绪 论

什么是统计学?《不列颠百科全书》上定义:“统计学是关于收集和分析数据的科学和艺术.”

统计学是一门关于数据的学科,它涉及数据的收集、处理、分析和解释,旨在通过数据解决实际问题.这门学科利用数学模型,特别是以概率论为理论基础,来分析和解释受到随机因素影响的数据,从而作出推断和预测.统计学不仅是一门科学,而且是一门艺术,它涉及如何有效地收集和解释数据,以便为决策提供支持和参考.

统计学家弗朗西斯·高尔顿说:“统计学处理各种复杂现象的能力是非凡的,它是追求科学的人从荆棘丛生的困难阻挡中开辟道路的最好工具.”统计学的应用范围广泛,从社会、经济、管理、教育到自然科学、工程技术和医疗卫生,以及工商业和政府决策,几乎涵盖了所有领域,各种事物所具有的内在数量规律性都可以借助统计方法加以探索.


0.1

0.1 统计学的产生和发展

0.1.1 统计学发展史

统计活动源远流长,人类社会有了数的概念,统计就开始了.但统计学作为一门独立的学科,多数人认为,大概兴起于17世纪,其发展大致经历了17世纪中叶至19世纪初的古典统计学萌芽时期、19世纪初至20世纪初的近代统计学形成时期和20世纪以来的现代统计学发展时期三个阶段.

17世纪中叶,欧洲各国为了适应经济发展的不同需要,从不同领域开始了统计学的奠基工作,相继形成了统计学的两大来源:国势学派和政治算术学派.国势学派,又称记述学派,产生于17世纪的德国,创始人是赫尔曼·康令.该学派主要用文字记述国家的显著事项来说明管理国家的方法;特点是偏重事物质的解释而不注重数量的分析,它为统计学的发展奠定了经济理论基础.另一代表人物戈特弗里德·阿亨瓦尔提出“统计学”(statistik)一词,并定义其为国家显著事项的学问,转译成英文statistics,为人们接受沿用至今.政治算术学派,产生于17世纪中叶的英国,代表人物是威廉·配第,1676年他的代表作《政治算术》的问世,标志着统计学的诞生.《政治算术》是一部用数量方法研究社会问题的著作.配第用“数字、重量和尺度”研究的方法为统计学的产生与发展奠定了方法论基础.另一位代表人物约翰·格朗特,1662年他发表了《关于死亡公报的自然和政治观察》,首次提出通过大量观察,可以发现新生儿性别比例具有稳定性等人口规律,被认为是人口统计学的创始人.

19世纪初至20世纪初是近代统计学形成时期，这一时期建设和完善了统计学的理论体系，并逐渐形成了以传统政治、经济现象描述为主要内容的社会统计学和以随机现象的推断统计为主要内容的数理统计学两大学派.社会统计学派产生于19世纪后半叶的德国，首倡者是克尼斯，主要代表人物还有恩格尔和梅尔.他们认为统计学是一门社会科学，是研究社会现象变动原因和规律性的实质性学科；强调统计研究必须以事物的质为前提和认识事物质的重要性，研究方法采用大量观察法.比利时人雅克·凯特莱认为，统计学既研究社会现象又研究自然现象，是一门独立的方法论科学.他把概率论引入统计学，根据大数定律，利用统计观察资料研究随机现象的数量规律性，开创了统计理论和实际应用的新领域，促进了数量研究由“算术水平”向“数理”阶段迅速转化，为数理统计学的形成和发展奠定了基础.凯特莱是承前启后的重要人物，按其贡献可以认为他是古典统计学的完成者、近代统计学的先驱，也是数理统计学派的奠基人.

19世纪末以来，欧洲自然科学的飞跃发展，促进了数理统计学的发展.进化论和能量守恒定律的出现促进了描述统计的完善，描述统计学派发展到顶峰.描述统计学主要研究资料的系统收集、整理、表述和计算，是以弗朗西斯·高尔顿为先导，以卡尔·皮尔逊为代表的用于对生物资料进行分析提出的一系列统计方法.以大量观察和正态分布为基础关于总体分布曲线的研究，确立了“大样本”统计理论，奠定了“描述统计学”的体系.20世纪20年代以后，在细胞学的发展推动下，统计学迈进推断统计的新阶段.推断统计学研究如何根据部分观察资料对总体情况作出具有一定可靠性的推断.20世纪50年代，是推断统计学发展最迅速的时期.这期间有影响的理论和大师很多，如世纪初的威廉·戈赛特(笔名Student)提出t分布理论，他提出的小样本理论成为统计推断思想的一块基石；20世纪20年代，艾尔默·费歇尔提出F分布理论，他在抽样分布、方差分析、试验设计等方面都有卓越的建树，成为推断统计学的真正创立者；20世纪30年代的乔治·奈曼、埃贡·皮尔逊等提出区间估计及假设检验等理论.20世纪50年代，经过几代大师的努力，推断统计的基本框架已经建成，并逐渐成为20世纪的主流统计学.

自20世纪五六十年代以来，统计决策、多元统计、时间序列、贝叶斯统计等都取得了重要进展.统计学发展史简图如图0.1所示.通过简单回顾统计学发展历史，可以看出，随着人们认识的不断深化、社会实践需要的推动，统计学不断地丰富和完善.它经历了从意义和概念不甚明确的阶段，到作为一门独立学科的转化；从数量研究的“算术”水平，到需要较丰富数学知识的“数理”阶段的转化；从大量观察消除误差干扰以达到对客观现象规律认识的大样本理论，到控制实验次数提高数据质量的小样本推断的转化；从社会科学领域的实质性科学到自然科学领域通用性方法论学科的转化.

0.1.2 我国的统计学教育

我国的统计学教育是从20世纪初清朝末年开始的，已有100多年的历史，大致可以分为三个阶段.第一阶段，20世纪初至解放之初，是我国统计学科建立时期.这一时期的特点是学习借鉴欧美统计理论和方法，主要是作为课程在理、工、农、医、商和社会科学等学科专业开设.第二阶段，解放之初到改革开放之初(1951—1978年).这一时期我国统计学科深受苏联的影响，将统计学一分为二，认为概率论与数理统计方法属于数学，社会经济统计属

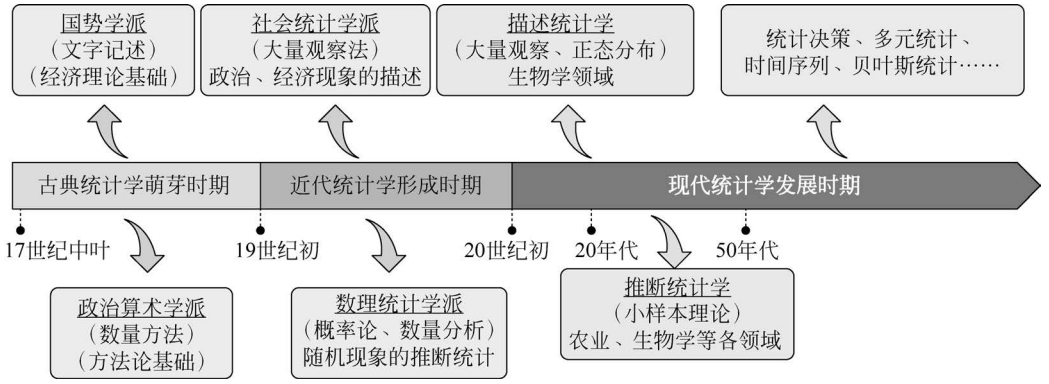

图 0.1　统计学发展史简图

于社会科学.第三阶段,改革开放之初至今,是我国统计学从拨乱反正到“大统计”,再到统计一级学科的建设时期,为追赶国际先进水平打下基础.国际科学界只存在一门统计学(即数理统计学),它是现代各国广泛应用的一门统计科学,也是我国对自然科学和社会、经济科学进行科学研究的一种必要的科学方法、技术(戴世光).目前,统计学学科在本科生层次、研究生层次上成为一级学科,设在理学门类下,既可以授理学学位,也可以授经济学学位;在我国科研科技统计专业目录上、在我国教育专业目录上成为一级学科,在形式上与国际统计学已经接轨,极大地促进国际学术交流和学生国际交流.

我们来了解一下我国著名统计学家的代表——生物统计学家吴定良先生,他师从卡尔·皮尔逊,创造了相关系数计算法和相关率显著性查表.他的突出成就为统计学界所认同,1931 年,吴定良和英国著名统计学家费歇尔等 21 人同时被推选为国际统计学会会员,他是该学会历史上第一位中国会员,成为我国首位有国际影响的统计学家.他既重视统计理论和方法的研究,又强调理论联系实际.1956 年,他用回归分析及其他科学方法成功鉴定了方志敏烈士的遗骸.

在数理统计和概率论方面第一个具有国际声望的中国数学家许宝騄先生,被视为奈曼最优秀的学生,被外国学者称赞为“20 世纪最深刻、最富有创造性的统计学家之一”.1933 年,他在清华大学获得理学学士学位,随后赴英攻读博士学位.博士毕业后,他谢绝了美国多所高校的聘任,回到北京大学担任教授.他在国际统计学界颇负盛名,他的肖像和国际其他著名统计学家的照片一起悬挂在斯坦福大学统计系的走廊上.许宝騄先生在北京大学设立了国内第一个统计学学习班,培养了一大批有影响力的统计人才.他研究兴趣广泛,涉及统计学的各主要领域,包括次序统计量、参数估计、矩阵微分和假设检验等.他最先发现线性假设的似然比检验(F 检验)的优良性,给出了多元统计中若干重要分布的推导,推动了矩阵论在多元统计中的应用.他的研究方法后来被称为“许方法”.在他的引导下,我国统计学各方向涌现出很多研究成果.他自幼体弱,后来又感染肺结核,但一直带病坚持工作,开展教学和科研活动.1970 年,许宝騄先生病逝于北京大学,但床头依然摆放着旧的钢笔和未完成的手稿.许宝騄先生热爱祖国,刻苦勤奋,取得了一系列原创性的理论成果,并将这些成果应用到我国的实际生产与生活中,大大地促进了我国统计学科的发展.

国际著名数理统计学家、中国科学院院士陈希孺先生,一生致力于我国数理统计学的

研究和教育事业，研究领域主要为线性模型、U 统计量、参数估计与非参数密度、回归统计和判据等数理统计学若干分支，并取得了多项重要成果. 他是我国线性回归大样本理论的开拓者，在参数统计及非参数统计领域做了具有国际影响的工作. 他在非参数计量，特别是极重要的 U 统计量的研究中获得 U 统计量分布的非一致收敛速度，具有国际领先水平，被 20 世纪 90 年代国际上几本专著和美国统计科学大百科全书所引述. 陈希孺先生在其《数理统计学简史》的序中说道："统计学不只是一种方法或技术，还含有世界观的成分——它是看待世界上万事万物的一种方法. 我们常讲某事从统计观点看如何如何，指的就是这个意思. 但统计思想也有一个发展过程. 因此统计思想（或观点）的养成，不单需要学习一些具体的知识，还能够以发展的眼光，把这些知识连缀成一个有机的、清晰的途径，获得一种历史的厚重感."

科学成就离不开精神支撑. 一代代统计学家胸怀祖国、服务人民的爱国精神，勇攀高峰、敢为人先的创新精神，追求真理、严谨治学的求实精神和淡泊名利、潜心研究的奉献精神，将激励我们不断攻坚克难、奋力前行.


0.2

0.2 现代统计学的性质和特点

现代统计学已成为与数学和一系列实质性学科互有交叉的综合性、通用性的方法论学科.

统计学有以下特点：①数量性. 统计的语言是数字，统计学是研究数量问题的学问. ②总体性. 数量有个体数量和总体数量之别，统计学主要研究后者，它对大量同类现象的数量方面进行综合反映. ③不确定性. 由于受到偶然随机因素的影响，客观事物的实际数量表现存在一定程度的"不可确知性"，也就是不确定性. 在现代统计学中，处理不确定性问题是统计学的主要课题和任务. ④统计方法有归纳推断的特点. 统计对总体的认识有两条途径：一是全面调查，对构成总体的全部事物逐一进行调查，取得全面资料；二是抽样调查，从总体中抽取部分事物组成样本，然后依据样本观察结果对总体进行推断. 对于前者，运用算术方法和统计描述手段就可达到目的，后者相对比较复杂，需要运用概率论知识和数理统计方法. 实际中，全面调查和非全面调查的抽样调查都会用到，但由于全面调查受到诸多因素的约束，从经济性、时效性、实用性和可行性方面考虑，利用样本资料进行推断的优势比较明显. 统计方法的归纳推断性主要是相对推断统计而言的，同逻辑学意义上的归纳推断有着明显的区别. 统计推断不是从假设、命题出发，按严格的逻辑推理程序进行推断，而是基于观察到的样本情况，对总体的可能情况作出判断.

统计学不是数学，它同数学其他分科相比有其特殊性，首先，统计学有较强的应用背景. 要正确使用统计方法，不仅要有数学基础，而且要懂得相关学科的知识，具备一定的实际经验和良好的判断力. 其次，统计学主要研究不确定性问题. 最后，现代统计学的本质是归纳推断，与数学演绎方式有较大的差别. 统计学与各专门学科存在必然的联系. 这个联系体现为统计学能为各个专门学科中带有普遍性的数据收集、整理、分析和解释活动提供方法与理论指导，帮助它们更精确、更深入地进行认识. 但统计方法只是定量分析的工具而已，不涉及各门学科中的具体问题.

目前，统计学已经发展成包括理论统计学和应用统计学庞大的学科体系. 理论统计学

侧重于从数学学科中汲取营养，研究数学方法和基础原理，以解决统计学自身发展中重点问题为目标；应用统计学从实际问题的背景出发，着重介绍如何使用统计方法. 按应用的学科性质不同，应用统计学分为应用于社会科学的应用统计学和应用于理工科的应用统计学. 依统计方法的构成，现代统计学分为描述统计和推断统计，描述统计是资料的系统收集、整理、表述和计算，是统计学的基础；推断统计由样本数据推断总体数量特征，是现代统计学的核心. 本书第 1 篇介绍统计推断，主要包括参数估计和假设检验；第 2 篇介绍线性统计模型，主要包括方差分析和回归分析模型；第 3 篇介绍多元统计分析，主要包括聚类分析、判别分析、主成分分析和因子分析.

应用统计方法需注意哪些问题呢？统计学的研究对象是数据，首先，我们要保证数据资料的准确性和有效性. 其次，统计学是一门科学，在应用统计方法前要清楚统计方法的适应对象和条件. 针对不同的问题、不同的资料，有选择地运用不同的处理方法. 统计方法用样本资料进行推断，并不总是保证不犯错误. 因此，对统计分析结果应有正确的认识，要给出合理的解释. 最后，统计学是一门艺术，要灵活使用统计方法，有时依赖人的判断甚至灵感. 不能以教条式的态度看待统计方法，生搬硬套一些公式和方法.

0.3

第1篇　统计推断

“统计推断”是统计学研究的核心问题，它是根据样本对总体的分布或数字特征等作出合理的推断，并为决策提供科学依据. 统计推断广泛应用于自然科学、社会科学、医学、工程等领域. 英国统计学家费歇尔认为常用的统计推断有三种基本形式：抽样分布、参数估计和假设检验.

第1章　统计学的基本概念及抽样分布

在概率论中，所研究的随机变量、概率分布都是假设已知的，在此前提下研究它的性质、特点和规律性. 但是对一个具体的随机变量来说，如何判断它服从哪种分布？如果知道它服从的分布类型又如何确定它的参数？这是统计学要研究的内容. 看下面的例子.


1.1

例 1.1　某公司要采购一批产品，每件产品可能是正品，也可能是次品，该批产品的次品率为 p，次品率的大小决定了该批产品的质量，它直接影响采购行为的经济效应. 人们会对 p 提出一些问题. 如 p 的大小如何？p 大概落在什么范围内？能否认为满足设定要求 $p<0.05$？

在此类问题中，我们需要对这批次品率未知的产品进行研究. 有些情况下，我们不可能对每件产品进行检测，判断它是正品还是次品，如一些破坏性的试验，或者需要花费大量时间、人力和物力的试验. 我们只能从中随机抽取少量产品进行检测分析，对此批产品的次品率进行推断.

统计学是以概率论为理论基础，根据试验或观察得到的数据来研究随机现象，对研究对象的客观规律性作出合理估计和判断的一门数学学科. 其内容包括：如何收集和整理数据资料，即试验的设计和研究；如何对所得的数据资料进行分析、研究，对研究对象的性质、特点作出推断，即统计推断. 本篇着重讨论统计推断. 本章首先介绍总体、随机样本和统计量等基本概念，其次介绍抽样分布，为后面的学习作准备.


1.2

1.1　统计学的基本概念

1.1.1　总体与样本

我们把研究对象的全体所构成的集合称为总体或母体，将总体中的每一个元素称为个体.

例 1.2　研究某厂生产的一批电子元件的寿命. 这批电子元件寿命指标的全部可能取值就构成一个总体，每个电子元件寿命的取值就是个体.

例 1.3　考察在某种工艺条件下织出的 8000 匹布的疵点数. 这 8000 匹布中每匹布疵点数的全部可能取值就构成一个总体，每匹布各自的疵点数是个体.

在实际中我们所研究的往往是总体中个体的某种数量指标，如电子元件的寿命 X，布匹的疵点数 Y 等，它们都是随机变量. 设随机变量 X 的分布函数为 $F(x)$. 如果我们主要关心的是数量指标 X，为方便起见，可以把这个数量指标 X 的可能取值的全体看作总体，并且

称这一总体为具有分布函数 $F(x)$的总体. 这样就把总体与随机变量联系起来,这种联系也可以推广到 k 维,$k\geqslant 2$. 如某种合金钢的硬度和韧度,某儿童群体的身高和体重等,我们可以把两个指标所构成的二维随机向量(X,Y)可能取值的全体看作一个总体,简称二维总体. 二维随机向量(X,Y)在总体上有联合分布函数 $F(x,y)$,称此总体为具有分布函数 $F(x,y)$的总体.

从总体中选取一些个体进行观测的过程称为抽样. 假如从总体 X 中抽取 n 个个体,这 n 个个体为$(X_1,X_2,\cdots,X_n)$,称其为样本或子样,n 称为样本容量. 在一次抽样后,把样本的观测值$(x_1,x_2,\cdots,x_n)$称为样本值或子样值. 从总体中抽取样本有多种方法,但都希望抽到的样本能很好地代表总体,能对总体作较可靠的推断. 最常用的方法是简单随机抽样法,它采用机会均等原则,满足以下两点:①独立性:样本中各个体 $X_1,X_2,\cdots,X_n$ 之间相互独立;②代表性:样本中各个体 $X_1,X_2,\cdots,X_n$ 与总体 X 同分布.

用简单随机抽样法得到的样本称为简单随机样本,简称样本. 以后所提的样本均为简单随机样本. 简单随机样本是有放回地抽取得到的样本,而在实际工作中,我们的抽样多数是无放回地抽样,从理论上来讲不再是简单随机样本. 由于总体中个体数目很大,抽取一些个体对总体影响不大,可近似看成有放回地抽样,其样本仍可看成独立同分布的. 简单随机抽样有很多好处,除了随机选取排除了观察者的偏见,小样本还能减少成本,小样本的数据质量更容易监控.

设总体 X 的分布函数为 $F(x)$,$(X_1,X_2,\cdots,X_n)$是来自总体 X 的一个样本,则样本的联合分布函数为

$$\begin{aligned}F(x_1,x_2,\cdots,x_n)&=P\{X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_n\leqslant x_n\}\\&\overset{\text{独立性}}{=}P\{X_1\leqslant x_1\}P\{X_2\leqslant x_2\}\cdots P\{X_n\leqslant x_n\}\\&\overset{\text{代表性}}{=}P\{X\leqslant x_1\}P\{X\leqslant x_2\}\cdots P\{X\leqslant x_n\}\\&=F(x_1)F(x_2)\cdots F(x_n)\\&=\prod_{i=1}^{n}F(x_i).\end{aligned}$$

当总体 X 是离散型随机变量,且分布律为 $P\{X=x^{(i)}\}=p(x^{(i)})$,$i=1,2,\cdots$,则样本的联合分布律为

$$\begin{aligned}P\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}&\overset{\text{独立性}}{=}P\{X_1=x_1\}P\{X_2=x_2\}\cdots P\{X_n=x_n\}\\&\overset{\text{代表性}}{=}P\{x=x_1\}P\{x=x_2\}\cdots P\{x=x_n\}\\&=p(x_1)p(x_2)\cdots p(x_n)\\&=\prod_{i=1}^{n}p(x_i),\end{aligned}$$

其中 $x_1,x_2,\cdots,x_n$ 的每一值都在 X 的所有可能取值 $x^{(1)},x^{(2)},\cdots$ 中.

当总体 X 是连续型随机变量,且分布密度函数为 $f(x)$时,则样本的联合分布密度函数为