。

克莱姆相关系数的值之意义

克莱姆相关系数的值		若细分……	若大略上划分……
1.0 ~ 0.8	⇨	相关性非常强	相关
0.8 ~ 0.5	⇨	相关性有点强	
0.5 ~ 0.25	⇨	相关性有点弱	
未满0.25	⇨	相关性非常弱	不相关



1. 独立性检验：Test of Independence。

例题和解答

例题

经营家庭餐馆的A公司，最近经营状况并不太好。因此必须用心倾听顾客的声音，所以针对“居住在日本的20岁以上居民”以随机抽样进行问卷调查。结果如下表所示。

	…	你在家庭餐馆常点哪类料理？	…	若附免费的餐后饮料，咖啡和红茶哪一种比较好？	…
回答者1	…	中式料理	…	咖啡	…
回答者2	…	西式料理	…	咖啡	…
:	…	⋮	…	⋮	…
回答者250	…	日式料理	…	红茶	…

用上表做成的交叉资料表如下所示。

		咖啡和红茶哪一种比较好?		合 计
		咖啡	红茶	
常点的料理种类	日式料理	43	33	76
	西式料理	51	53	104
	中式料理	29	41	70
合 计		123	127	250

请求出“在家庭餐馆常点的料理种类是？”和“若附免费的餐后饮料，咖啡和红茶哪一种比较好？”的克莱姆相关系数值。

解答

步骤1

准备交叉资料表。

		咖啡和红茶哪一种比较好？		合计
		咖啡	红茶	
常点的料理种类	日式料理	43	33	76
	西式料理	51	53	104
	中式料理	29	41	70
合计		123	127	250

步骤2

求出期望次数。

		咖啡和红茶哪一种比较好？		合计
		咖啡	红茶	
常点的料理种类	日式料理	$\frac{76\times123}{250}$	$\frac{76\times127}{250}$	76
	西式料理	$\frac{104\times123}{250}$	$\frac{104\times127}{250}$	104
	中式料理	$\frac{70\times123}{250}$	$\frac{70\times127}{250}$	70
合计		123	127	250

步骤3

计算出各个表格里的 $\frac{(\text{观测次数}-\text{期望次数})^2}{\text{期望次数}}$。

		咖啡和红茶哪一种比较好		合计
		咖啡	红茶	
常点的料理种类	日式料理	$\frac{\left(43-\frac{76\times123}{250}\right)^2}{\frac{76\times123}{250}}$	$\frac{\left(33-\frac{76\times127}{250}\right)^2}{\frac{76\times127}{250}}$	76
	西式料理	$\frac{\left(51-\frac{104\times123}{250}\right)^2}{\frac{104\times123}{250}}$	$\frac{\left(53-\frac{104\times127}{250}\right)^2}{\frac{104\times127}{250}}$	104
	中式料理	$\frac{\left(29-\frac{70\times123}{250}\right)^2}{\frac{70\times123}{250}}$	$\frac{\left(41-\frac{70\times127}{250}\right)^2}{\frac{70\times127}{250}}$	70
合计		123	127	250

步骤4

求出步骤3的表中粗框内的值之总和，意即皮尔森的卡方统计量χ_0^2之值。

$$\chi_0^2=\frac{\left(43-\frac{76\times123}{250}\right)^2}{\frac{76\times123}{250}}+\frac{\left(33-\frac{76\times127}{250}\right)^2}{\frac{76\times127}{250}}$$

$$+\frac{\left(51-\frac{104\times123}{250}\right)^2}{\frac{104\times123}{250}}+\frac{\left(53-\frac{104\times127}{250}\right)^2}{\frac{104\times127}{250}}$$

$$+\frac{\left(29-\frac{70\times123}{250}\right)^2}{\frac{70\times123}{250}}+\frac{\left(41-\frac{70\times127}{250}\right)^2}{\frac{70\times127}{250}}$$

$$=3.3483$$

步骤5

求出克莱姆相关系数的值，即

$$\sqrt{\frac{\chi_0^2}{\text{数据个数}\times(\min\{\text{交叉资料表的行数},\text{交叉资料表的列数}\}-1)}}$$

$$\sqrt{\frac{3.3483}{250\times(\min\{3,2\}-1)}}=\sqrt{\frac{3.3483}{250\times(2-1)}}=\sqrt{\frac{3.3483}{250}}=0.1157$$

总整理

• 相关系数为表示数值数据和数值数据的关联程度之指标。

• 相关比为表示数值数据和分类数据的关联程度之指标。

• 克莱姆相关系数（也可称作克莱姆关联系数或克莱姆V）为表示分类数据和分类数据的相关程度之指标。

• 相关系数、相关比和克莱姆相关系数的特征如下表所示。

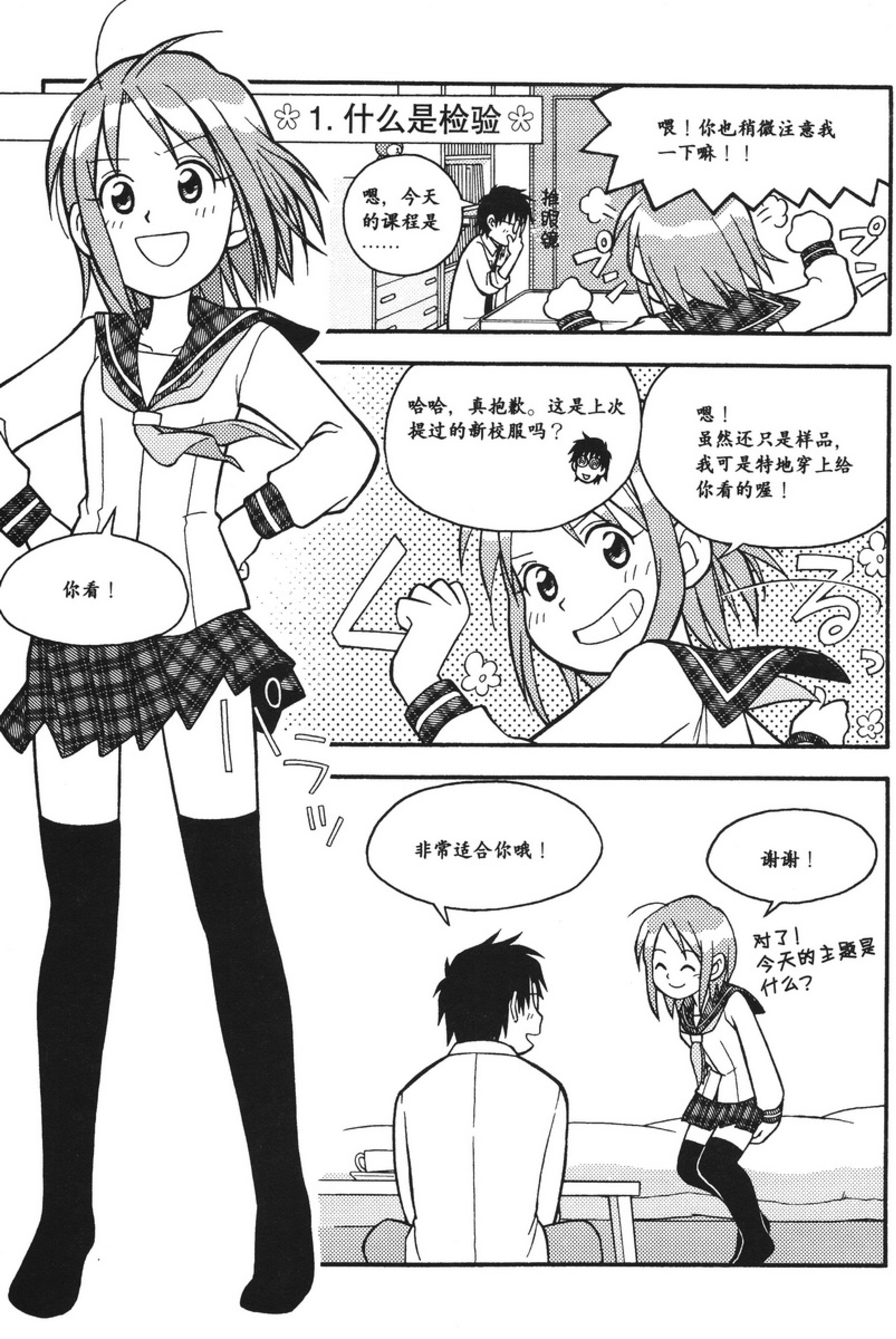
	最小值	最大值	两变量完全不相关时的值	两变量相关性最强时的值
相关系数	−1	1	0	−1或1
相关比	0	1	0	1
克莱姆相关系数	0	1	0	1

• 相关系数、相关比和克莱姆相关系数中，在统计学上，并无“其值若在XX以上时，则两变量的关联性较强”的标准。

◆第7章◆

深入理解独立性检验

1. 什么是检验

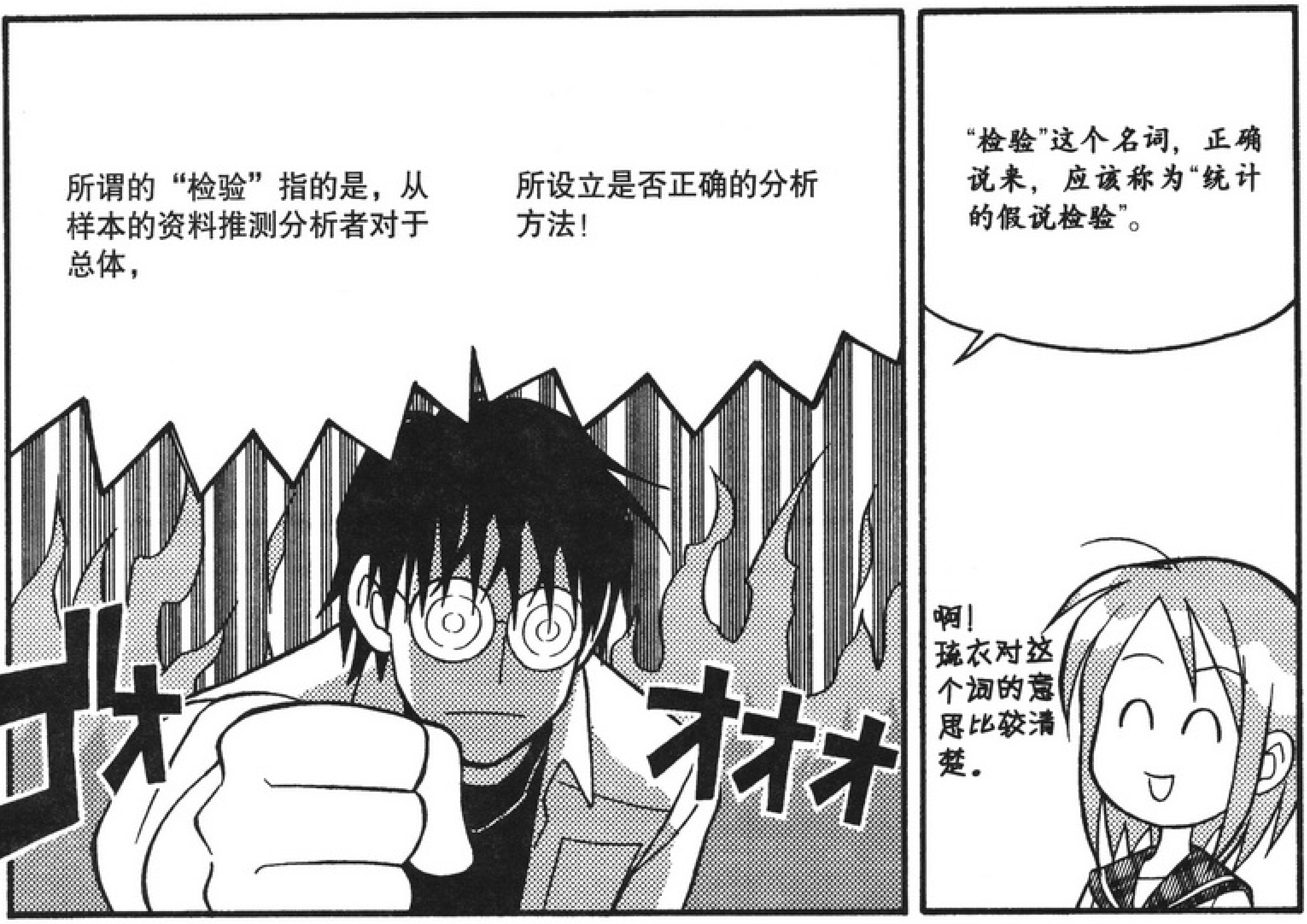

嗯，今天的课程是……
推眼镜
喂！你也稍微注意我一下嘛！！
你看！
哈哈，真抱歉。这是上次提过的新校服吗？
嗯！虽然还只是样品，我可是特地穿上给你看的喔！
非常适合你哦！
谢谢！
对了！今天的主题是什么？

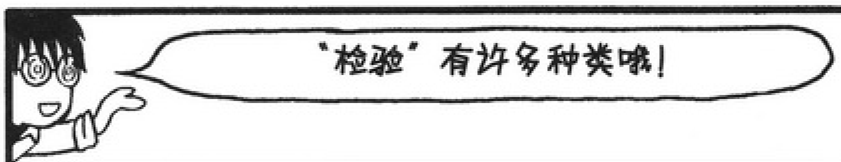

嗯……
上次的课程中，我们学习了克莱姆相关系数，对吧！
征询300名高中生
你希望对方用什么样的方式向你表白？
表白的话题嘛。
那个例子的克莱姆相关系数值是0.1634。
结论是——“相关性非常弱”。
是呀。
那么，请你仔细想想。
那份问卷调查是从“居住在日本的全体高中生”中随机抽样的结果，
只不过是300人的资料所得的推论结果。
如果再抽样调查另外的300人，
克莱姆相关系数的值应该绝对不会是0.1634吧！
这么说来，的确是这样……


你认为原始的总体——“居住在日本的全体高中生”的克莱姆相关系数是多少呢？
嗯……不晓得耶！
没错。如果没有收集到“居住在日本的全体高中生”的资料的话，
很抱歉！任何人都无从得知。
是呀。
因此，我们只能……
不仅限于那个例子，一般而言，我们是不可能知道总体的克莱姆相关系数的值。

对于总体的克莱姆相关系数做出……
“因为从随机抽出300人的资料中，所推论的克莱姆相关系数为0.1634，
所以总体的克莱姆相关系数大约为这个数值。
0.1634
也只能这样主观地判断了。
相当的模糊啊……
不过利用统计学，或许可以做些什么吧？
灵机一动
不！即使运用统计学，很可惜地，我们还是无法严谨地得知克莱姆相关系数的值。
啊，是这样吗？


光芒四射
但是，“总体的克莱姆相关系数的值——
究竟是否为0”，是可以知道的！
这很厉害吗？
究竟是否为
那当然啦！因为可以得到客观的总体信息呀！
那么，该怎么做呢？
只要使用之前提过的名为“独立性检验”的分析方法即可。
是类似英检的东西吗？
哈哈……不！完全不同喔！
独立性检验
检验
独立性检验是统计学上总称为“检验”的分析方法之一。
首先就什么是“检验”做一下说明吧！
无相关检验
相关比检验
总体比例
好的。





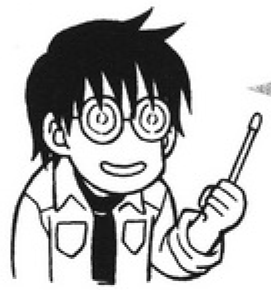
“检验”的实例	
名　称	可使用的情况之实例
独立性检验	推测总体中，“性别”和“希望的表白方式”的克莱姆相关系数之值是否为0。
相关比检验	推测总体中，“喜欢的服装品牌”和“年龄”的相关比之值是否为0。
无相关检验	推测总体中，“1个月使用的化妆品费用”和“1个月使用的置装费”的相关系数之值是否为0。
总体平均数差检验	推测东京都的女高中生和大阪府的女高中生“每月的零花钱”是否不同。 ※注意，这个例子中设定了两个总体。
总体比例差检验	推测居住于都市的有投票权者和居住于农村的有投票权者中，对“××内阁的支持率”是否不同。 ※注意，这个例子中设定了两个总体。

“检验”的程序	
【步骤1】	定义总体。
【步骤2】	建立虚无假说[1]和对立假说[2]。
【步骤3】	选择要进行的“检验”种类。
【步骤4】	决定置信水平[3]。
【步骤5】	从样本资料求出检验统计量的值。
【步骤6】	调查【步骤5】所求出的检验统计量值，是否在拒绝域*4之中。
【步骤7】	若【步骤6】的检验统计量在拒绝域之中，则结论为“对立假说正确”。反之，则结论为“无法判定虚无假说为误”。

1. 虚无假说：Null Hypothesis。 2. 对立假说：Alternative Hypothesis。 3. 置信水平：Confidence level。 4. 拒绝域：Rejection Region。


2.独立性检验
那么，现在开始讲今天的主题“独立性检验”。
所谓的“独立性检验”指的是，推测“总体的克莱姆相关系数的值究竟是否为0”的分析方法。
了解。
换句话说，就是推测“交叉资料表中的两变量是否相关”的分析方法。
原来如此，那就是问卷调查分析啦！
希望的表白方式
打电话 发短信 当面
合计
性别
女性 34 61 53 148
男性 38 40 74 152
合计 72 101 127
独立性检验也可称为“卡方检验”哦！
又来了！真麻烦！

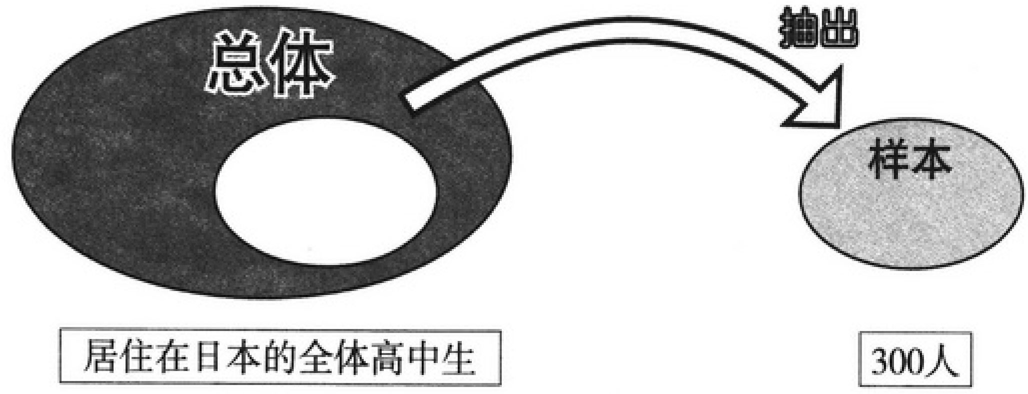
解说　皮尔森卡方统计量[1] χ_0^2 和卡方分布



在开始解说独立性检验的实例前，先为各位解说独立性检验基础的重要事实。虽然现实中是不可能成立的，但我们假设以下的实验已经完成。

步骤1

从总体“居住在日本的全体高中生”中随机抽取300人。



步骤2

对步骤1中抽出的300人进行127页的问卷调查，以求出皮尔森卡方统计量 χ_0^2。

步骤3

将随机抽出的300人送回总体。

步骤4

持续重复步骤1~3。

如此一来，若做为总体的“居住在日本的全体高中生”中，其克莱姆相关系数为0，则实验中皮尔森卡方统计量 χ_0^2 之图形为自由度为2的卡方分布[2]。换句话说，若做为总体的“居住在日本的全体高中生”中，克莱姆相关系数为0，则“实验中的皮尔森卡方统计量 χ_0^2”服从自由度为2的卡方分布。

1. 皮尔森的卡方统计量 χ_0^2 的算法，请参照130~133页。
2. 自由度为2的卡方分布，请参照100页。

试着进行实际实验。请注意，在实验进行时，我们设定了以下的限制条件。

- 由于真正以“居住在日本的全体高中生”为对象的实验是不可能实现的，因此将表7.1中记载的1万人的集合，解释为“居住在日本的全体高中生”。
- 将“居住在日本的全体高中生”中的克莱姆相关系数设为0。意即，女性和男性在“想在电话中表白：想在短信中表白：想当面表白”的比例是相等的（请参照135页）。实际将表7.1的交叉资料表做成表7.2。
- 由于实验永无止境，因此重复步骤1~ 3的步骤20000次后就结束。

◆表7.1　希望的表白方式（居住在日本的全体高中生）

	性别	希望的表白方式
1	女	当面
2	女	打电话
⋮	⋮	⋮
10000	男	发短信

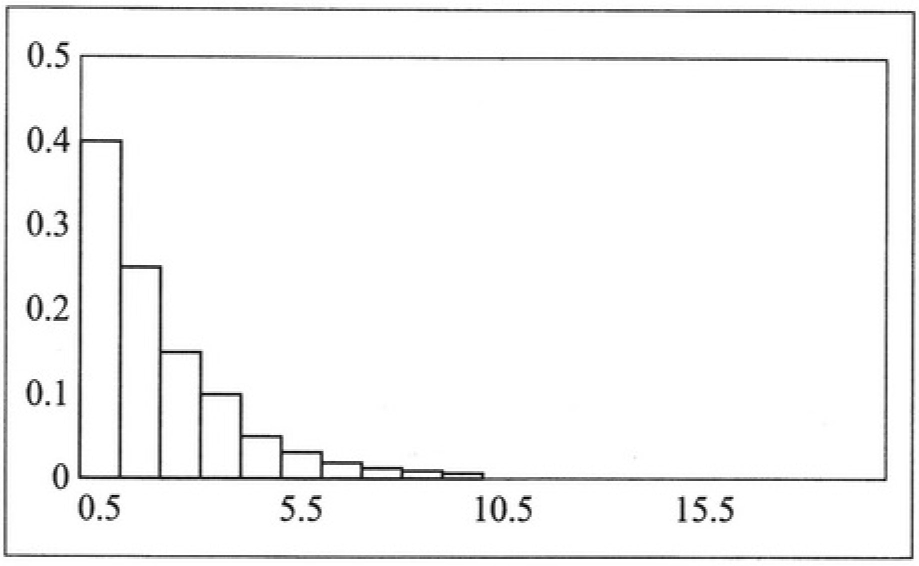
◆表7.2　“性别”和“希望的表白方式”之交叉资料表

		希望的表白方式			合 计
		打电话	发短信	当面	
性别	女性	400	1600	2000	4000
	男性	600	2400	3000	6000
合 计		1000	4000	5000	10000

实验结果如表7.3。图7.1是以表7.3为基准所绘出的直方图。

◆表7.3　实验结果

	皮尔森卡方统计量χ_0^2
第1次	0.8598
第2次	0.7557
⋮	⋮
第20000次	2.7953



◆图7.1　以表7.3为基准之直方图（组距为1）

图7.1确实和100页的“■自由度为2”时的图形非常相似。看来“皮尔森卡方统计量χ_0^2确实服从自由度为2的卡方分布。

虽然实验将就此结束了，但有一点必须注意。即，自由度的是由

$$(2-1)\times(3-1)=1\times2=2$$

“女性”“男性” 共两种时为2

“打电话”“发短信”“当面” 共三种时为3

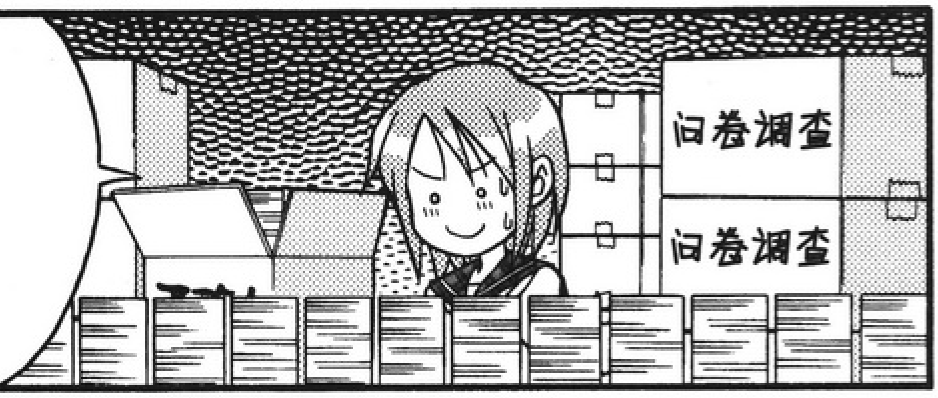
而来的。至于为何用这样不可思议的计算方式，由于这已经超出本书的讨论范围，因此就先略过。即使不了解这个计算方法的来龙去脉，在实务上并不会造成任何影响，所以请各位放心。

"居住在日本的全体高中生"的克莱姆相关系数的值为0……

意即"性别"和"希望的表白方式"并无关联。

那么，从"居住在日本的全体高中生"中选出300人进行问卷调查……

做了一次又一次……又一次！



求出皮尔森卡方统计量χ_0^2后……

这个图形就是自由度为2时的卡方分布呀！


琉衣又有所长进了呀！
哇哈哈！
那么，就以表白方式的问卷调查为例，
进行独立性检验吧！
遵命！
从现在起，我们的课程就以
“例题” → “思考” → “解答”
这样的方式来进行啦……
找找
“解答”
分析者
解说者
就由我一人分饰两个角色来说明吧！
分成“分析者”和“解说者”来说明，会让“解答”更易于理解哦！
琉衣应该更可爱啊！

例题

凛凛出版社将“询问300名高中生！你希望对方用什么样的方式向你表白？”的报道刊载于女性杂志“P-girls”中。凛凛出版社从“居住在日本的全体高中生”中，随机抽出300人，进行了问卷调查。其结果如下表所示。

	希望的表白方式	年龄	性别
回答者1	当面	17	女
回答者2	打电话	15	女
⋮	⋮	⋮	⋮
回答者300	发短信	18	男

然后，“性别”和“希望的表白方式”之交叉资料表如下。

		希望的表白方式			合 计
		打电话	发短信	当面	
性别	女性	34	61	53	148
	男性	38	40	74	152
合 计		72	101	127	300

总体“居住在日本的全体高中生”中，“性别”和“希望的表白方式”的克莱姆相关系数的值是否大于0，也就是“性别”和“希望的表白方式”是否有关联，请利用独立性检验来推测。此外，我们将置信水平（待后说明）设为0.05。

思考

如同152~154页中的解说，若总体“居住在日本的全体高中生”中的克莱姆相关系数为0，则“皮尔森卡方统计量χ_0^2”是服从自由度为2的卡方分布。因此，若总体“居住在日本的全体高中生”中的克莱姆相关系数的值为0，则由随机抽出的300人的资料所求出的χ_0^2若为5.9915以上的机率，则能从103页的卡方分布表中清楚得知，其值为0.05。

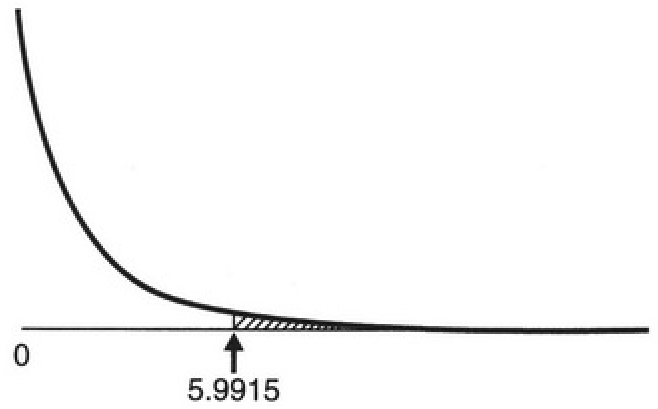


图7.2　χ_0^2为5.9915以上的机率

本例题的χ_0^2在132页就已计算完毕，其值为8.0091。怎么会这样呢？虽然是由随机抽出300人的资料所求出的值，看起来似乎还是太高了吧！若以132页的评论为基础来思考，总体“居住在日本的全体高中生”的克莱姆相关系数的值大于0的想法，是不是就很自然呢？

不仅限于这个例题，在说明独立性检验时，我会以

① 暂且解释为“总体的克莱姆相关系数的值为0”

② 由样本的资料求出χ_0^2

③ 若χ_0^2过大，则结论为“总体的克莱姆相关系数的值大于0”

这样的流程进行说明，请先记下来。

接下来，将为前一段落的③做补充。

χ_0^2越大，则下图斜线部分的机率理应越小。

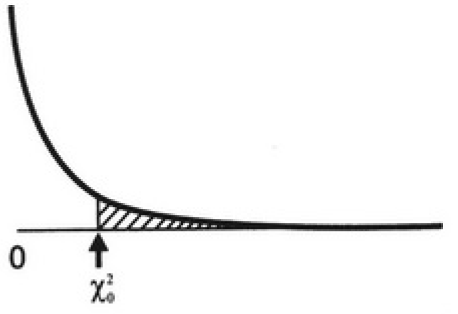


图7.3　对应χ_0^2的机率

独立性检验中，若上图斜线部分的机率在名为置信水平的值以下，则可做“总体的克莱姆相关系数的值大于0”的结论。置信水平一般设为0.05或0.01，采用何者则完全取决于分析者的判断。

现在假设采用0.05的置信水平。实际上，所谓的置信水平就是指下图斜线部分的机率。

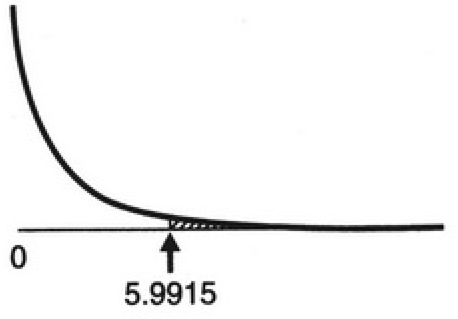


图7.4　再现图7.2（$=\chi_0^2$在5.9915以上的机率）

此外，下图的范围称为拒绝域。

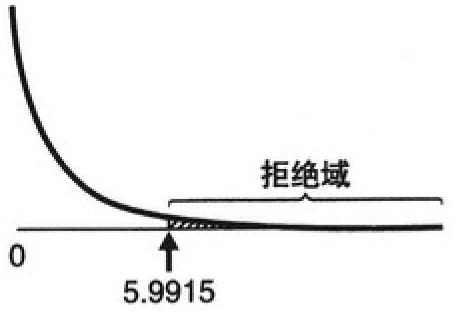


图7.5　（置信水平0.05时）拒绝域

❗解答

步骤1

定义总体。

总体

总体
=居住在日本的全体高中生

由于本例题中的总体一开始就定义为“居住在日本的全体高中生”。因此在本例题中，步骤1当然是不需要的。

举例来说，149页的表“总体比例差检验”中，设定“居住在都市的有选举权者”和“居住在农村的有选举权者”为总体。那么，“都市”具体上到底指哪里呢？“东京都和大阪府”吗？“各都道府县的地方政府所在地”吗？这是由分析者所决定。没错，实际上执行“检验”时，总体必须由分析者自行定义。

无论是何种“检验”，若没有清楚地定义总体，则易陷于“奇怪！我当初到底想推测什么？”的状况之中。陷于这种状况的分析者并不在少数。请各位务必特别注意这一点。

步骤2

建立虚无假说和对立假说。

虚无假说为

总体的克莱姆相关系数的值为0
="性别"和"希望的表白方式"不相关。

对立假说为

总体的克莱姆相关系数的值大于0=
"性别"和"希望的表白方式"相关。

关于虚无假说和对立假说，随后将进行讲解。

步骤3

选择进行的“检验”种类。

进行独立性检验。

本例题原先就设定为进行独立性检验。因此，本例题当然不需要步骤3。实际上，进行“检验”之际，分析者必须选择符合分析目的的“检验”。

步骤4

决定置信水平。

设定置信水平为0.05。

本例题原先就设定置信水平为0.05，因此，本例题也不需要步骤4。

实际进行“检验”之际，分析者必须自己决定置信水平。如同先前所述，置信水平一般会设为0.05或0.01。

置信水平通常以“α”这个符号来表示。

步骤5

从样本资料求出检验统计量的值。

我想做的是独立性检验，因此检验统计量为皮尔森的卡方统计量χ_0^2。本例题中的χ_0^2值已在132页计算完毕，χ_0^2=8.0091。

所谓的检验统计量，是指将样本资料转换成1个值的公式。

依所进行“检验”的种类不同，检验统计量也会有所不同。独立性检验的情况，则如同上述，采用χ_0^2，而无相关检验（请参照149页）的情况，则采用下述的值。

$$\frac{\text{相关系数}^2 \times \sqrt{\text{数据个数}-2}}{\sqrt{1-\text{相关系数}^2}}$$

步骤6

调查步骤5所求出的检验统计量值，是否在拒绝域中。

检验统计量–皮尔森卡方统计量χ_0^2的值为8.0091。

由于置信水平为0.05，因此，拒绝域根据103页的卡方分布表得知，其值为“5.9915以上”。

如下图所示，检验统计量的值在拒绝域之中。

拒绝域依置信水平α不同而变化。如果本例题中α不是0.05而为0.01时，则拒绝域根据103页的卡方分布表所示，其值为“9.2104以上”。

步骤7

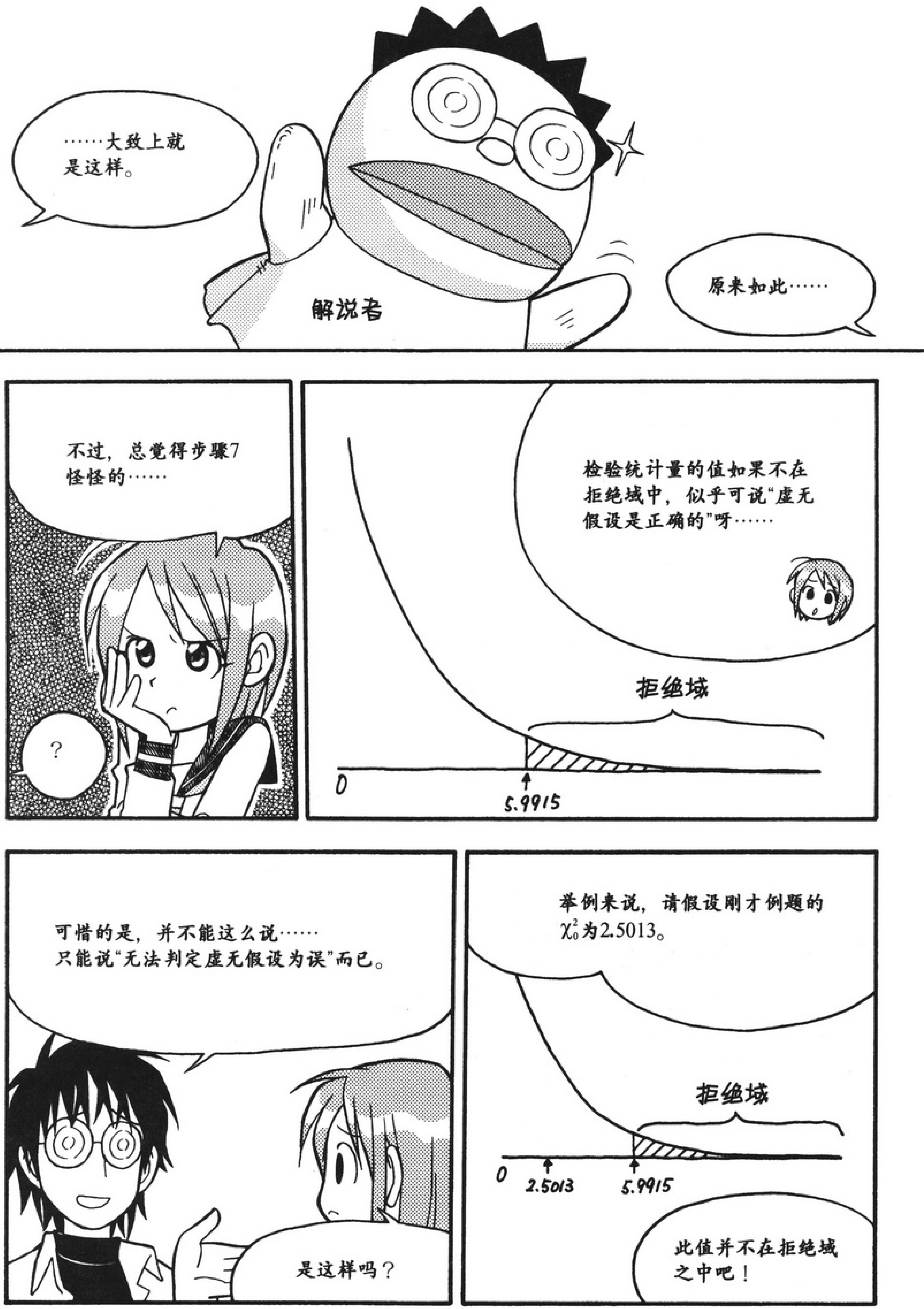
若步骤6的检验统计量值在拒绝域之中，则结论为“对立假说正确”。反之，则结论为“无法判定虚无假说为误”。

检验统计量的值在拒绝域之中，因此

总体的克莱姆相关系数的值大于0
=“性别”和“希望的表白方式”有关联。

这样的对立假说为正确！

检验统计量即使在拒绝域中，单以“检验”并无法给出“对立假说‘绝对’正确……但是，只能作虚无假说存在正确的机率。其值最大为（α×100）%”的结论。



……大致上就是这样。
解说者
原来如此……
不过，总觉得步骤7怪怪的……
?
检验统计量的值如果不在拒绝域中，似乎可说“虚无假设是正确的”呀……
拒绝域
0
5.9915
可惜的是，并不能这么说……只能说“无法判定虚无假设为误”而已。
是这样吗？
举例来说，请假设刚才例题的χ_0^2为2.5013。
拒绝域
0
2.5013
5.9915
此值并不在拒绝域之中吧！

因此，当然不可以作出“总体的克莱姆相关系数的值大于0”的结论。
唔……
然而,却不能断言“总体的克莱姆相关系数为0”。
再来举个更容易理解的例子吧！
假设琉衣想吃的布丁被某人吃掉了。
谁呀！是谁这么可恶！？
嫌疑犯由美出现了。
由美，太过份了！
只是举例啦！
再设定“检验”的种类和置信水平等细节……
虚无假说 由美是犯人
对立假说 由美不是犯人
咚
以这个假说为对象来进行“检验”。

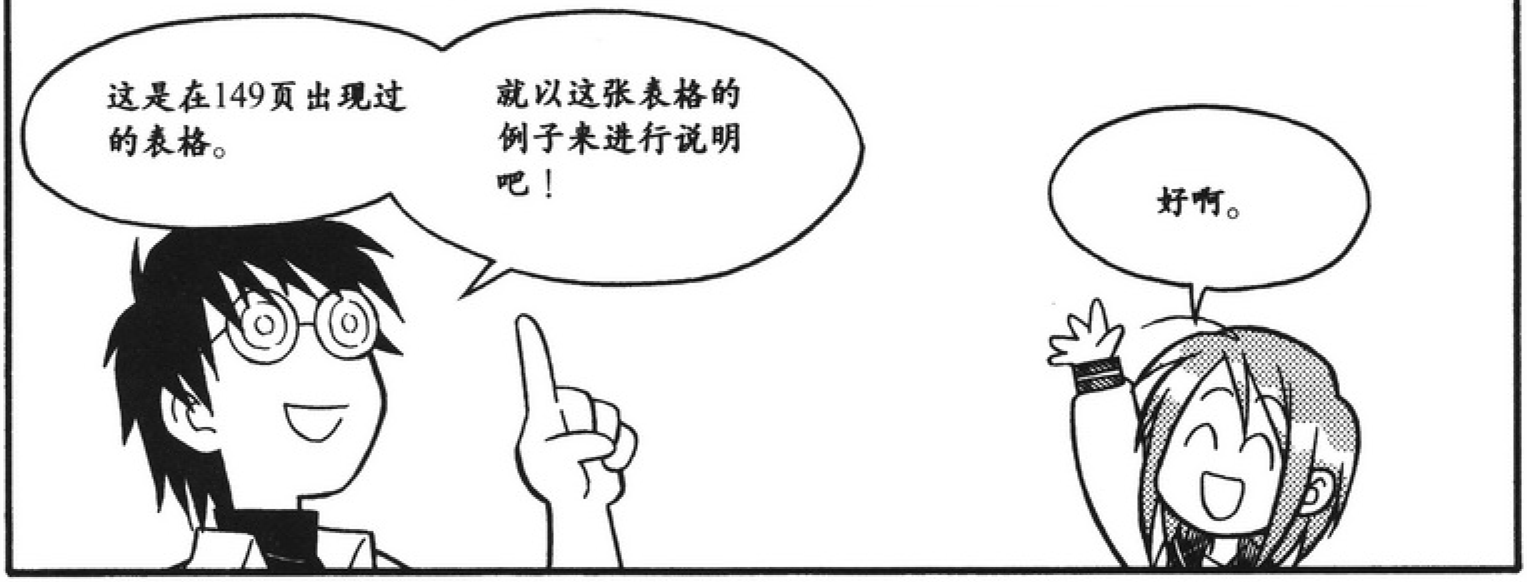

假设由美具有非常有利的不在场证明。
那个时间，
我去补习班了。
如果真是这样，就没有余地反驳“由美不是犯人”的结论。
我先走一步。
警察
对不起
是呀……
那么，假设由美只能举出令人怀疑的不在场证明。
那个时间我在附近散步
太可疑了…
若为如此，则当然无法做出“由美不是犯人”的结论。
然而，也不能因此就断定“由美就是犯人”。
拜托……
证据呢？
原来如此……
就是这么一回事。那么就继续接下来的课程吧！
对了
等我一下！
哒
？

✿3.虚无假说和对立假说✿

ぷるん~
多亏了你，让我想起冰箱里还有布丁。
没有被偷走真是太好了。
那么，在进行“检验”时，
必须建立虚无假说和对立假说。
话说回来，虚无假说和对立假说是什么？
刚才你说等一下要说明，但我还没听到？
其实，很难用三言两语来说明虚无假说和对立假说。
噴？

“检验”的实例

名　称	可使用的情况之实例
独立性检验	推测总体中，“性别”和“希望的表白方式”的克莱姆相关系数之值是否为0。
相关比检验	推测总体中，“喜欢的服装品牌”和“年龄”的相关比之值是否为0。
无相关检验	推测总体中，“1个月使用的化妆品费用”和“1个月使用的置装费”的相关系数之值是否为0。
总体平均数差检验	推测东京都的女高中生和大阪府的女高中生“每月的零花钱”是否不同。 ※注意，这个例子中设定了两个总体。
总体比例差检验	推测居住于都市的有投票权者和居住于农村的有投票权者中，对“××内阁的支持率”是否不同。 ※注意，这个例子中设定了两个总体。



■独立性检验

虚无假说	总体中“性别”和“希望的表白方式”之克莱姆相关系数的值为0。
对立假说	总体中“性别”和“希望的表白方式”之克莱姆相关系数的值大于0。

■相关比检验

虚无假说	总体中“喜欢的服装品牌”和“年龄”之相关比的值为0。
对立假说	总体中“喜欢的服装品牌”和“年龄”之相关比的值大于0。

■无相关检验

虚无假说	总体中“1个月使用的化妆品费”和“1个月使用的装置费”之相关系数的值为0。
对立假说	总体中“1个月使用的化妆品费”和“1个月使用的装置费”之相关系数的值不为0。 或 总体中“1个月使用的化妆品费”和“1个月使用的装置费”之相关系数的值大于0。 或 总体中“1个月使用的化妆品费”和“1个月使用的装置费”之相关系数的值小于0。

■总体平均数差检验

虚无假说	东京都的女高中生和大阪府的女高中生的“每个月零用钱”相等。
对立假说	东京都的女高中生和大阪府的女高中生的“每个月零用钱”不相等。 或 比起东京都的女高中生，大阪府的女高中生的“每个月零用钱”较多。 或 比起东京都的女高中生，大阪府的女高中生的“每个月零用钱”较少。

■总体比例差检验

虚无假说	居住在都市的有投票权者和居住在农村的有投票权者中，对“XX内阁的支持率”相等。
对立假说	居住在都市的有投票权者和居住在农村的有投票权者中，对“XX内阁的支持率”不相等。 或 比起居住在都市的有投票权者，居住在农村的有投票权者，对“××内阁的支持率”较高。 或 比起居住在都市的有投票权者，居住在农村的有投票权者，对“××内阁的支持率”较低。


感觉怎么样？
虚无假说中，被推论为并非“总体的克莱姆相关系数的值‘越接近’于0”，而是“总体中的克莱姆相关系数的值为0”的难以证明的假说，你明白了吗？
嗯！听起来是极端的理论嘛！
因此，虚无假说就是“是！”
“××为相等”的肯定假说；
而对立假说则为“不是××”“××不相等”的否定假说。
你了解了吗？
没错耶！
看来难以证明的假说做为虚无假说，而和虚无假说对立的假说则称为对立假说。
嗯！原来如此……
大概有这样的了解就够了。


4.P值和“检验”的顺序
为“检验”下结论时的根据有……
① 检验统计量是否在拒绝域中
② P值是否小于置信水平
以上这两种。
①已经在刚才说明过了，而②则尚未说明。
“P值”是什么呢？
依“检验”的种类不同，思考方式多少会有所不同……
独立性检验的“P值”即为，
在虚无假说为真的情况下，则本次求出的值为大于或等于χ_0^2值的机率。
以先前的例子来说……
0
$\chi_0^2=8.0091$
就是这个部分的机率。


原来如此！
是指这个斜线部分
呀……
在电脑普及前，要算出P值可是相当浩大的工程哦！
直到20世纪90年代初期。
哇！
因此，以往作为“检验”结论的根据，
几乎都采用①。
那现在呢？
现在由于使用EXCEL就可求出P值，所以采用②的情况也越来越多了。
原来如此！
②和刚才解说过的①在步骤上有些差异……
因此，让我再为你分析一下吧！
分析者
别用那种怪声音

步骤6P

调查在步骤5所求出的检验统计量值相对应的P值，是否比置信水平小。

置信水平为0.05。

由于检验统计量的皮尔森卡方统计量χ_0^2的值为8.0091，因此P值为0.0182。

0.0182＜0.05。也就是说，P值比较小。

如同先前所述，虽然依“检验”种类不同，结果也会不同，但是只要使用Excel，仍可以求出P值。

值得庆幸的是，独立性检验的P值可经由Excel来求得。详情请参照208页。

步骤7P

在步骤6p所得的P值若小于置信水平，即可作出“对立假说为正确”的结论。反之，则结论为“无法判定虚无假说为误”。

P值小于置信水平。因此

总体的克莱姆相关系数的值大于0
=“性别”和“希望的表白方式”有关联。

这样的对立假说正确！

即使P值小于置信水平，以“检验”并无法作出“对立假说‘绝对’正确”的结论。只能作出“虽然想说对立假说‘绝对’正确，但是，只能作虚无假说存在正确的机率为(P值×100)%”的结论。

原来如此，和①一样嘛。
大致上了解啦！
哦！你真是帮我打了一剂强心针呢。
呵呵！
听清楚了呦！琉衣，请回想①的布丁的例子。
即使P值大于置信水平，虽看似可以作出“虚无假说为正确”的结论，但实际上并不可以。
只能做出……
“无法判定虚无假说为误”的结论。
微笑

琉衣努力到最后一刻呢！
谢谢！
山本老师，真是太感谢你了。
其实虽然觉得统计学很难，但现在觉得懂了不少。
我觉得问卷调查的次数分布表非常有趣啊！
Pairls
呜呜~~~~~
听到你这样说，我也很开心哦！
真想做更多的分析！
哈哈哈……
琉衣也来从事我们的工作，如何呀？
对……耶！
这好像就是我最初的目的……
那么，我回去啦。

好吧！
去见五十岚先生。
啊！
五十岚先生
现在……
正在度
蜜月哦！
什么！？……结婚！？
结婚！？
我……
到底是为了什么才那么拼命
地学统计学呀？
ガーン
ガーン
他竟然有老
婆了……
啊？
难道你不是真
的对统计学感
兴趣吗？
吵，吵死啦……
哇—
哇—
危险！
晕倒
对不…


你还好吧？
哇哇哇哇哇哇哇……


从今以后请教我更多
知识吧！
山本老师……
啊？
啊？
咱们两个人的课程还会继续……
是吧？

✿ 5.独立性检验和齐性检验 ✿

齐性检验（test of homogeneity）与独立性检验是非常类似的“检验”方法。齐性检验的例子如下所示。请一边阅读，一边思考和独立性检验的差异。

例

“询问300名高中生！你希望对方用什么样的方式向你表白？”

- 打电话
- 发短信
- 当面

的报道，凛凛出版社刊载在女性杂志“P-girls”之中，然而凛凛出版社早已设立下列假说。

假说

打电话：发短信：当面

的人数比，女高中生和男高中生有所不同。

因此，为了确定上述的假说是否正确。凛凛出版社从“居住在日本的全体女高中生”和“居住在日本的全体男高中生”中，各随机抽出一些人进行实际的问卷调查。其结果如下表。

	希望的告白方法	年龄	年龄
回答者1 ⋮ 回答者148	当面 ⋮ 发短信	17 ⋮ 16	女 ⋮ 男
回答者149 ⋮ 回答者300	打电话 ⋮ 发短信	15 ⋮ 18	女 ⋮ 男

然后“性别”和“希望的表白方式”的交叉资料表如下。

		希望的表白方式			合 计
		打电话	发短信	当面	
性别	女性	34	61	53	148
	男性	38	40	74	152
合 计		72	101	127	300

请用齐性检验来推测上述的假说是否正确。而其置信水平设为0.05。

解

步骤1	定义总体。	假设“居住在日本的全体女高中生”和“居住在日本的全体男高中生”为总体。
步骤2	建立虚无假说和对立假说。	虚无假说为 “‘打电话：发短信：当面’的比例，两者相等”。 对立假说为 “‘打电话：发短信：当面’的比例，两者不相等”。
步骤3	选择要进行的“检验”种类。	进行齐性检验。
步骤4	决定置信水平。	假设置信水平为0.05。
步骤5	从样本资料求出检验统计量的值。	本例题中欲进行的是齐性检验。因此检验统计量为皮尔森卡方统计量χ_0^2。本例题中的χ_0^2值已在132页计算完毕。$\chi_0^2=8.0091$。且本例题中，若虚无假说为真，则皮尔森统计量χ_0^2为服从自由度（2-1）×（3-1）=1×2=2的卡方分布。
步骤6	调查在步骤5所求出的检验统计量值，是否在拒绝域之中。	检验统计量χ_0^2的值为8.0091。由于置信水平α为0.05，因此根据103页的卡方分布表，拒绝域为“5.9915”以上。检验统计量的值在拒绝域之中。
步骤7	若步骤6的检验统计量值在拒绝域之中，则结论为“对立假说正确”。反之，则结论为“无法判定虚无假说为误”。	检验统计量的值在拒绝域之中。因此对立假说为“‘打电话：发短信：当面’的比例，两者不相等”为正确。

如何？例题和解答都和独立性检验的例子几乎相同。

下面我们来确认独立性检验和齐性检验的相异之处。

相异处有3点。首先，定义的总体不同。前者是“居住在日本的全体高中生”的一群总体，后者则是“居住在日本的全体女高中生”和“居住在日本的全体男高中生”的两类总体。此外，假说也不相同。前者是

虚无假说	总体的克莱姆相关系数的值为0 =“性别”和“希望的表白方式”不相关。
对立假说	总体的克莱姆相关系数的值大于0 =“性别”和“希望的表白方式”相关。

而后者是

虚无假说	（打电话：发短信：当面）的比例，两者等。
对立假说	（打电话：发短信：当面）的比例，两者不相等。

另外，顺序也不太一样。前者是收集资料后才建立假说，而后者是在收集资料前就先建立假说。

如同前段所说明的，独立性检验和齐性检验有明确的相异点。然而，实际上，通常的情况是，本来想做独立性检验，却误做了齐性检验，或是想要两种都做做看，之所以想进行独立性检验，通常是因为已经进行了齐性检验，或是想进行齐性检验时，通常是因为已经进行了独立性检验。因此，请特别注意。

✿6. “检验”的结论表现✿

到目前为止“检验”中的结论都是以

> 若检验统计量的值在拒绝域之中，则做出“对立假说为正确”的结论。反之，则作出“无法判定虚无假说为误”的结论。

来表现。但实际上，这样的表现方式并非一般性的。

“检验”的结论的表现形式有很多种，兹总整理于下表。

◆表7.4 “检验”的结论表现

检验统计量的值在拒绝域之中	检验统计量的值不在拒绝域之中
•对立假说为正确。 •有信心。 •放弃虚无假说。	•无法判定虚无假说为误。 •无信心。 •无法放弃虚无假说。 •保留虚无假说 •无法判定虚无假说为不真。 •采纳虚无假说。

“有信心”“无信心”的表现不是比较易于使用吗？那么，为什么我要故意使用非一般性的表现？真正的理由如下所述。

我想恐怕只是想确认检验统计量的值和P值的大小吧！我已经注意到，学习“检验”的人之中，有些人在完全不了解用途的状况下，就轻易地将“有信心”时常挂在嘴边。这些人完全不了解“有信心”的意义，事实上他们是在未确立虚无假说和对立假说之下，就直接进行“检验”。我认为这些人根本不明白总体的定义。以前我也曾想过：对于才刚开始学统计学的人再怎么吹毛求疵也没用。然而，若对虚无假说和对立假说的意义不明了，又怎么下结论？果然，吹毛求疵并不是这么无理的要求。因此，本书为了让虚无假说和对立假说可以永存于读者脑海中，特别使用了“对立假说为正确”和“无法判定虚无假说为误”的表现方式进行处理。

例题和解答

例题

下表为沿用前一章138页的交叉资料表。

		咖啡和红茶哪一种比较好?		合 计
		咖啡	红茶	
常点餐的料理种类	日式料理	43	33	76
	西式料理	51	53	104
	中式料理	29	41	70
合 计		123	127	250

请用独立性检验推测总体为“居住在日本20岁以上的人”之中，“常点的料理种类”和“咖啡和红茶哪一种比较好?”的克莱姆相关系数的值是否大于0，意即“常点的料理类别”和“咖啡和红茶哪一种比较好?”是否有关联。另外，置信水平设为0.01。

解答

步骤1	定义总体。	设“居住在日本20岁以上的人”为总体。
步骤2	建立虚无假说和对立假说。	虚无假说为“常食用的料理种类”和“咖啡和红茶哪一种比较好?”有相关。 对立假说为“常食用的料理种类”和“咖啡和红茶哪一种比较好?”不相关。
步骤3	选择要进行“检验”的种类。	进行独立性检验。
步骤4	决定置信水平。	设置信水平为0.01。
步骤5	从样本资料求出检验统计量的值。	本例题欲进行的是独立性检验。因此检验统计量为皮尔森卡方统计量χ_0^2。本例题中的χ_0^2值已在141页计算完毕。χ_0^2=3.3483。
步骤6	调查步骤5所求出的检验统计量的值，是否在拒绝域之中。	检验统计量χ_0^2的值为3.3483。由于置信水平α为0.01，因此根据103页的卡方分布表，拒绝域为“9.2104”以上。检验统计量的值不在拒绝域之中。
步骤7	若步骤6的检验统计量在拒绝域之中，则结论为“对立假说正确”。反之，则结论为“无法判定虚无假说为误”。	检验统计量的值不在拒绝域之中。因此无法判定虚无假说——“常点的料理种类”和“咖啡和红茶哪一种比较好?”两者有相关为误。

总整理

• 所谓“检验”指的是，由样本数据来推测分析者针对总体所建立的假说是否正确的分析方法。

• “检验”的正确名称为统计的假说检验。

• 检验统计量是将样本数据转换为1个数值的公式。

• 置信水平一般都设为0.05或0.01。

• 拒绝域为对应置信水平的范围。

• 独立性检验为推测“总体的克莱姆相关系数的值是否为0”的分析方法。也可说是推测“交叉资料表中的两变量是否有关联”的分析方法。

• 若总体的克莱姆相关系数的值为0，则“皮尔森卡方统计量χ^2”为遵守卡方分布。

• 虚无假说若为真，独立性检验中的P值，为求出大于或等于本次所求出的皮尔森卡方统计量χ^2之机率。

• 在“检验”中，下结论的根据有2种：

① 检验统计量的值是否在拒绝域中。

② P值是否小于置信水平。

• 无论是否为独立性检验，其“检验”分析顺序均相同。具体来说，如下所述。

步骤1	定义总体。
步骤2	建立虚无假说和对立假说。
步骤3	选择要进行的“检验”种类。
步骤4	决定置信水平。
步骤5	从样本数据求出检验统计量的值。
步骤6	调查在步骤5所求出的检验统计量值，是否在拒绝域之中。
步骤7	若在步骤6中检验统计量的值在拒绝域之中，则结论为“对立假说成立”。若非如此，则结论为“无法判定虚无假说为误”。
步骤6P	调查与在步骤5所求出的检验统计量值相对应的P值，是否比置信水平小。
步骤7P	步骤6P所得的P值若小于置信水平，则可作出“对立假说正确”。反之，则结论为“无法判定虚无假说为误”。

◆附 录◆

运用 EXCEL 计算

在此，利用Excel函数功能进行解说。

1. 做成次数分布表（的一部分）
2. 算出平均数、中位数、标准差
3. 做成“次数分布表”（的一部分）
4. 算出标准分数、离差
5. 算出标准正态分布的机率
6. 算出卡方分布的横轴刻度
7. 算出相关系数的值
8. 独立性检验

已经熟悉Excel函数功能的读者，建议你先从“2算出平均数、中位数、标准差”入手。

1 做成次数分布表（一部分）

使用33页的资料

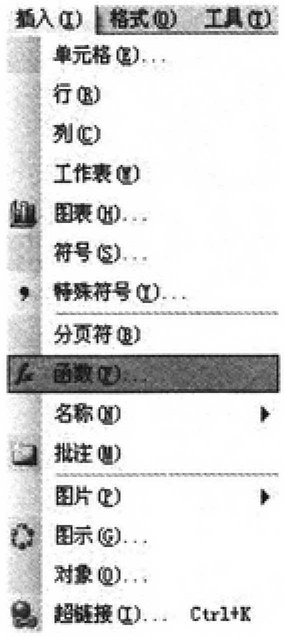
步骤1

选取“J3”单元格。

A	B	C	D	E	F	G	H	I	J
	价格（日元）			价格（日元）					
拉面馆1	700		拉面馆26	780		以上	未满	（以下）	次数
拉面馆2	850		拉面馆27	590		500	600	599	
拉面馆3	600		拉面馆28	650		600	700	699	
拉面馆4	650		拉面馆29	580		700	800	799	
拉面馆5	980		拉面馆30	750		800	900	899	
拉面馆6	750		拉面馆31	800		900	1000	999	
拉面馆7	500		拉面馆32	550					
拉面馆8	890		拉面馆33	750					
拉面馆9	880		拉面馆34	700					
拉面馆10	700		拉面馆35	600					
拉面馆11	890		拉面馆36	800					
拉面馆12	720		拉面馆37	800					
拉面馆13	680		拉面馆38	880					
拉面馆14	650		拉面馆39	790					
拉面馆15	790		拉面馆40	790					
拉面馆16	670		拉面馆41	780					
拉面馆17	680		拉面馆42	600					
拉面馆18	900		拉面馆43	670					
拉面馆19	880		拉面馆44	680					
拉面馆20	720		拉面馆45	650					
拉面馆21	850		拉面馆46	890					
拉面馆22	700		拉面馆47	930					
拉面馆23	780		拉面馆48	650					
拉面馆24	850		拉面馆49	777					
拉面馆25	750		拉面馆50	700					

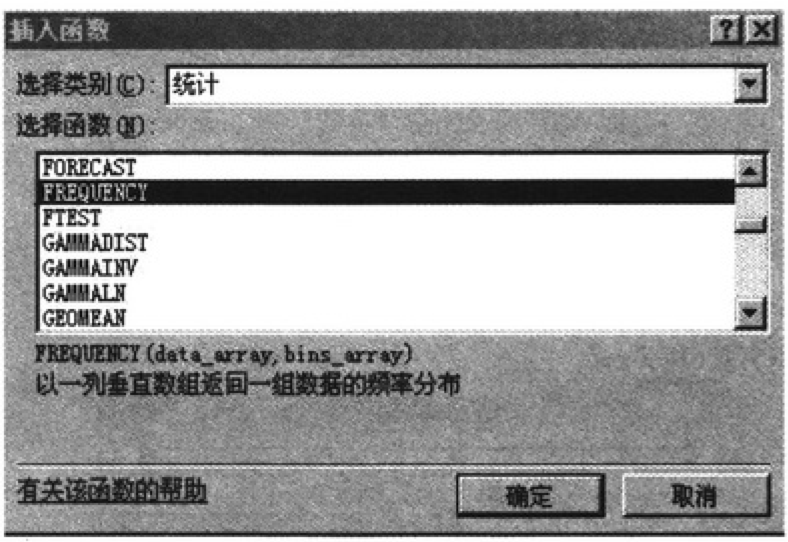
步骤 2

从工具栏的“插入”中选“函数”一项。



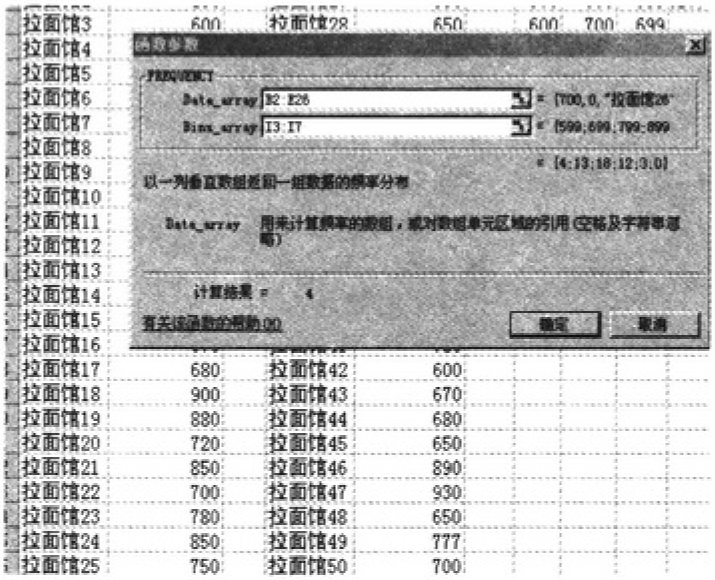
步骤 3

在“选取类别”中选择“统计”，再从“选择函数”中选择“FREQUENCY”。



步骤 4

选取下图所示的范围，点“确定”按钮。



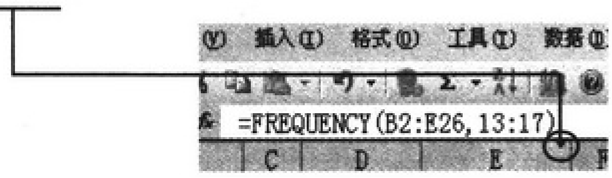
步骤 5

从单元格“J3”为起点，如同下图，选取单元格“J3”到“J7”的范围。

G	H	I	J
以上	未满	（以下）	次数
500	600	599	4
600	700	699	
700	800	799	
800	900	899	
900	1000	999	

步骤 6

点数学式中的这个部分。



步骤 7

同时按下“Shift”键和“Ctrl”键后，按“Enter”键。

步骤 8

计算完成！

G	H	I	J
以上	未满	（以下）	次数
500	600	599	4
600	700	699	13
700	800	799	18
800	900	899	12
900	1000	999	3

2 算出平均数、中位数、标准差

步骤 1

选取单元格“B10”。

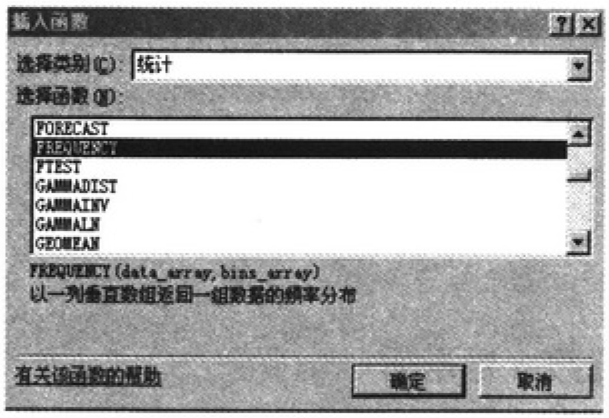
	A	B
1		A队
2	琉衣	86
3	小润	73
4	由美	124
5	小静	111
6	桃子	90
7	小枫	38
8		
9	平均数	
10	中位数	
11	标准差	
12		

步骤 2

从工具栏的“插入”中选取“函数”一项。

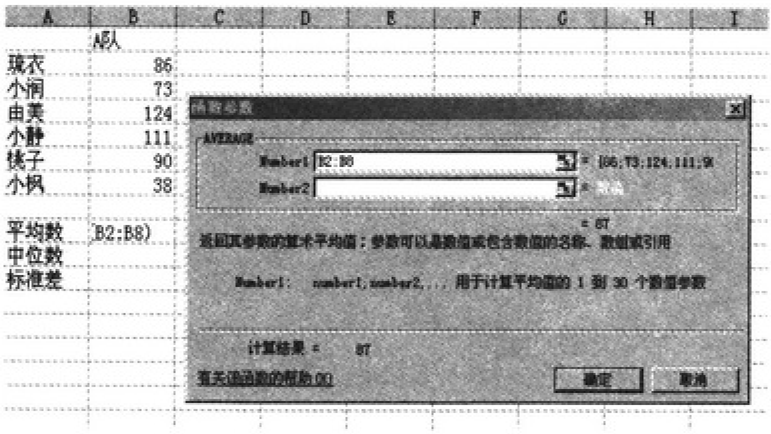
步骤 3

在“选取类别”中选择“统计”，再从“选取函数”中选择“AVERAGE”。



步骤 4

选取下图的范围，点击“确定”按钮。



步骤 5

计算完成！！

	A	B
1		A队
2	琉衣	86
3	小润	73
4	由美	124
5	小静	111
6	桃子	90
7	小枫	38
8		
9	平均数	87
10	中位数	
11	标准差	
12		
13		

步骤 6

与【步骤1】到【步骤5】相同步骤，求中位数和标准差。求中位数时，利用“MEDIAN”函数，求标准差时，则利用“STDEVP”函数。

3 做成“次数分布表”（一部分）

使用61页的资料。

步骤 1

选取单元格“F20”。

	A	B	C	D	E	F	G	H
1		新校服			新校服			新校服
2	1	喜欢		16	普通		31	普通
3	2	普通		17	喜欢		32	普通
4	3	普通		18	喜欢		33	喜欢
5	4	普通		19	喜欢		34	讨厌
6	5	讨厌		20	喜欢		35	喜欢
7	6	喜欢		21	喜欢		36	喜欢
8	7	喜欢		22	喜欢		37	喜欢
9	8	喜欢		23	讨厌		38	喜欢
10	9	喜欢		24	普通		39	普通
11	10	喜欢		25	喜欢		40	喜欢
12	11	喜欢		26	喜欢			
13	12	喜欢		27	讨厌			
14	13	普通		28	喜欢			
15	14	喜欢		29	喜欢			
16	15	喜欢		30	喜欢			
17								
18								
19						次数		
20					喜欢			
21					普通			
22					讨厌			
23								

步骤 2

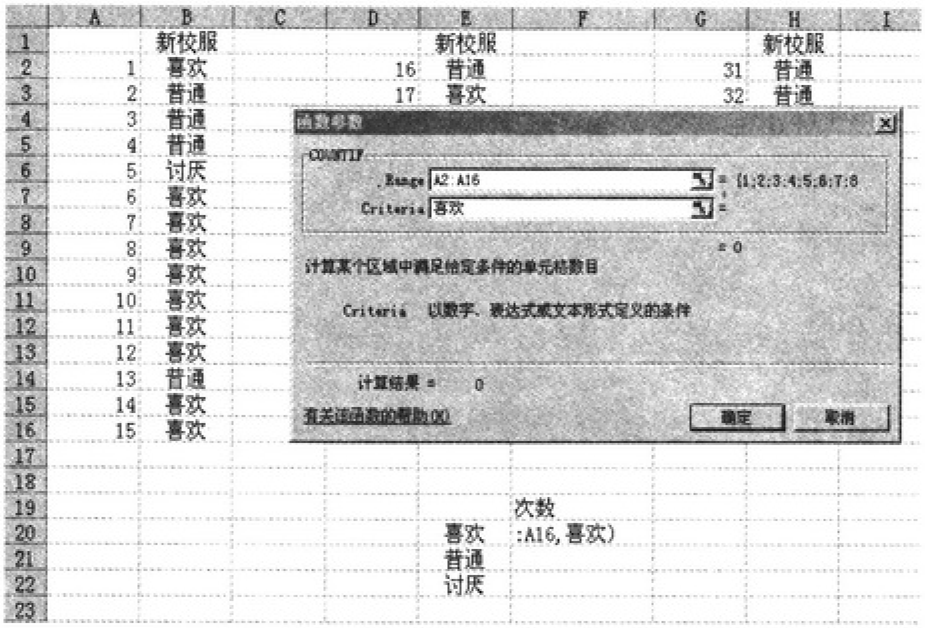
从工具栏的“插入”中选取“函数”一项。

步骤 3

在“选取类别”中选择“统计”，再从“选取函数”中选择“COUNTIF”。

步骤 4

选取下图的范围，在“Crjteria”直接输入“喜欢”，点“确定”按钮。



步骤 5

计算完成！

	A	B	C	D	E	F	G	H
1		新校服			新校服			新校服
2	1	喜欢		16	普通		31	普通
3	2	普通		17	喜欢		32	普通
4	3	普通		18	喜欢		33	喜欢
5	4	普通		19	喜欢		34	讨厌
6	5	讨厌		20	喜欢		35	喜欢
7	6	喜欢		21	喜欢		36	喜欢
8	7	喜欢		22	喜欢		37	喜欢
9	8	喜欢		23	讨厌		38	喜欢
10	9	喜欢		24	普通		39	普通
11	10	喜欢		25	喜欢		40	喜欢
12	11	喜欢		26	喜欢			
13	12	喜欢		27	讨厌			
14	13	普通		28	喜欢			
15	14	喜欢		29	喜欢			
16	15	喜欢		30	喜欢			
17								
18								
19						次数		
20					喜欢	28		
21					普通			
22					讨厌			

步骤 6

与【步骤1】到【步骤5】相同步骤，求“普通”和“讨厌”的次数。

4 算出标准分数、离差

使用72页的资料。

从【步骤1】到【步骤5】是标准分数的相关程序。而从【步骤10】到【步骤12】为离差的相关程序。

虽然Excel中存在可求出标准分数的函数，然而并不存在可求出离差的函数。但是，如果利用标准计分的结果，将能更快求出离差。因此，本书使用Excel求离差。

步骤 1

选取单元格“E2”。

	A	B	C	D	E	F	G
1		历 史			标准分数	离差	
2	琉衣	73		琉衣			
3	由美	61		由美			
4	A	14		A			
5	B	41		B			
6	C	49		C			
7	D	87		D			
8	E	69		E			
9	F	65		F			
10	G	36		G			
11	H	7		H			
12	I	53		I			
13	J	100		J			
14	K	57		K			
15	L	45		L			
16	M	56		M			
17	N	34		N			
18	O	37		O			
19	P	70		P			
20	平均数	53					
21	标准差	22.7					
22							

步骤 2

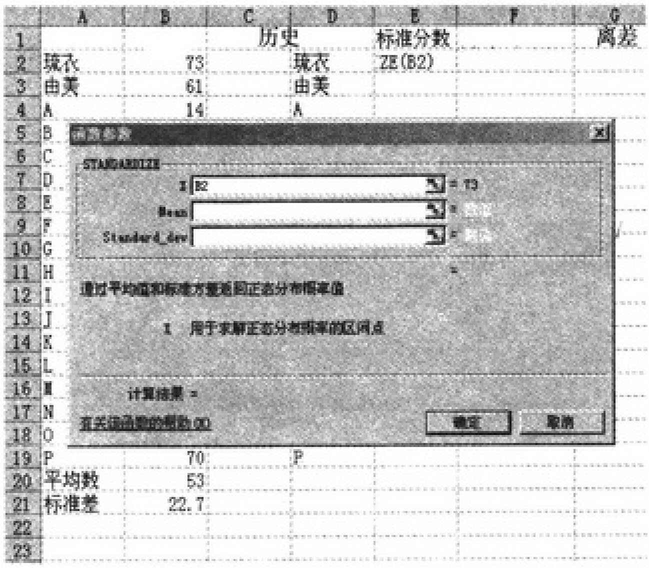
从工具栏的“插入”中选取“函数”一项。

步骤 3

在“选取类别”中选择“统计”，再从“选取函数”中 选择“STANDARDIZE”。

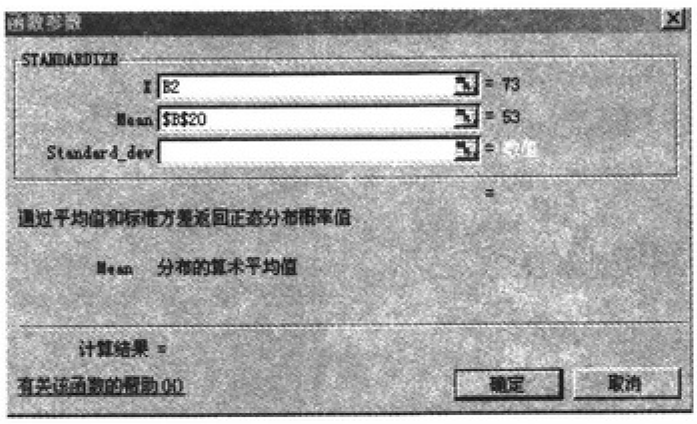
步骤 4

选取单元格“B2”。



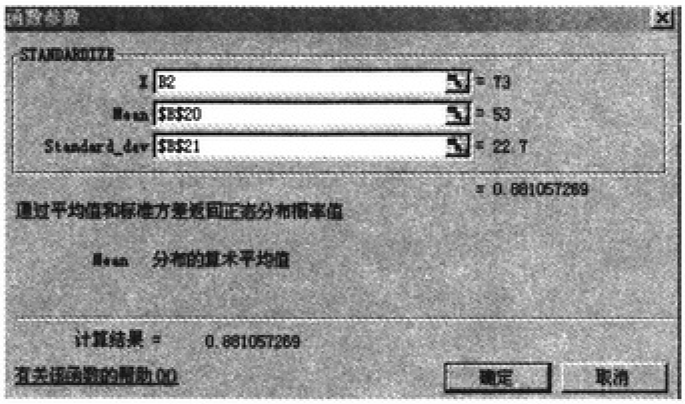
步骤 5

于“Mean”中选择单元格“B20”后，按一次“F4”键，并确认“B20”是否变成“B20”。



步骤 6

于“Standard__dev”中选取“B21”后，按一次“F4”键，并确认“B21”是否变成“B21”后，点“确定”按钮。



红历 离差

步骤 7

确认是否已求出琉衣的标准分数。

	A	B	C	D	E	F
1			历史		标准分数	
2	琉衣	73		琉衣	0.88	
3	由美	61		由美		
4	A	14		A		
5	B	41		B		
6	C	49		C		
7	D	87		D		
8	E	69		E		
9	F	65		F		
10	G	36		G		
11	H	7		H		
12	I	53		I		
13	J	100		J		
14	K	57		K		
15	L	45		L		
16	M	56		M		
17	N	34		N		
18	O	37		O		
19	P	70		P		
20	平均数	53				
21	标准差	22.7				
22						

离差

步骤 8

将鼠标移近单元格“E2”的右下角，待鼠标变为“黑色十字游标”后，按下鼠标左键，拖拉至“E19”后放开左键。

D	E	F
	标准分数	离　差
琉衣	0.88	
由美		
A		
B		
C		
D		
E		
F		
G		
H		
I		
J		
K		
L		
M		
N		
O		
P		

步骤 9

标准差计算完成！！

D	E	F
	标准分数	离　差
琉衣	0.88	
由美	0.35	
A	1.71	
B	-0.53	
C	-0.18	
D	1.49	
E	0.7	
F	0.53	
G	-0.75	
H	-2.02	
I	0	
J	2.07	
K	0.18	
L	-0.35	
M	0.13	
N	-0.84	
O	-0.7	
P	0.75	

步骤 10

选取“F2”，在单元格内输入“=E2*10+50”，然后按下“Enter”键。

D	E	F
	标准分数	离　差
琉衣	0.88	=E2*10+50
由美	0.35	
A	1.71	
B	-0.53	
C	-0.18	
D	1.49	
E	0.7	
F	0.53	
G	-0.75	
H	-2.02	
I	0	
J	2.07	
K	0.18	
L	-0.35	
M	0.13	
N	-0.84	
O	-0.7	
P	0.75	

步骤 11

重复【步骤8】的操作。

步骤 12

离差计算完成！！

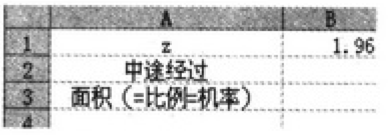
D	E	F
	标准分数	离　差
琉衣	0.88	58.8
由美	0.35	53.5
A	1.71	32.9
B	-0.53	44.7
C	-0.18	48.2
D	1.49	64.9
E	0.7	57
F	0.53	55.3
G	-0.75	42.5
H	-2.02	29.8
I	0	50
J	2.07	70.7
K	0.18	51.8
L	-0.35	46.5
M	0.13	51.3
N	-0.84	41.6
O	-0.7	43
P	0.75	57.5

5 算出标准正态分布的机率

使用93页的资料。

步骤 1

选取单元格“B2”。



步骤 2

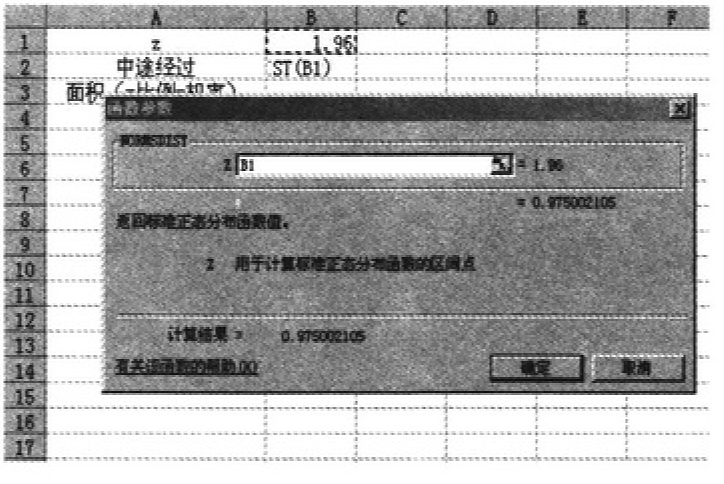
从工具栏的“插入”中选“函数”一项。

步骤 3

在“选取类别”中选择“统计”，再从“选取函数”中选择“NORMSDIST”。

步骤 4

选取单元格“B1”，按下“确定”键。



步骤5

其实“NORMSDIST”为求出下图机率的函数。
在此，于单元格“B3”内输入“=B2-0.5”。

	A	B
1	z	1.96
2	中途经过	0.975002
3	面积（=比例=机率）	=B2-0.5

步骤6

计算完成！！

	A	B
1	z	1.96
2	中途经过	0.975002
3	面积（=比例=机率）	0.475002

6 算出卡方分布的横轴刻度

使用104页的资料。

步骤1

选取单元格“B3”。

	A	B
1	P	0.05
2	自由度	1
3	卡方分布	

步骤2

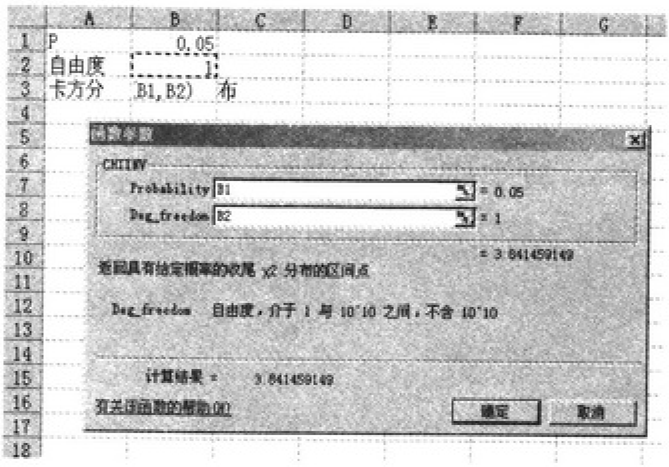
从工具栏的“插入”中选“函数”一项。

步骤3

在“选取类别”中选择“统计”，再从“选取函数”中选择“CHIINV”。

步骤 4

选取单元格“B1”和单元格“B2”，按下“确定”键。



步骤 5

计算完成！

	A	B
1	P	0.05
2	自由度	1
3	卡方分	3.84布459

7 算出相关系数的值

使用116页的资料。

步骤 1

选取单元格“B14”。

	A	B	C
1		化妆品费（日元）	置装费（日元）
2	A小姐	3000	7000
3	B小姐	5000	8000
4	C小姐	12000	25000
5	D小姐	2000	5000
6	E小姐	7000	12000
7	F小姐	15000	30000
8	G小姐	5000	10000
9	H小姐	6000	15000
10	I小姐	8000	20000
11	J小姐	10000	18000
12			
13			
14	相关系数		

步骤 2

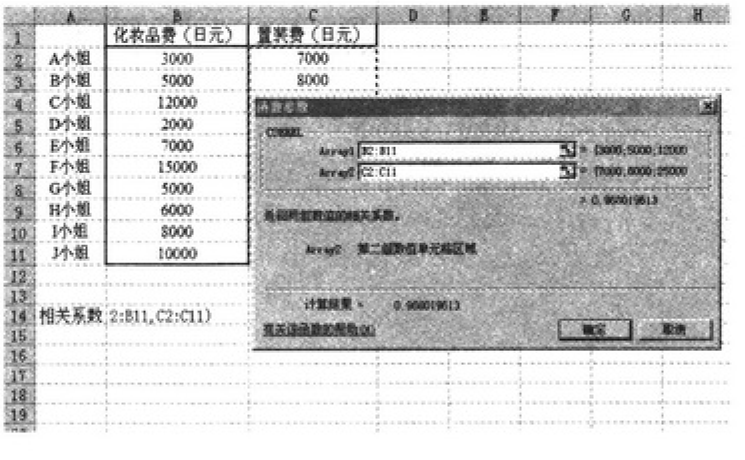
从工具栏的“插入”中选“函数”一项。

步骤 3

在“选取类别”中选择“统计”，再从“选取函数”中选择“CORREL”。

步骤 4

选取下图的范围后，按下“确定”键。



步骤 5

计算完成！！

	A	B	C
1		化妆品费（日元）	置装费（日元）
2	A小姐	3000	7000
3	B小姐	5000	8000
4	C小姐	12000	25000
5	D小姐	2000	5000
6	E小姐	7000	12000
7	F小姐	15000	30000
8	G小姐	5000	10000
9	H小姐	6000	15000
10	I小姐	8000	20000
11	J小姐	10000	18000
12			
13			
14	相关系数	0.968019613	

参考

非常可惜，并不存在可求出相关比和克莱姆相关系数的Excel函数。

8 独立性检验

使用157的资料。

步骤 1

选取单元格“B8”。

	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性				
9	男性				
10					
11					
12	P值				
13					
14					

步骤 2

于单元格“B8”内，输入“=E2*B4/E4”。然后按下“Enter”键。

	B	C	D	E
	打电话	发短信	当面	合计
性	34	61	53	148
性	38	40	74	152
	72	101	127	300
	打电话	发短信	当面	
性	=E2*B4/E4			
性				

步骤 3

选取单元格“B8”内的“E2”文字部分，连按3次“F4”键，并确认“E2”是否变为“$E2”后，按下“Enter” 键。

	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性	=$E2*B4/E4			
9	男性				
10					
11					
12	P值				
13					
14					

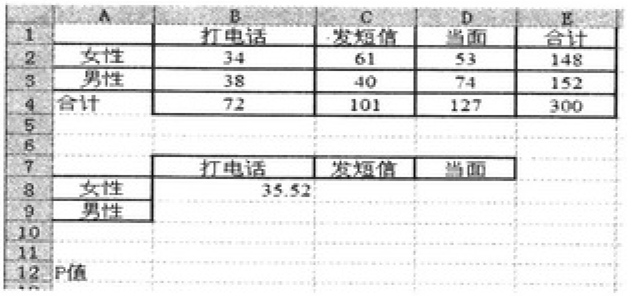
步骤 4

选取单元格“B8”内的“B4”文字部分，连按2次“F4”键。并确认“B4”是否变为“B$4”。选取单元格“B8”内的“E4”文字部分，按1次“F4”键，确认“E4”是否变为“E4”后，按下“Enter”键。

	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性	=$E2*B$4/E4			
9	男性				
10					
11					
12	P值				
13					

步骤 5

选取单元格“B8”，将鼠标移近单元格“B8”的右下角，待鼠标变为“黑色十字鼠标”后，按下鼠标左键，拖来至“D8”后放开左键。



	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性	35.52			
9	男性				
10					
11					
12	P值				

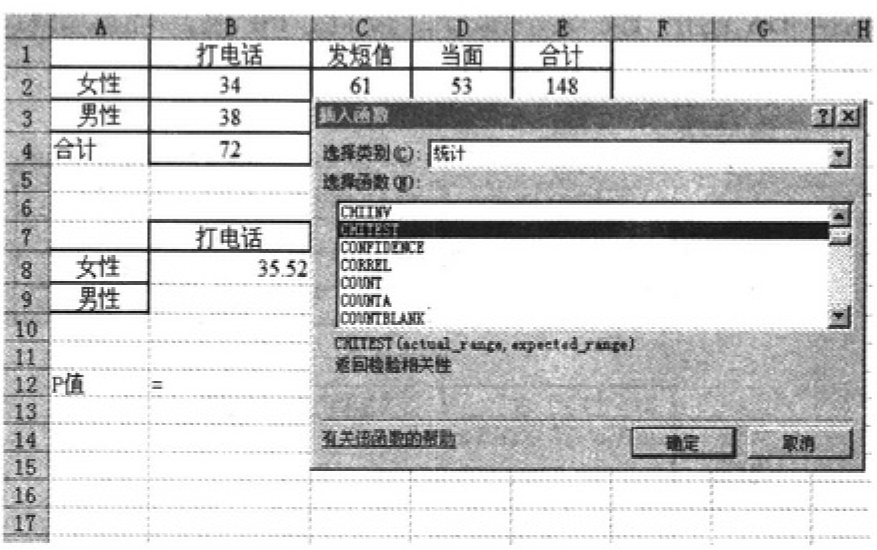
步骤 6

从单元格“B8”选取单元格“D8”，将鼠标移近单元格“D8”的右下角，待鼠标变为“黑色十字鼠标”后，按下鼠标左键，拖拉至单元格“D9”后放开左键。

	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性	35.52	49.82667	62.65333	
9	男性				
10					

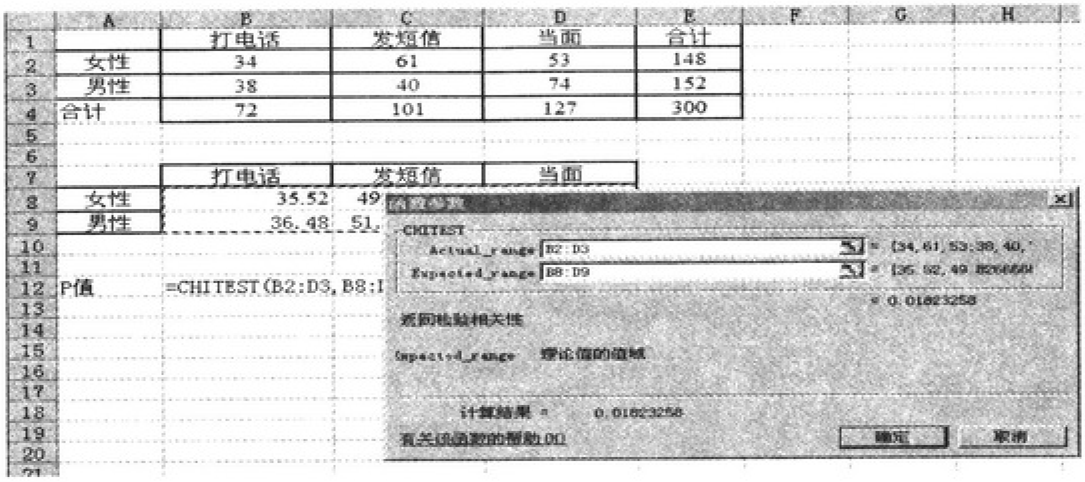
步骤 7

选取单元格“B12”。从工具列的“插入”中点选“函数”。在“选取类别”中选择“统计”，再从“选取函数”中选择“CHITEST”。



步骤 8

选取下图的范围，按下“确定”按钮。



步骤 9

计算完成！！（※请确认此值是否与177页的P值一致。）

	A	B	C	D	E
1		打电话	发短信	当面	合计
2	女性	34	61	53	148
3	男性	38	40	74	152
4	合计	72	101	127	300
5					
6					
7		打电话	发短信	当面	
8	女性	35.52	49.82666667	62.65333333	
9	男性	36.48	51.17333333	64.34666667	
10					
11					
12	P值	0.01823258			
13					

◆ 参考文献 ◆

・石村貞夫『統計解析のはなし』(東京図書) 1989

・内田治/菅民郎/高橋信『EXCELアドインによる多変量解析』(東京図書) 2003

・仮谷太一『医歯系・生物系のベーシック統計学』(共立出版) 1988

・菅民郎『新版　アンケートデータの分析』(現代数学社) 2000

・菅民郎『Excelで学ぶ統計解析入門(第2版)』(オーム社) 2003

・杉山髙 一『統計学入門』(絢文社) 1984

・鈴木武/山田作太郎『数理統計学―基礎から学ぶデータ解析―』(内田老鶴圃) 1996

・豊田秀樹『調査法講義』(朝倉書店) 1998

・東京大学教養学部統計学教室編『統計学入門』(東京大学出版会) 1991

・東京大学教養学部統計学教室編『自然科学の統計学』(東京大学出版会) 1992

・東京大学教養学部統計学教室編『人文・社会科学の統計学』(東京大学出版会) 1994

・永田靖『統計的方法のしくみ』(日科技連) 1996

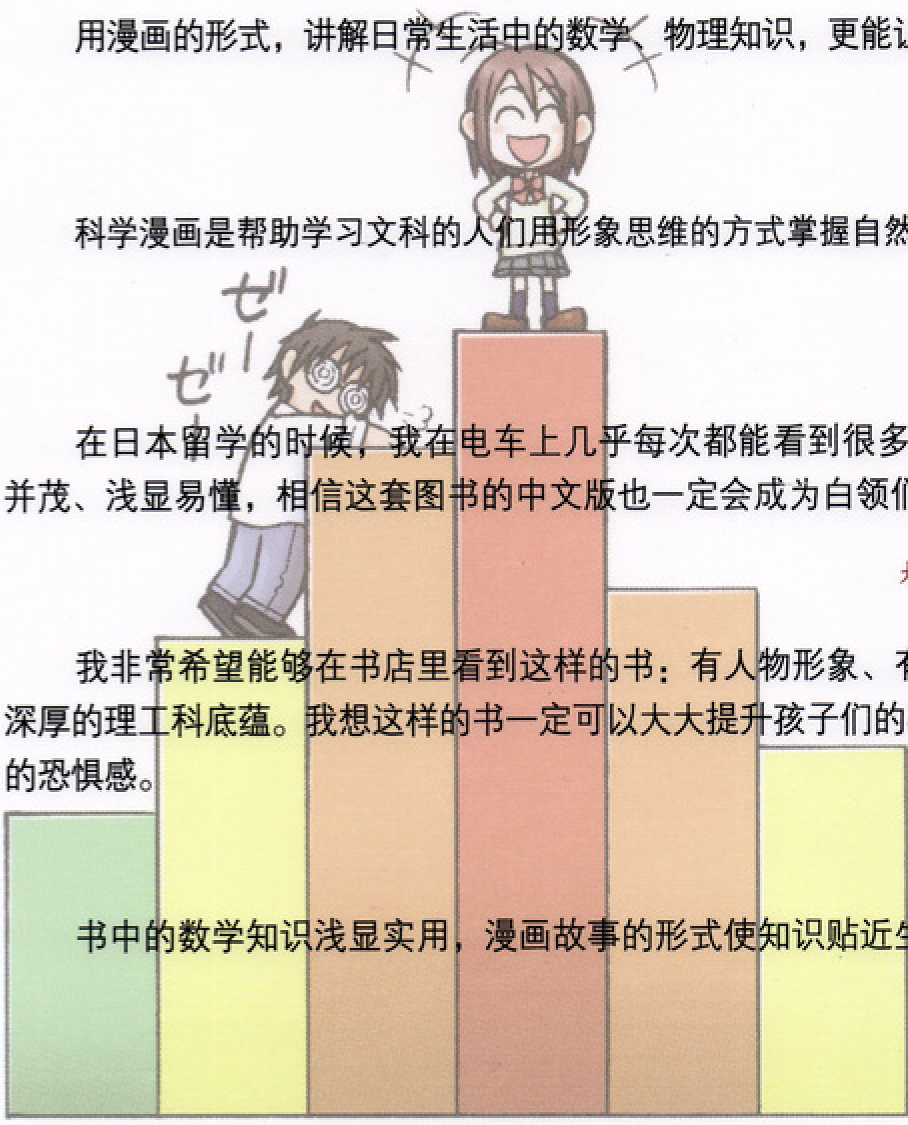
・永田靖/棟近雅彦『多変量解析法入門』(サイエンス社) 2001

・野田一雄/宮岡悦良『入門・演習　数理統計』(共立出版) 1990

・L.ゴニック/W.スミス(中村和幸　訳)『マンガ　確率・統計が驚異的によくわかる』(白揚社) 1995


换换眼镜吧！
不用！
推眼镜

（N-0349.0103）
责任编辑：张丽娜　赵丽艳
责任制作：董立颖　魏　谨
封面制作：【铭轩堂设计：13671110894】



用漫画这种形式讲数学、物理和统计学，十分有利于在广大青少年中普及科学知识。

周恩来、邓颖超秘书，周恩来邓颖超纪念馆顾问
中日友好协会理事，《数理天地》顾问，全国政协原副秘书长

用漫画和说故事的形式讲数学，使面貌冷峻的数学变得亲切、生动、有趣，使学习数学变得容易，这对于提高全民的数学水平无疑是功德无量的事。

《数理天地》杂志社　社长 总编
“希望杯”全国数学邀请赛组委会 命题委员会主任　周国镇

用漫画的形式，讲解日常生活中的数学、物理知识，更能让大家感受到数学殿堂的奥妙与乐趣。

《光明日报》 原副总编辑
中华炎黄文化研究会 常务副会长　鲁谆

科学漫画是帮助学习文科的人们用形象思维的方式掌握自然科学的金钥匙。

中国人民大学外语学院日语专业 主任
大学日语教学研究会 会长　成同社

在日本留学的时候，我在电车上几乎每次都能看到很多年轻的白领看这套图书，经济实惠、图文并茂、浅显易懂，相信这套图书的中文版也一定会成为白领们的手中爱物。

大连理工大学 能源与动力学院 博士 副教授

我非常希望能够在书店里看到这样的书：有人物形象、有卡通图、有故事情节，当然最重要的还有深厚的理工科底蕴。我想这样的书一定可以大大提升孩子们的学习兴趣，降低他们对于高深的理工科知识的恐惧感。

北京启明星培训学校 校长

书中的数学知识浅显实用，漫画故事的形式使知识贴近生活，概念更容易理解。

北京大学 数学科学学院 博士　张磊

上架建议：科普/漫画
ISBN 978-7-03-024796-4

科学出版社 东方科龙
http://www.okbook.com.cn
zhaoliyan@mail.sciencep.com

定价：29.80元