数学概貌丛书

统计学概貌

陈希孺 著

科学技术文献出版社

数学概貌丛书

统 计 学 概 貌

陈希孺 著

科学技术文献出版社

内 容 提 要

本书对统计学的基本面貌作了概括性的介绍. 全书分六章. 第一章围绕"统计学是讨论以有效的方式收集和分析带随机性的数据"这个命题,通过大量实例,阐明了统计学的学科性质、任务和主要问题. 以下三章是这一章的展开,分别讨论数据的收集、整理及统计推断. 第五章介绍统计学在各方面的应用,而最后一章则简略地回顾了统计学发展的历史和当前的发展趋势. 书中对、西方关于统计学的不同理解以及统计学与数学的关系也有所说明.

本书基本上是通俗性的,只用到中学数学,适合于广大关心这门学科的读者阅读.

数学概貌丛书

统计学概貌

陈希孺 著

*

科学技术文献出版社出版

(北京市复兴路15号)

上海市新华印刷厂印刷

新华书店上海发行所发行 各地新华书店经售

*

开本850×1156 1/32 印张2.5 字数47,000

1989年3月第1版 1989年3月第1次印刷

印数: 1—5,000本

定价: 1.05元

ISBN 7-5023-0663-3/O·53

目　录

一、什么是统计学 …… 1
二、抽样调查与试验设计 ……18
　1．抽样调查……19
　2．试验设计……26
三、数据的整理 ……39
　1．一维数据的重要统计量……41
　2．样本协方差，样本相关系数与回归 …44
四、统计推断 ……49
五、统计学的应用 ……56
六、简单的历史与现状 ……63

本书所说的“统计学”，在我国习惯上常称之为“数理统计学”.对这些名词，存在着不尽相同的理解，这些将在本书的第 1 节中加以说明.

本书试图对统计学的学科性质、基本内容和发展历史，作一简要介绍.作为一本通俗性的书，又受篇幅和所使用的数学工具的限制，这个介绍在理论方面难于达到充分的严谨、全面和系统化，还请读者见谅.

本书前三节在数学上要求很少，一般具备高中程度数学知识的人，都可以读懂.第 4 节则要求读者具有一定的概率论知识，不过其中的基本思想，在第 1 节中已有所交代.最后一节，因系讲述历史，不能不提到某些在前几节中未充分解释的概念，不具备统计学初步知识的读者，浏览一过，大致有所了解就行了.

一、什么是统计学

什么是统计学？什么是数理统计学？这个问题，难于用简短的语言作一个明确、严谨而全面的回答.我们打算

先用通俗的语言作一个概括的论述，然后在适当的地方加以补充、解释，以使读者对此问题有一个比较清楚的理解.

实践是认识的来源. 所以,在研究一个问题时,往往首先要收集必须的资料.比方说,少年犯罪与家庭情况的关系如何?跑步对中、老年人的健康是否有益? 或更细致一些,对各种年龄的人,在什么时间,以怎样的速度,跑多长的距离为好?受教育时间的长短与其收入的关系如何?吸烟是否增加患肺癌的机会?增加多少?对一种产品的制造工艺作一些改变,是否有助于改善其质量，改善多少?凡此种种,都是很有现实意义的问题. 这类问题中,有的前人已作过一些研究,提出过种种见解.但前人的研究可能是在不同的条件下进行的(例如,对不同的国家，受教育时间长短与收入的关系就有不同),有的还受到当时科技发展水平和其他因素的局限(例如,某项研究由于经费的关系,收集资料的规模很小)，因此他们的结论不能照搬.如果所研究的是一个前人未接触过的新问题,那当然更不用说,收集资料这步工作是必不可少的.

收集资料的方法有两种:观察和试验.这两个词的含义的差别在于,在"观察"时,观察者可以说是处在被动的地位,他只是对所感兴趣的事物,记录下"自然而然地"发生的结果,而不去企图改变他所观察的事物.天文观察是一个典型的例子.在吸烟与患肺癌的关系的问题中,情况也是如此:你可以观察一个人是否吸烟,吸多少，观察他是否患肺癌;但你不能也不会去设法改变他的状况,这是

因为一个从不吸烟的人，不会应你的研究工作的要求而去吸烟．而在“试验”中，试验者则处在主动的地位，可在一定范围内自由地控制某些因素，以考察它们对其他因素的作用．典型的例子如在工业试验中，工艺参数如何取，原料配方如何选择，出自试验者的主动，以考察它们对产品产量和质量的影响．

从统计学的眼光看，观察和试验都是收集资料的方法．因此，许多统计学著作混用这两个词．但也应注意到，有些统计方法的合理使用和解释，与资料是来自观察还是试验有关．

在不少情况下，收集的资料可以用数量的形式表达．如一个人的月收入以人民币多少元计，是一个数字．有时需要研究事物的若干个方面，则资料可以用若干个数字（即一个向量）表达．如同时观察一个人的身高和体重，结果是一个二维向量．也有些情况，观察或试验所得只是事物所属的等级、类别等．例如观察一人的血型，结果为A、B、AB、O四类中之一；对一种酒品尝结果，列入甲、乙、丙三等之一．这些，在必要时可以进行“数量化”．例如，约定把A、B、AB、O四种血型分别给以数字1、2、3、4．因此，在统计工作中，习惯上把所收集来的资料称为“数据”，或者用“样本”这个专门术语，意思都是一样的．

但是，认识并不是实践的直接产物．为研究一个问题而收集的资料，一般是一大堆杂乱无章的数字，从中看不出什么道理来．比如说，为研究吸烟与患肺癌的关系，观

察了 5,000 人，逐一记下每人日抽烟多少支，抽烟史多长，是否肺癌患者，患病多久等资料，订成一大厚册；泛泛翻阅这本册子，得不出多少东西．因此，需要把数据加以整理，从其中提取出与所研究的问题最有关的信息，并以简明醒目的方式表达出来．例如，一种可能的整理方式如下：把所观察的 5,000 人，按“不吸烟”、“每天吸 10 支以下”、“每天吸 10～20 支”、“每天吸 20 支以上”分组，从收集的资料中算出各组肺癌的发病率，并画成一张图；则很易看出肺癌发病率随吸烟量增加而上升的趋势，以及这个趋势的大小的概念．再举一个例子：为考察毕业后工作了 10 年至 30 年的大学生的工资收入状况，在这类人中抽取了 10,000 名进行观察，记下每人目前月工资数，得 $x_1, x_2, \cdots, x_{10000}$ 等 10,000 个数据．计算其算术平均，即

$$\bar{x} = (x_1 + \cdots + x_{10000}) \div 10000,$$

就可以对这批人的收入的总的状况，或平均状况，有一个了解．如果要进一步了解收入参差不齐的情况，需引入另外的指标，例如

$$s^2 = [(x_1 - \bar{x})^2 + \cdots + (x_{10000} - \bar{x})^2] \div 10000.$$

s^2 愈大，参差不齐的程度就愈大．这个指标 s^2 能给我们一些启示：s^2 太小，说明收入没有适当拉开档次，可能与平均主义有关；s^2 太大，则说明资历较浅的工作者收入太低，可能是反映了某种问题．因此，通过整理数据得到的这两个指标 $\bar{x}$ 和 s^2，以易于理解的方式告诉了我们不少东西（或者说，以简明的方式总结了这一大批资料的信

息）. 如果想了解得更细致些，可以用一定间隔作单位（如1年，5年等），算出在这10,000人中，毕业年限在此间隔内的人，目前的月平均工资，用列表或画图的方式给出结果. 自然，随着所研究的问题的不同、数据形式的不同，整理的方法也会有差别. 可以说，在什么情况下该用怎样有效的方式去整理数据，属于统计学的研究范围.

按一定的方式整理数据，往往也就构成对数据的一种分析. 例如，分析上例中的数据，可得出：毕业后工作年限每长一年，平均月工资增长多少. 在吸烟与患肺癌关系的例中，分析所收集的数据，可以知道日吸烟支数每增加5支，或吸烟史每增加5年，肺癌的发病率增加多少. 但无论对数据进行整理或分析，都没有越出所得数据的范围. 就是说，分析所得的结果只对现有这批数据有效. 就上例来说，设想分析结果是：大学毕业后每多工作一年，平均月工资多2元. 这结果只是针对所抽查的这10,000人来说的. 在我国，大学毕业后工作10～30年的，何止1万人. 而我们真正感兴趣的，是这些工作人员的全体，而不止于所抽出作调查的这一小部分. 这样，我们就需要向前跨一大步：基于所收集到的数据及对它进行整理分析的结果，对数据所来自的总体的有关情况，作出一定的论断. 这种论断叫做“统计推断”. 其具体形式，依问题中要求的不同而不同. 所谓“数据所来自的总体”，就是指与所研究的问题有关的所有个体的集合. 如在上例中，总体就是我国目前尚在的、大学毕业后工作了

10～30 年的全体工作人员. 若这样的工作人员有二百万，则本问题的总体中包含二百万个体. 被抽出作调查的那一万个体(即样本)，是其一部分. 由于本问题中我们关心的不是具体的人而是其月工资，也可以说总体和样本分别由二百万个数和一万个数构成（其中可以有重复的). 这实际上就构成了一种抽象，且是很重要的抽象. 因为这使我们可以摆脱总体及样本的具体属性，便于运用数学的方法，对不同的具体问题作出统一的处理方法.

如上文所述，统计推断的对象，是总体的有关情况，即因我们的研究目的而对之感兴趣的那些情况. 在上例中，我们感兴趣的可以是总体平均值——即总体中二百万个工作人员的月平均工资 a，如果所抽查的那一万名工作人员的月平均工资为 $\bar{x}=90$(元)，那么未知的 a 是否就等于 90？当然不一定.但也易了解，这二者之间会有些关系.关系的大小和性质，取决于这一万名工作人员是如何抽得的.取决于从总体中所抽出的个体的数目(在此为 10,000)，它在统计学上称为“样本大小”——不难明白，样本大小愈大，$\bar{x}$ 与 a 一般就应愈接近. 此外，还与总体的数学性质，即所谓数学模型有关. 这一点留待后面再作解释.

统计推断的具体形式，依问题的要求而异.就此例而言，所要求的可能就是：根据样本，对总体平均值 a 作一估计. 这种问题称为估计问题，是在理论上研究得最深入，在应用上最常见最重要的一类统计推断问题.总体平

均值 a 刻划了总体一个方面的性质，它称为总体的“参数”. 因此，在统计学中，像估计总体平均 a 的这类问题，常称为“参数估计问题”. 直观上觉得，应当用样本平均值 $\bar{x}$(在本例为 90)去估计 a. 这个方法，即按样本算出的值去估计总体的相应值，是一个重要而常用的估计方法.

当我们声明，采用 $\bar{x}$ 去估计 a 时，我们就作出了一种统计推断.因为在这样做时，我们已越出了所掌握的样本的范围，而论及了样本所来自的总体.有的读者可能会问：这样一个看来似乎是纯粹形式上的步骤，能有多大的意义？其实不然. 跨出这一步是不简单的. 比方说，“用 $\bar{x}$ 估计 a”会有误差，这误差有多大，用怎样的形式表达出来，需要用到以深刻的数学理论为基础的统计方法. 又如，用 $\bar{x}$ 估计 a 也并非理所当然的，唯一可行的方法. 兹举一种可以设想的估计方法如下：把所得的 10,000 个数据按大小排序，取出居于正中的那两个，设为 x' 和 x''(若数据个数为奇数，则只有一个恰居正中，就以之代替下文的 x^*)，取其平均

$$x^* = \frac{1}{2}(x' + x'')$$

去估计 a. x^* 称为样本的“中位数”. 用 x^* 估计 a，在直观上也是讲得通的. $\bar{x}$ 和 x^*这两个估计那个较好？在什么意义下较好？这是深刻的理论问题. 除此而外，还可设想出其他种种在直观上看来也合理的估计方法. 我们需要证明：$\bar{x}$ 这个估计在理论上具有某种优良性，这

样，用 $\bar{x}$ 估计 a 才有坚实的基础. 彻底解决这个问题，牵涉统计学上多方面的理论问题. 由此可见，正如我们所曾指出的，跨出这一步并非易事.

再举一例. 在研究吸烟与患肺癌的关系问题时，我们首先感兴趣的，可能是一个初步的问题：这二者到底是否有关，而暂不计其关系的深浅与其确切性质. 这问题可以较具体地解释如下：设如前所述，我们观察了 5,000 人，记录了各人是否吸烟与是否患肺癌的情况，经对数据作初步整理分析，觉得二者似有些关系. 但是，由于这 5,000 人只是地球上的成年人的很少一部分，仅凭这 5,000 人的数据而推及地球上全体成年人，有多大的可靠性？或更清楚地说，你在这 5,000 人中分析出的关系，是纯出于偶然性呢，还是确实反映了一种适用于全体成年人的规律性. 这个问题与上例中估计 a 的问题不同，它只要求回答一个“是”或“否”（是纯出于偶然，或否）. 这种问题在统计学上称为“假设检验”问题. 它与参数估计，并列为统计学中两类基本推断问题，在理论上有深入发展且有重要应用. 名称的由来，是因为在统计学上处理这类问题时，先引进一个有关的假设. 如在本例中，引进“吸烟与患肺癌无关”这个假设. 然后，用样本去“检验”这假设是否成立. 具体说，我们通过分析所掌握的数据，看二者关系的大小如何：若关系不大，则不能排斥它是来自抽样的偶然性，因而断言二者有关的理由不充足，这导致我们接受上述假设；反之，若关系甚大，则仅以偶然性去解释是勉强的，因而有足够的理由断言二者有关，

这导致我们接受上述假设. 这里,"关系大小"如何刻划,导致接受或否定假设的界限如何划分, 都需要统计学的理论.除上述两种以外,还有许多形式更复杂的统计推断问题,需要以深刻理论为背景的不同处理方法.

由此可见,撇开收集数据的问题不谈,统计学的中心问题,或者说其主要内容,就是统计推断. 统计学之所以有如此广泛的应用, 正因为在数学上成功地发展了一整套有关的理论,并在其基础上,制定出了针对一些常见的重要问题的统计推断方法.就是收集数据的问题,也在一定程度上与统计推断的理论和方法有关. 因为只有当数据的结构(这取决于用怎样的方式去收集数据)满足一定的条件时, 才能对它运用适当的统计推断方法. 不然的话,所收集的数据就不好处理.

以上在谈论统计推断问题时, 我们是从一种科学研究的眼光去看待它,即它是以弄清事实为目的,不计较什么利害关系.有一类问题,通称为"统计决策问题",或"统计判决问题",与统计推断问题有关但又有差异. 有关的地方是:统计决策问题的解决也要基于收集的数据,并使用统计推断理论中提供的种种方法.不同之处在于:决策(也常称为判决,或行动)要产生经济上的后果*). 在实际作出决策时,不仅要考虑到统计推断上的结果,还必须把经济上可能的后果结合进来.例如,有一批产品包含很多

---

*) 自然,决策的后果不必限于经济方面,但在统计决策理论中, 只考虑那种问题,其后果可以用一定方法归结为经济上的得失.

件，要估计该批产品的废品率 p，则可以在该批产品中抽取若干个作检查，以样品中的废品率 $\hat{p}$ 去估计 p 即可．但是，如果这批产品是工厂对商店的供货，而商店经理要决定是否接收这批货，则问题并非简单地估计废品率 p 即可．因为，接收或拒收该批产品，都有经济上的后果．例如，若拒收，则当日无该货可出售，要损失利润；但如接收这批货，则有可能废品率 p 较大，而得不偿失．该经理作出的决策，除了考虑到 p 的估计值 $\hat{p}$ 以外，还须把每件废品的损失和出售每件合格品的利润结合考虑进来．另举一例：某工厂的设计试验部门，通过适当安排的试验并使用一定的统计推断方法，搞清楚了原料配方与产品性能之间的关系．但不同的配方涉及成本、原材料来源(这与运输费用也有关系)与消费者喜好，即市场前途问题，在最后作决策(即选用一种确定的配方用于生产)时，统计推断的结果自然是重要的．这只有在统计学家、专业人员、经济师和市场分析人员的共同参与下，才能作出适当的决策——当然，这中间涉及到的问题并非全是统计性的．

到这里，我们已说明了统计学是干什么事的．现在把它小结一下，而对统计学提出一个比较完整的定义：统计学是一门科学，它研究怎样以有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议．在这个定义中，有两点在上文未作仔细交代：一是“有效的方式”一语的含义，这涉及在收集数据的工作中具体的

作法问题．这个重要问题将在下文第二节作仔细论述．二是“带随机性的数据”一语的含义，对概率论初步知识略知一二的读者，自然明白其意义，下文我们还将略加解释．

此处引进的统计学定义，是依照《中国大百科全书·数学卷》中对“数理统计学”所下的定义，这个定义与《不列颠百科全书》上关于“统计学”的说法，基本精神也是一致的．后者把统计学定义为收集和分析数据的艺术．这个定义嫌过于简略一些．不过，其中“分析”一词兼有我们定义中整理、分析、推断的含义．它没有明确指出数据应带随机性，这是一个弱点(见下文)．至于此定义中称统计学是“艺术”，尽管有其不够严谨之处，却也有独到的地方：它提醒人们，统计学并不是一堆在应用时可以机械地照搬的公式，而是在应用上要发挥灵活性以至灵感，需要积累充分的经验．

按这个定义，统计学是一门与数字打交道的学科．在这个意义上，可以把它看成是数学的一个分支．它当然不是社会科学．还有一点要着重说明：像这样定义的统计学，在我国常称为“数理统计学”．而在西方，“统计学”和“数理统计学”有明确的区别，即数理统计学是统计学的数学理论那一部分．所以，在我国，数理统计学等于西方的统计学加数理统计学．其所以产生这个差别，与苏联对这个问题的看法有关．在苏联，把统计学定义为一门研究大量社会现象的社会科学，有很强的阶级性和党性；而数理统计学则被看成是在统计学中使用的数学方法及其理

论基础.这个看法对我国至今仍有很大的影响. 因此,在我国至今仍有不少人采取这样的看法:统计学是一门社会科学,数理统计学则是一门数学学科.

作者不打算在此对上述观点之间的分歧发表评论. 然而,读者不难看出,本书是按照西方的观点来写的. 对此持异议的读者可以这样看待本书:它讨论了统计学与数学有关的那一部分.

按我们所讲的方式去理解统计学,自然地得出它的一个特点:它是通过事物的外在的数量上的表现,去揭示事物可能存在的规律性.它不能确认和解释,为什么事物会存在这样或那样的规律性,后者要依靠有关专门学科的研究. 不过,在探求这种规律性的解释的研究工作中,统计方法也有其作用.例如,用种种统计方法对一些统计资料进行分析的结果,都使人相信吸烟者中患肺癌的比率较高. 但是,究竟吸烟是引发肺癌的一个原因,还是这二者都受到同一遗传基因的控制?如果是后者,则统计资料分析的结果只是表明这二者有一种先天的联系,而不表明这二者有因果关系.要确定这种因果关系的存在,需要从医学上弄清吸烟引发肺癌的机制问题.

统计方法的这个特点,划清了统计学和其他学科的界线.例如,经济学、人口学、社会学、工程学、生物学……等学科,都用到统计学提供的方法.但统计学在这些学科中,只起着一个辅助性质的作用. 统计学自有其研究对象,即超脱了具体含义的数据的收集和分析问题. 当然,统计方法的这种辅助性质并不降低它的意义,恰恰相反,

由于事物的本质规律性往往隐藏很深，不易为人们所察觉，而其外在数量上的表现则易于引起人们的注意，以此，统计方法在揭示事物规律性的过程中，常能起到先导的作用.

按照上述观点，可以说统计方法是一种数学方法.在为数众多的数学方法中，统计方法有什么特点呢？因为，如果把统计学说成是一种处理数据的数学方法，那末，它与算术，一般讲与计算数学，就划不清界线. 这里就要用到前面给统计学下定义时所加的那个限制词：随机性.统计学是处理带随机性的数据的问题. 所谓随机性（又称偶然性），是“随机会而定”的意思. 从实际应用的角度去看，统计学中考虑的数据随机性有两种形式.一种形式的例子是前面提到的吸烟与肺癌关系问题，以及大学毕业后工作 10～30 年的人员的收入问题. 在这些例子中，总体是由一些实在的个体（在此两例是人）组成. 数据的随机性来源於，那些个体被抽出（以组成样本），是随机会而定. 举一个极端的例子. 如果碰巧在你抽出的那 10,000 人中，大多数都是工龄短而工资高，或工龄长而工资低的人，则你会得出“工作年限愈长，收入愈少”的结论. 虽则“碰巧”出现这类情况的机会不大，但既是抽查，你在逻辑上就不能绝对否定其可能性.由此也可以看到，统计推断有产生错误的可能.事实上，统计推断理论中的一个重要课题，就是计算在种种情况下，各种推断方法可靠的程度如何.

大体上说，这种随机性是与“观察”联系在一起的.另