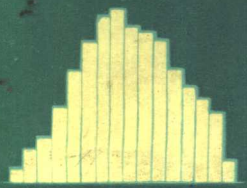
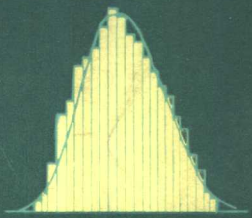
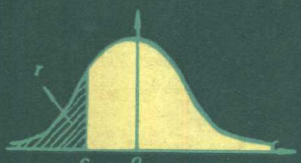
统计学漫话

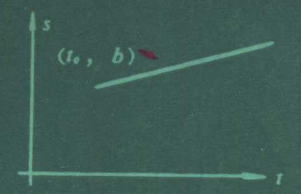
陈希孺　苏淳 编著

科学出版社









统计学漫话

陈希孺　苏　淳　编著

科学出版社

1987

内 容 简 介

本书以漫谈的方式，用大量通俗的例子引入和说明了统计学的基本概念、基本思想和基本方法．使读者对统计学的全貌有所了解，并且学会用统计观点去看待现实世界中的许多事物．

本书内容深入浅出、通俗易懂，可供具有中等文化程度的读者、从事实际工作的统计工作者阅读，也可供大专院校统计专业师生参考．

统 计 学 漫 话

陈希孺　苏　淳　编著

责任编辑　毕　颖

科学出版社出版

北京朝阳门内大街 137 号

中国科学院印刷厂印刷

新华书店北京发行所发行　　各地新华书店经售

*

1987 年 10 月第 一 版　开本：787×1092　1/32

1987 年 10 月第一次印刷　印张：9 1/2

印数：0001—5,000　字数：214,000

统一书号：13031·3685

本社书号：5201·13—1

定价：1.80 元

2.572

序　言

本书试图以漫谈的方式，用通俗的语言，向具有中等文化程度的读者介绍统计学的基本思想和方法。统计学在许多领域中都有广泛的应用，其重要性无需在此处强调。作者想表达一个想法：从一定程度上看，统计学的初步知识已构成一个人文化修养中必要的部分。这是因为，在现实世界中有许多事物需要用正确的统计观点去看待，才能得到恰当的理解，即使在日常生活中碰到的一些事情也不例外。

目前，由国内专家编写的统计学著作已出版不少，但对具有中等文化程度的读者来说，它们大都过于专深，需要用到高等数学和概率论的知识。另一个问题是，统计学不同于纯粹数学，不能单纯从公式的数学论证中去正确理解它，而必须对其基本概念和问题提法的实际背景、方法的思想、结果的解释、使用统计方法应注意的种种问题等，作深入的思考才行。正因为如此，我们在本书的编写过程中，对以上提到的诸方面问题作了仔细论述。

本书的编写方式，是在介绍每一个主题时，先从大家都理解的一些事物入手，经过分析，提出一些想法和问题，由此逐步展开，从而引入明确的概念、方法和理论。在每一步中，我们都尽量用一些通俗的比喻，以便把一些艰深的概念形象地表达出来，但又不失去科学的严谨性。这是作者定下的目标，但达到了多少，只能由广大读者来评判。

为适应具有中等文化程度的广大读者的需要，本书力图避免高深的数学知识。当然，讲统计学不能只空谈思想而不

涉及具体方法。相反，只有通过介绍具体方法，才能把观点和思想讲清楚。因此，书中介绍了一些重要的统计方法，仔细交代了方法的步骤及使用时要注意的地方。读本书自然不能像读一本通俗小说那么轻松，所以要求读者必须深入的思考。本书虽然力图避免枯燥，但由于作者写作能力有限，未免力不从心，不到之处望读者谅解。

除上述读者外，我们还希望本书对学习统计学的大学生，以及具有一定统计知识和实践经验的应用工作者，能多少有点用。在统计学的教学中，由于学时限制等原因，重点都放在介绍方法内容及其数学论证上，而对上面提到的若干问题则注意较少，本书希望能起一点拾遗补阙的作用。

我们还希望本书能对作者的广大同行——统计学教师有一点用处。也许作者的同行们都有这样的体会：这门课不好教，讲起来总觉得不容易使学生信服，而其难点又不在数学论证上。本书有些内容，包含了作者在教学实践中关于这些难点的若干思考和看法，虽不一定确切，但也许能引起广大同行注意这些问题。书中的叙述肯定会有不当之处，恳请广大读者不吝赐教。

在编写本书的过程中，项可风同志对第三章的写作提供了不少帮助。吴启光同志审阅了原稿，良多补益。在写作的最后阶段，陆传荣、林正炎同志安排了良好的条件，使作者得以如期完成。搁笔之际，感触颇多，除致以衷心的感谢外，特书于此，以志不忘。

作　者

1985年10月24日于西子湖畔

试读结束：需要全本请在线购买：　www.ertongbook.

目　录

一、什么是统计学 …… 1
1. 统计方法和统计学 …… 1
2. 通过事物的外在数量表现考察事物的规律性 …… 3
3. 由部分推断整体、总体和样本 …… 6
4. 统计性推断的错误和误差 …… 11
5. 学一点统计学 …… 13
二、获取数据(Ⅰ)——抽样调查 …… 16
1. 抽样调查的意义 …… 16
2. 要注意的问题 …… 18
3. 简单随机抽样、随机数表 …… 20
4. 集团抽样 …… 26
5. 分层抽样 …… 28
6. 随机化的重要性再议 …… 31
三、获取数据(Ⅱ)——试验的设计 …… 35
1. 前言 …… 35
2. 完全随机化设计 …… 37
3. 随机区组设计 …… 42
4. 平衡不完全随机区组设计 …… 47
5. 拉丁方设计 …… 51
6. 多因子试验 …… 57
7. 拉丁方用于多因子试验 …… 60
8. 正交拉丁方 …… 62
9. 正交表 …… 67
四、平均值与比率的精度 …… 76
1. 平均值的代表性问题 …… 76

2、总体方差 …… 79
3．样本均值的方差 …… 82
4．方差的估计、样本方差 …… 89
5．均值之差的估计 …… 93
附录 …… 96
五、分布与区间估计 …… 99
1．方差的局限性 …… 99
2．分布的概念 …… 101
3．分布的列表形式 …… 103
4．直方图与密度函数 …… 106
5．标准正态分布 …… 112
6．正态分布均值的区间估计(方差已知) …… 118
7．t 区间估计 …… 125
8．大样本情况 …… 131
六、概率初步知识 …… 135
1．什么是概率 …… 135
2．事件 …… 136
3．古典概率 …… 137
4．频率与统计定义 …… 145
5．主观概率 …… 151
6．随机变量 …… 152
7．概率分布 …… 155
8．均值和方差 …… 163
9．均值的大数定律 …… 167
七、假设检验 …… 169
1．原假设和对立假设 …… 169
2．拟合优度 …… 173
3．检验的水平 …… 176
4．两类错误 …… 183
5．卡·皮尔逊的 χ^2 检验 …… 186

6. 无关联性的检验 …… 196
7. u 检验 …… 201
8. 一样本 t 检验 …… 208
9. 与区间估计的关系 …… 211
10. 两样本 t 检验 …… 213
11. 非参数统计方法 …… 215
八、相关与回归 …… 218
1. 事物的联系 …… 218
2. 相关系数 …… 220
3. 相关系数的估计和检验 …… 224
4. 偏相关和其它 …… 228
5. "平均相关"的谬误 …… 230
6. 回归与回归方程 …… 232
7. 回归方程的估计(最小二乘法) …… 235
8. 残差、残差平方和与偏相关 …… 238
9. 回归用于估计条件平均值 …… 242
10. 回归名称的由来 …… 246
11. 几点注意事项 …… 248
附录 …… 251
九、方差分析法 …… 255
1. 基本思想 …… 255
2. 完全随机化设计 …… 258
3. 随机区组设计 …… 262
4. 对比试验 …… 268
5. 拉丁方设计 …… 270
6. 正交表设计 …… 275
7. F 检验 …… 278
8. 估计问题和最优处理的选定 …… 282
9. 交互效应 …… 284
附表 …… 289

一、什么是统计学

1. 统计方法和统计学

统计这个字眼大概对一般人都不陌生，因为恐怕谁都听见过这类说法：明天组织去游山，需要把去的人统计一下；工厂年终发奖金，要统计一下发千元以上的有几人，五百元至千元的有几人等等。可见"统计"成了一个常用词。总而言之，不少单位设有统计员，国家设有各级统计机构，以收集关于经济、人口和社会等方面的资料——称为**统计数据**。对这些数据要进行整理和分析的工作，以作出种种结论和预测，所用的方法就是**统计方法**。研究这种方法的学问，就叫做**统计学**。统计学在我国也常称为**数理统计学**。其实，后者应是前者的数学理论基础的部分，又称**理论统计学**。它是关于统计方法如何建立，及其正确性和有效性的数学论证。概括地说，统计方法是有关收集和取得数据资料，并对之进行整理、分析，以对所研究的问题作出一定的结论的那些方法，这样说基本上是正确的，但必须附加大量的补充解释，需要指出统计方法的特点何在，正是这些特点划清了统计方法和其他方法的界线。

如果使用高深的数学概念，可以用很少几行字把这个问题说清楚，但这对本书读者未必有益。在本书的整个叙述中，我们将力求回避抽象的数学概念，而尽量用实例、大家都熟悉的事物和形象化的比喻来说明问题，因而在语言上难免失之累赘，这是要请读者见谅的。现在先列举几个例子。

例 1 一个生产灯泡的工厂，以往一直采用某种工艺，现

在，厂里的技术人员对此提出一些改进措施，以期能改善产品质量．为了验证这个想法，取在新、老工艺下生产的灯泡各若干个去使用，记下每个参加试验的灯泡的寿命（即从开始使用到损坏所经历的时间）．所得数据可以这样处理：算出老工艺下生产的灯泡的平均寿命，例如为 420 小时；对新工艺下生产的灯泡也这样做，结果为 440 小时．于是作出结论：新工艺的确有助于改善质量，使用时间约可增加 20 小时．

例 2 要调查某县的个体农户在某年每户使用化肥的平均数量．全面的普查将涉及到数以十万计的农户，这是人力、物力和时间所不允许的．于是从这县的农户中抽选出若干户，比如 400 户，调查出这些户共用化肥 32,000 公斤，户均 80 公斤，以这个数字作为全县户均化肥用量的估计值．

例 3 一物件的重量 a 未知，放到天平上去称，由于天平有些误差，而对结果又要求有很高的精度，于是觉得称一次不够，就把它重复称三次，分别得出结果2.45、2.46和2.41（克），以其平均数 2.44 克作为 a 的估计．

例 4 为探索吸烟与患肺癌二者之间有否关联，调查一大批人，按是否吸烟和是否患肺癌分成四类：不吸烟也不患肺癌的，吸烟患肺癌的，不吸烟患肺癌的，吸烟不患肺癌的．根据这些数据，用一定的（统计）方法，作出像“吸烟与患肺癌二者有显著关联”之类的结论（这方法性质复杂些，待到后面第七章再作解释）．一个易于理解的处理方法如下：算出在抽选出的这批人中，不吸烟者患肺癌和吸烟者患肺癌的比率，比方说分别为十万分之三与十万分之十二，于是作出结论说，吸烟者患肺癌的危险性是不吸烟者患肺癌的四倍．这类报道，及其它性质类似的医学报道，常见于各种报章杂志．

以上是几个用统计方法研究问题的范例，其中都有获取、整理和分析数据的工作．由此可以总结出统计方法的哪些特

点呢？这就是我们在以下几段中要讨论的问题。应当交代清楚的是，由于我们力图少用抽象的数学论证，以下几段的说明还是不全面的，有的问题将在本书以后的叙述中再行补充。

2. 通过事物的外在数量表现考察事物的规律性

统计方法只是从事物外在的数量表现上去研究问题，不涉及事物的质的规定性。通俗些说，统计方法可能告诉你，从试验或观察结果来看如何如何，而不能回答为什么会如何如何。如在例1，试验结果有可能显示新工艺有助于改善质量，但其原因何在？也可能一目了然，也可能涉及专门学科领域中深奥的道理。在例4中，虽则许多统计资料都表明吸烟与患肺癌之间有关联，就是说吸烟的人看来更倾向于易得肺癌，但这种结论目前看来仍只能算是一种统计规律性——由表面上的数量关系而归纳出来的规律性。因为，不仅吸烟何以引发肺癌的机制在目前尚未确切研究清楚，甚至这二者之间表面上的联系是否真正反映了一种因果关系，在学者中也有分歧。有的学者认为，这二者表面上的关联，可能不过是由于它们受同一遗传基因的控制，其作用使那些易于染上吸烟嗜好的人，同时也倾向于易患肺癌，若这种看法被证明为确实，则戒烟既不会减少也不会增加患肺癌的危险性。更多的学者则认为，二者的联系是因果性的，尽管其机制目前没有充分弄明白。

这点值得作为统计方法的一个特点（或称"性质"也可以）提出来，是因为它划清了统计学和其他专门学科的界线，如在遗传学、医学……等中都用了不少统计方法，但统计学绝不能代替这些专门学科，而只是有助于它，可以说只是一个辅助性

试读结束：需要全本请在线购买： www.ertongbook.

的工具而已。了解这一点，就不致对统计方法和统计学者提出过高的期望，以为他们掌握的方法是万能的，可以在许多专门领域中单枪匹马地解决种种实际问题。一个从事于实际应用问题的统计工作者，其知识面愈广，就愈易与种种专门学科领域的人员取得共同语言，因而也就愈能对他们的工作提供一些帮助。

我们说统计方法只是一个辅助性的工具，仅是就以下一点而言：单纯的表面上的数量关系是否反映事物的本质，这本质究竟如何，必须依靠专门学科的研究才能下定论。这个提法不能理解为，统计方法的作用完全是被动性的，恰恰相反，事物的本质，其根本规律性的东西，一般都是隐藏得很深，它不时地在一些场合下有所表现。学者们注意和收集了这些资料，初看起来杂乱无章，而他如果具有一些统计的眼光，就有可能透过这些纷繁的数据而发现某种规律性的东西。这诚然还是表面上的，但可以作为专门研究的出发点，好比在一个刑事案件中，罪犯往往隐藏很深，但他总会多少留下一些痕迹，受过训练而有经验的侦察人员，能据此对案犯作案的动机和过程提出一些设想，作为破案工作的起点。所以我们说，统计方法在研究自然界和人类社会的规律性方面，是起着积极的、主动的作用。科学史上有大量这样的例子。下面我们以遗传学上的一项伟大发现为例，在这问题上再多说几句。

奥地利生物学家孟德尔在1865年发表了一篇文章，其中事实上提出了基因的学说（“基因”一词是英国学者贝特松在1909年提出的），从而奠定了现代遗传学的基础。他这项伟大发现的过程很足以说明统计方法在科学研究中所起的作用。孟德尔是用豌豆作试验，这种豆的果实有黄、绿两种颜色，孟德尔分别培养了一个黄色的纯系和一个绿色的纯系，其每一代所结的豆子分别全部是黄色的和全部是绿色的，孟德

尔然后将这两个纯系进行杂交，发现这种黄-绿杂交品种所结的豆子全部是黄色的，与黄色纯系无不同。但在将这种杂交体再进行一次杂交而产生第二代时，孟德尔发现某些这种"第二代杂交豆子"呈黄色，而另一些呈绿色，其数目的比例大致接近3:1。孟德尔把他的试验重复了多次，每次都得到类似的结果，到这里为止，所得到的还只是一个表面上的统计规律性，但这个表面上的规律性启发了孟德尔去发展一种理论，以解释这个现象。他假定存在一种现在称之为基因的实体以控制豆子的颜色。这实体有两个状态 y（黄）和 g（绿），共组成四种配合：yy，yg，gy，gg（称为基因型）。前三种配合使豆子呈黄色，而第四种配合使豆子呈绿色（在遗传学上称 y 为显性的而 g 为隐性的）。根据这个学说，孟德尔就容易给他的试验结果以圆满的解释：黄色纯系和绿色纯系的基因型分别是 yy 和 gg。杂交第一代种子的基因型则只有一个可能性，即 yg，而根据 y 为显性的假设，具有这个基因型的豆子呈黄色，在外观上与 yy 无异，但若对 yg 再行杂交，则呈现四种可能性，即 yy，yg，gy 和 gg。前三种是黄色而后一种是绿色，这解释了杂交第二代豆子中颜色黄绿之比近似为 3:1 的观察结果。为什么只是近似 3:1 而非严格 3:1 呢？这好比有两个极大的盒子，每个盒子中放入为数极多的黑白两种颜色的球，每盒中两种颜色的球的个数相同，然后你每次从两盒中各抽出一球配成一对，这样重复多次，得出许多个（个数很大，但比起合中所有的球的个数则很小）对子，在这许多个对子中，"黑-黑"对子的个数只是接近全部对子数的四分之一，而不见得恰好是四分之一。自然，孟德尔理论的伟大意义不是在于它给这个特殊的观察结果提供了理论解释，而是在于，用这个理论（当然是经过大大发展了的）可以解释生物体的很多遗传现象，从而形成了遗传学中的基因学派，到本世纪50年代，基因的存在

已经在分子水平上获得了证实。关于统计方法在建立孟德尔理论的过程中所起的作用，我们还可以补充一点：在从分子的水平上观察到基因的存在因而完全证实这个理论以前，曾经用统计方法对依这个理论推出的大量结论进行过检验，检验的结果都证实了这个理论与观察结果符合（这个问题在后面第七章中还要讨论）。这本身就是统计方法在科学上的一项重要应用——用于客观地评价某项理论上的结论是否与观察结果相符，以作为该理论是否站得住脚的印证。

3. 由部分推断整体、总体和样本

统计方法都具有**部分推断整体**这个性质。如在例2中，**整体**就是全县的所有个体农户——由于我们只关心其化肥使用量，也可以说整体是由该县所有个体农户每户化肥用量组成。若此县有10万个个体农户，则整体包含10万个数字，所要考察的问题（化肥用量户均值）是关系到这个整体，而不是关于其中某些户的。**部分**就是被抽选出的那些农户（也可以说，是抽选出的这些农户化肥用量的全部数据）。我们的方法是算出这"部分"的平均值。如果停留在此地，则所得结果还只与这个"部分"有关。若再往前跨一步，而声称"以这部分的平均去估计整体的平均"，则我们工作的意义越出了这部分之外而达到整体。这一步工作称为统计推断，它是关键的一步，构成统计方法的一个重要特点。举一反三，读者不难按这个精神，对其他三例作类似的分析。

为什么要把这个强调为统计方法的一大特点呢？原因有二：一是它把统计方法与其他数学方法区别开来，一是它把大量日常工作以至生活中与数字打交道的工作，和统计方法

区别开来。

先说第一点。统计方法要用到许多数学工具，尽管在学者中对统计学是否可算作数学的一个分支存在分歧，但对于统计方法中使用大量数学工具、统计方法的原理依靠高深数学的论证这些事实，却不容抹煞。那么，相对其它数学方法而言，统计方法的特征何在？关键就在"部分推断整体"这一点上。举一个极简单的例子：有两块矩形木板 A 和 B，要比较其面积谁大，大多少？量得 A 的长宽分别为 1.52 米和 1.425 米，B 为 1.79 米和 1.21 米，如果测量绝对准确，则根据矩形面积＝长×宽这个公认的数学公式，即算出 A 的面积比 B 大，大 0.0001 米^2。这个问题用数学方法解决了，但不是用的统计方法，因为你已掌握了与问题有关的全部资料，不存在"部分推断整体"的因素。然而，你可能觉得测量有一定误差，而二者面积测量值之差（0.0001）又很小，只测一次就下结论未必可靠，为了增加可靠性，你把 A、B 的长宽各量 100 次，算出 A，B 面积百次测量结果的平均值之差，以此为准来定何者面积大，大多少，这就是一个典型的统计方法。为什么？就在于你只掌握了与问题有关的部分信息而非全体。因为，你既可以量 100 次，又何尝不可再量到 200 次，300 次，……直观上告诉我们，测量次数愈多，平均数愈可靠。理论上说，要"绝对"可靠，只有测量无穷多次求平均。设想你真这么做了（当然事实上不可能），你就掌握了问题的全部信息。因此在本问题中，"无穷次测量结果的全部记录"构成一个整体，你实际作了的那 100 次测量只是这个整体中的一部分，这仍是一个由部分推断整体的格局。

再说第二点，若在例 2 中问题不是一个县而是一个村，则我们大可不必从其中挑出一部分农户，而可以逐户搞清楚，算出其平均就可以了。从严格的统计学观点说，这里谈不上用

到了什么统计方法，只是例行公事地作一些加法和除法的运算，就得到确实的结果（统计方法由于只用到部分资料，结果不见得确实，即有误差，这误差可能多大，是统计学的任务，这构成统计方法一个特点，将在下文论述）。这类工作很多，虽然在习惯上也无妨承认它们用及统计方法，但这个差别却不可忽视。你把今年家里每个月所花的伙食费都记下来，到年终一平均，就得到你家今年每人每月的平均伙食费。这一切都一目了然，提不上统计方法这个大文章。可是如果你在以往五年中都作了这个计算，分析所得结果，总结出这五年伙食费以年率15%的幅度增加，并进而推断在今后三年内，你家及情况类似的人家，其每人每月平均伙食支出仍将以这个幅度上升，则这整个过程就可以看作是统计方法的使用。因为，所作出的结论越出了你掌握的数据资料的范围，而构成一项统计推断。

在这一段的最后我们介绍几个在统计学上常用的专门术语，并作些补充说明。

统计方法有"部分推断整体"的特征。这个整体在统计学上常称为**总体**，也有叫**母体**的。总体依所研究的问题而定。前面已指出，如在例2中总体由该县的全部个体农户组成。总体里的每一分子称为一个**个体**或**单元**。在例2中，每一农户构成一个个体或单元。如果你要对小学生中的近视眼比率作调查，则随着你研究规模的不同，总体可以是全国所有的小学生，或是人口在20万以上的城市中所有的小学生，或者是指定城市中的全部小学生。总体取得不同，研究结果适用的范围当然也就有别。

从总体中抽选出的那部分个体，统计学上称之为**样本**，也有叫**子样**的。如在例2中，抽选出的那400户就构成样本。有时也把样本中单个或一部分个体称为样本。样本中所含的

个体数，在统计上称为**样本大小**，也有叫**样本容量**的，例 2 中的样本大小为 400。从总体中抽选出样本的过程叫**抽样**，也有叫**取样**的。不论在任何问题中，由于与问题有关的往往只是个体的某项（或某几项）指标，也可把个体的指标值就说成是该个体。这时，不论总体和样本，都是由一些数字构成（这在本节开头已就例 2 说明过）。这个看法突出了与问题有关的数量方面，便于在数学上作统一的处理。

把某一个体算作是所研究问题的总体之内，有一个明显的前提，即该个体的指标值（这项指标是问题中所关心的）必须是可知的，必要时得加以明确的规定。如在例 4 中，每一个体（一个人）的指标值，就是他属于四类中的哪一类（见例 4 开头的说明），若你没有必要的医学设备，就无法检定一个人是否患肺癌，则每一个人所属的类别就无法确定，研究工作也就无从着手了。有时，一个体的指标值如何，需要根据一定的规则才能定下来。比方说，某甲在今日确诊患有肺癌，但他是一星期前才开始吸烟的，难于设想，这一段吸烟史与某甲患此病有关，故某甲的指标值似以定为“不吸烟，患肺癌”为合理。故这里存在一个“怎样的人算作吸烟者”的问题，这不见得是很容易解决的问题。在例 2 中，要求该县每户化肥用量都可知。比方说，以该户户主所报数目为准，即使有些误差也不计较，当然，若所报数目很不准，或对其误差性质有些了解，可以在问题中把这种误差考虑进来，这时总体和个体就不能像原来那样子，而是大大复杂化了，这一点在此不能细论。

在有些问题中，总体是由一些看得见、摸得着的个体构成的。例 2 是一个典型例子。在另一些问题中则不然，它只存在我们的想像中。例 3 是这类情况的一个代表，可能会认为，在本例中，总体就是由这个（其重量 a 未知）物体构成，其实不

试读结束：需要全本请在线购买： www.ertongbook.