

高等学校适用教材

应用统计学

刘春雨 鲁统宇 陈育蕾 刘 林 王宏宇 编著



中国质检出版社
中国标准出版社

013048778

C8
235

高等学校适用教材

应用统计学

刘春雨 鲁统宇 陈育蕾 刘 林 王宏宇 编著



中国质检出版社
中国标准出版社

北 京

C8
235

277810810

图书在版编目(CIP)数据

应用统计学/刘春雨等编著. —北京:中国质检出版社, 2013. 6

高等学校适用教材

ISBN 978-7-5026-3775-0

I. ①应… II. ①刘… III. ①应用统计学—高等学校—教材 IV. ①C8

中国版本图书馆 CIP 数据核字 (2013) 第 020970 号

内 容 提 要

本书介绍应用统计学的基本理论和方法,包括绪论、统计数据的收集、统计数据的整理与描述、抽样与抽样分布、参数估计与假设检验、相关分析与回归分析、方差分析与试验设计、时间序列分析及指数分析,并以附录的形式介绍了 Excel 在统计中的应用。

本书可作为高等院校管理类专业的教材,也可供统计人员和相关管理人员学习和参考。

中国质检出版社
中国标准出版社 出版发行

北京市朝阳区和平里西街甲 2 号 (100013)

北京市西城区三里河北街 16 号 (100045)

网址: www.spc.net.cn

总编室: (010)64275323 发行中心: (010)51780235

读者服务部: (010)68523946

中国标准出版社秦皇岛印刷厂印刷

各地新华书店经销

*

开本 787×1092 1/16 印张 22 字数 546 千字

2013 年 6 月第一版 2013 年 6 月第一次印刷

*

定价 38.00 元

如有印装差错 由本社发行中心调换

版权专有 侵权必究

举报电话: (010) 68510107

前 言

统计学是高等学校管理类专业主要课程之一,为适应这门课程各类教学需要,我们曾经出版了几种版本的统计学教材。随着我国社会主义市场经济体制不断发展和完善,加之近年来统计学术界对一些基本理论问题的反思和重新认识,统计学教学和教材也面临革故鼎新的任务。国内已有不少学者认识到现有的统计学教材中存在的问题,进行了很多积极有益的探索,出版了不少新版的统计学教材。本教材的编写是我们在教学过程中结合实践的探索做出的努力和尝试。

按照《不列颠百科全书》的定义,统计学是“收集和分析数据的艺术”。这“数据”是带有随机性的数据,即受统计规律的支配。这一特点使统计学与诸如数学、数据分析之类的学科既有明显的区别又有密切的联系。系统完善的统计知识结构对目前在管理学科大门类下的学习是必须的,尽管各细分专业门类或方向(如质量管理、财务管理、行政管理等)对统计学知识可能有不同的特殊要求。为适应这一需要,我们在内容的选择方面突出重点、方法全面、紧密结合统计工具软件,适应不同类型的专业统计课程需要。

本书不同于数理统计(纯粹的统计理论),也不同于那些把统计学定义为一门研究大量社会现象的社会科学的教材。本书的宗旨是编写研究现象总体(既包括自然现象,也包括社会现象)数量关系的方法论科学著作。采用合适的统计方法研究现象或总体数量之间关系的方法论学科。本书的特点是:(1)全面系统地介绍统计方法,适当引入抽样分布推断内容;(2)密切结合生产和生活的实际;(3)注意吸收国内外基本理论和基本方法的最新成就;(4)文字通俗简练,结构合理,具有较强的可读性。它既可以作为学习统计学的基础理论读物,又可作为统计专业的基础课教材。

全书共分九章,主要内容有绪论、数据收集方法、数据整理与描述、抽样分布、参数估计与假设检验、相关分析与回归分析、方差分析与正交试验设计、时间序列分析及指数分析。本书正文部分由中国计量学院的老师编写,其中第一、第二、第三、第四章由刘春雨老师编写;第五章由刘林老师编写;第六、第七章由鲁统宇老师编写;第八、第九章由陈育蕾老师编写;附录部分由杭州服装职业高中的王宏宇老师编写;习题及答案由各章作者拟出。全书由刘春雨老师负责统稿。陶靖轩教授审阅了全部书稿,给出了很多有益的指导意见。在此表示诚挚的谢意。

鉴于统计运算与大量数据处理的密切联系,本书除了正文每章中结合统计软件 SPSS 实现统计方法外,还在附录中加入了“统计方法在 Excel 上的实现”和常用工具数表,在便于巩固基础统计知识的同时,可以方便地学习和使用各种统计软件 and 数据处理方法,以便更加

有效地进行统计分析。

本书是我们在多年的统计学教学实践过程中，不断总结过去的教学经验和教训基础上写作而成。各位作者除了自己的讲义和心得外，还参考了不少中外统计学基础著作及学术研究成果，在一些统计学术会议上与同行专家的交流也使我们深受启发。在此，对所有给予我们帮助的专家和同仁，表示衷心的感谢！

限于编者的水平，书中谬误之处肯定存在，期盼得到读者宝贵、有益的指教，以便再版时修正。

编著者

2013年6月

目 录

第一章 绪 论	(1)
第一节 统计与统计工作	(1)
第二节 统计和统计学的发展	(2)
第三节 统计数据	(5)
第四节 统计学的基本概念	(10)
第五节 SPSS 简介	(13)
习题一	(18)
第二章 统计数据的收集	(19)
第一节 统计调查方案	(19)
第二节 数据的收集方式和方法	(23)
第三节 调查问卷设计	(28)
习题二	(37)
第三章 统计数据的整理与描述	(38)
第一节 统计整理	(38)
第二节 数据特征的描述	(54)
第三节 SPSS 的应用	(78)
习题三	(86)
第四章 抽样及抽样分布	(88)
第一节 关于抽样	(88)
第二节 抽样中的常用分布及定理	(92)
第三节 抽样误差	(98)
第四节 抽样方法和抽样的组织方式	(102)
第五节 SPSS 的应用	(108)
习题四	(117)
第五章 参数估计与假设检验	(120)
第一节 参数的点估计	(120)
第二节 参数的区间估计	(127)
第三节 假设检验的一般方法	(135)
第四节 常用参数的假设检验	(138)
第五节 非参数检验	(143)
第六节 SPSS 的应用	(149)
习题五	(165)

第六章 相关分析与回归分析	(167)
第一节 相关分析	(167)
第二节 回归分析的基本概念	(171)
第三节 一元线性回归模型	(173)
第四节 多元线性回归模型	(180)
第五节 非线性回归模型	(185)
第六节 SPSS 的应用	(189)
习题六	(196)
第七章 方差分析与正交试验设计	(198)
第一节 方差分析	(198)
第二节 正交试验设计	(209)
第三节 SPSS 的应用	(221)
习题七	(225)
第八章 时间序列分析	(228)
第一节 时间序列概述	(228)
第二节 动态比较分析	(231)
第三节 动态平均分析	(234)
第四节 时间序列影响因素的测定	(241)
第五节 时间序列的现代分析	(262)
第六节 SPSS 的应用	(267)
习题八	(276)
第九章 指数分析	(278)
第一节 指数概述	(278)
第二节 总指数	(280)
第三节 指数体系与因素分析	(286)
第四节 功效系数法	(293)
第五节 指数的应用	(295)
习题九	(301)
附录 1 统计方法在 Excel 上的实现	(304)
第一节 Microsoft Office Excel 概述	(304)
第二节 Excel 基本操作	(307)
第三节 Excel 在统计中的应用 (一)	(311)
第四节 Excel 在统计中的应用 (二)	(316)
附录 2 常用工具数表	(330)
习题参考答案	(342)
参考文献	(346)

第一章 绪 论

在人们的日常生活中，统计问题以及与统计有关的问题随处可见。例如：我们在选购某种商品时通常会货比三家，其目的就是要获取所购买的同类商品的相关信息，如质量、花色、服务、产地、信誉、占有率等，便于决策；证券市场上投资者关注、收集、分析公司发布的统计信息以确定如何投资；而国家统计局关心的是国民经济的增长速度、人们的消费水平、企业生产投资状况等。这些数据的取得都离不开统计。统计已经渗透到人们日常生活的各个方面，成为人们认识世界的有力工具。

人类的行为、自然界的现象与科学研究通常会产生大量的数据信息，要想从这些数据资料中提取有价值的信息，只有依靠统计学的方法。统计学作为方法论在经济、教育、医疗、金融、企业管理等各领域得到广泛的应用。著名经济学家马寅初曾说过：“……学者不能离开统计而研学，政治家不能离开统计而施政，事业家不能离开统计而执业。”

第一节 统计与统计工作

一、“统计”的含义

统计作为一种社会实践活动有着悠久的历史，它是适应国家宏观管理和社会经济发展的需要而产生和发展起来的。

统计包括三个方面的含义：统计工作（或统计活动）、统计资料和统计学。这已成为国际上通行的有关统计的习惯术语。

1. 统计工作

统计工作是指有关机构为满足各方面管理或从事科学研究的需要，而对总体或现象的数量方面进行搜集、整理和分析有关数据的工作过程。

2. 统计数据

统计数据，也称统计资料，是指统计工作过程中所产生的原始的或加工过的统计数据、统计报表、统计图表、统计分析报告，以及与之相联系的其他资料的总称。对统计资料一般有准确、客观、及时、连续、系统等要求。

3. 统计学

统计学是指系统地阐述统计实践活动基本原理和研究方法的理论。它是一门研究如何搜集、整理和分析统计数据的理论和方法论科学。统计方法是自然与社会各研究领域和工作部

门必要的基本的数量分析手段,是从大量数据资料中提取有价值信息的重要工具。

统计的三种含义密切相关。统计工作和统计资料是过程和结果的关系。统计资料是统计工作达到的成果或结晶,它来源于统计工作又服务于统计工作。统计工作和统计学是理论与实践的关系。统计学是统计活动发展到一定阶段的产物,是统计实践经验的理论概括,并用于指导统计实践,为统计工作提供理论依据。统计工作的发展会不断完善和丰富统计理论,而统计理论的丰富又促进了统计工作的发展,使之更有效、更科学。

以上对“统计”一词含义的理解,有助于对统计学的初步认识。但这还不够。要真正了解统计学这一门科学的性质、特点、功能,还要从其产生的历史来考察。

二、统计工作的基本过程

统计工作的一般过程包括统计设计、统计调查、统计整理、统计分析等。

统计设计,是统计工作的初始阶段,严格地说是统计工作实际进行之前的准备阶段,是根据统计研究的目的,对统计工作各个环节的统筹考虑和安排。统计设计的结果表现为各种统计方案,比如确定调查对象与调查单位、统计指标体系、分类分组、统计报表、调查方案、汇总和整理、分析方案等。

统计调查,就是按照统计设计和调查方案,有计划、有组织地向调查单位搜集统计资料的工作过程。具体说,统计调查就是按照统计设计规定的统计调查对象、分类标准和调查方法,对总体(社会经济现象或者自然现象)的各个单位有关标志值进行有计划、有组织的搜集,取得原始资料的过程。统计调查是整个统计工作的基础,是统计整理与统计分析的前提,是保证统计资料准确、及时、全面而系统的重要环节,是关系到整个统计活动成败的关键。

统计整理,是根据统计的目的和要求,对统计调查阶段搜集到的大量数据资料进行加工使之条理化、系统化的过程。统计整理为统计分析打下基础,起到承前启后的作用。

统计分析,就是指根据研究的目的,运用统计方法,以统计资料为依据进行科学的分析和研究,揭示总体或现象的特征、本质和规律性,提出解决问题方法的过程。

第二节 统计和统计学的发展

一切科学都来源于实践,统计这门科学的产生和发展也是如此,它是适应社会经济的发展的需要而产生和发展起来的。统计的产生和发展与人类社会的文明、社会进步是密切相关的。正如德国历史学家 A. L. V. Schlozer (1735~1809) 所言:“统计是动态的历史,历史是静态的统计。”

统计的产生有着很悠久的历史。最早的统计活动可以追溯到原始社会末期,当时的人们打猎捕鱼、结绳记事等活动,就孕育了统计的最初含义。由于生产力发展缓慢,当时统计还只是简单意义上的登记数字和比较,还没有被当作一门学问去研究,尚处于萌芽阶段。

据历史记载,中国古代的一些政治家、军事家早就意识到统计的重要性,秦国的商鞅在《商君书》的《去强篇》中这样说:“强国当知十三数:境内仓、口(府)之数,壮男、壮女之数,老、弱之数,官士之数,以言说取食者之数,利民之数,马、牛、刍、藁之数。欲强国,不知国十三数,地虽利,民虽众,国愈弱至削。”秦汉之时,已有地方田亩和户口记录;

唐宋时则有计口授田田亩鱼鳞册等土地调查和计算；到明清时代，已经建立了经常的人口登记和保甲制度等。同样在欧洲，古希腊和古罗马时代，就已开始了人口数字和居民财产的统计调查工作。封建君主国家根据其需要也进行了有关人口、军队、世袭领地和财产等统计。但是那个时代的统计活动和统计资料，无论中国还是外国，都是一些原始的、简单的汇总和计算。

随着社会的不断进步，统计在社会经济管理中的作用更为重要，从而得到了更广泛的应用和发展。随着统计实践的发展和经验的积累，各种统计理论和学说应运而生。但是直到 17 世纪以后统计才逐渐发展成一门科学，并逐步形成了不同的学派。

(1) 国势学派。这个学派属于古典统计学时期，产生于 18 世纪的德国，代表人物是 Hensmanu Corning (1606~1681)，继承者主要有 G. AchenWall (1719~1772)。《近代欧洲各国国势学概论》是其代表作。他们在哥廷根大学中开设了一门新的课程，以叙述国家显著事项和国家政策关系为内容，最初叫做国势学。因为在德文中“国势”与统计的含义相通，后正式定为统计学。它收集大量的实际资料分门别类记述有关国情国力，包括土地、人口、政治、军事、财政。这个学派主要采用文字记述为主，始终没有把数量对比分析作为这门科学的基本特征，故又称为记述学派。这一学派最为人称道的是 G. AchenWall 在 1749 年第一个把国势学定名为统计学，“统计学”这一名词以后就一直沿用至今。

(2) 政治算术学派。与国势学派同属于古典时期，产生于 17 世纪资本主义的英国，创始人是威廉·配第 (William Petty, 1623~1687)。威廉·配第在《政治算术》一书中第一次用计量和比较的方法，即大量观察法、分类法以及对比综合法等，从整体上分析了英、法、荷三国的经济军事实力及其内在潜力，论证了英国称雄世界的条件和地位。书中提出了一整套比较系统的反映社会经济现象数量方面、分析其数量关系的方法，创造性地进行了建立一门统计科学的尝试。马克思称配第在某种程度上也可以说是统计学的创始人。这里的“政治”是指政治经济学，“算术”是指统计方法。这本书中采用了不同于过去传统的方法，而是利用数字、重量、尺寸来说话，找出事物变化的规律。

政治算术学派的另一个代表人物是约翰·格朗特 (John. Graunt, 1620~1674)。17 世纪上半叶，英国多次发生严重的瘟疫，政府定期公布有关人口出生和死亡的数字。约翰·格朗特利用这些数字资料研究并发表了《关于死亡表的自然和政治的观察》的论著，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和不同死因的比例等人口规律；并且第一次编制了初具规模的生命表，对年龄性别死亡率与人口寿命作了分析，从而引起了普遍的关注。

(3) 数理统计学派。产生于 19 世纪中叶，其奠基人是比利时的物理学家和统计学者凯特勒 (L. A. J. Quetelet, 1796~1874)。凯特勒把德国的国势学派、英国的政治算术学派和意大利、法国的古典概率论加以融合改造为近代意义的统计学。他是法国数学家、物理学家拉普拉斯的学生，其代表作是《论人类》、《概率论书简》和《社会物理学》。凯特勒在研究社会现象时主张采用自然科学的方法，把古典概率论引进统计学，发展了概率论，推广了概率论在统计中的应用。他最先用大数定律论证了社会经济现象纷繁复杂的偶然性中隐藏着规律性，并且提出了误差理论，用以解决统计上的准确性问题。他明确地指出统计学是一门既研究社会现象又研究自然现象的独立的方法论科学。随着统计研究的领域不断扩大，概率论方法的运用日趋成熟。到 19 世纪末数理统计学慢慢地独立，自成一派。

(4) 社会统计学派。19 世纪后半叶，正当数理统计学派开始发展的时候，在欧洲以德国

为中心又兴起了社会统计学派。由大学教授克尼斯 (K. G. A. Knies, 1821~1898) 首创, 主要代表人物有恩格尔 (C. L. E. Engel, 1821~1896)、梅尔 (G. V. Mayr, 1841~1925) 等人。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科学, 以此同数理统计学派的通用方法论相抗衡。这是社会统计学派的“实质性科学”的显著特点。他们强调统计学是研究社会现象的科学, 包括统计资料的搜集、整理和分析研究, 目的是要揭示现象内部的联系。强调在统计研究中必须以事物的质为前提和认识质的重要性。从学术渊源上看, 他们融合了记述学派和政治算术学派观点, 又继承和发扬了凯特勒强调研究社会现象的传统, 把政府统计和社会调查结合起来形成自己的特点。德国的社会统计学派在国际学术界占有一席之地, 日本等国深受其影响。

现代统计学时期是指 20 世纪初至今的时期。这一时期统计科学出现了新的趋势。

由于自然科学和工程技术科学的发展日新月异, 数理统计学与其紧密结合, 获得了广泛的应用, 使统计学进入了鼎盛时期。首先它在随机抽样的基础上建立了推断统计的理论和办法。统计学的主流从描述统计学转向推断统计学。20 世纪 30 年代 R·费希尔的推断统计理论标志着现代数理统计学的确立。所谓推断统计也就是通过随机样本来推断总体数量特征的办法。这种方法源于英国数学家格赛特 (N. S. Gosset, 1876~1936) 的小样本 t -分布理论, 其后美国学者费希尔 (R. A. Fisher)、波兰统计学家内曼 (J. Neyman, 1894~?) 以及 E. S 比尔森等人进一步发展, 建立了统计假设理论。后来美国统计学家瓦尔德 (A. Wald, 1902~1952) 又将统计学中的估计和假设理论加以归纳, 创立了决策理论; 美国的威尔克斯 (S. S. Wilks, 1906~1964), 英国的威史特 (J. Wishart, 1898~1956) 等学者对样本分布理论又加以充实和发展; 美国的科克伦 (W. G. Cochran, 1909~1980) 等在 1957 年提出了实验设计的理论和方法, 大大拓宽了统计学的范围。

20 世纪 60 年代以前深受流派的影响, 门户对立严重, 一定程度上影响了统计学的发展; 而同一时期的社会统计学发展则比较缓慢, 其基本趋势是由实质性科学向方法论转变。重要的人物如第二次世界大战后德国法兰克福大学教授弗拉斯克姆波 (P. Flaskamper, 1886~?)。60 年代以后数理统计学发展有三个明显的趋势: ①统计学对数学的依赖和吸收更多, 大量的采用数学方法; ②数理统计学的新分支和以数理统计为基础的边缘学科不断形成; ③与电子计算机技术相结合, 应用范围更广, 作用更大, 已成为现代统计学的主流。

在其他国家 (如前苏联) 的一些统计学者也在理论上作出了一定的贡献, 但囿于意识形态和其他方面因素的束缚没能及时吸取各国统计学发展的最新成果, 一味地加以排斥, 反而落后于西方学者。中国的统计学发展也深受影响进步迟缓, 经济统计和数理统计之争一直没有间断。进入到 20 世纪 80 年代以后, 局面有所改观, 大统计格局 (不分经济统计和数理统计) 深入人心。

从统计学的过程来看, 统计学是从研究社会经济现象开始的, 经过 300 多年的演变和发展, 统计学趋于成熟, 成为一门研究客观事物数量方面的独立的方法论科学, 从研究对象、研究方法和学科体系上看, 它是研究客观事物总体数量方面的方法论科学, 强调对客观事物总体进行大量的观察, 通过归纳推理等方法来获得总体数量方面的综合认识, 是一门多科性的科学。

统计学经过 300 多年的发展, 与各学科之间既有融合又有分化, 特别是在现代自然科学和社会科学相互影响和渗透, 边缘学科不断涌现的大趋势下, 统计学作为一门方法论科学应

用于各学科,同它们结合,产生了一系列专门的或专业的统计学分支,从而使统计学形成了一个多学科的家庭。对其主体部分,以研究领域为标志进行划分,统计学体系可分为两个部分:数理统计学和社会统计学。数理统计学是以概率论为基础,以抽样为核心,研究随机变量的数量规律,是在社会经济和自然科学及应用技术统计实践基础上抽象和概括出来的具有普遍性的纯方法论科学。数理统计学又分为理论数理统计学和应用数理统计学,其分支如抽样理论、非参数统计、多变量分析和时间序列分析等。社会统计学是以人口、社会、经济、科技、文化等人类自身及其活动为研究对象的统计方法论。包括众多学科分支,如人口统计学、管理统计学、商业统计学、金融统计学、教育统计学等。还有一些交叉学科如生物统计学、心理统计学、气象统计学、医疗卫生统计学、统计物理学,这些学科都是统计学与自然科学和应用技术科学相结合产生的新学科。此外还有统计史、比较统计学等。

第三节 统计数据

统计数据是统计工作活动过程中所取得的成果,是反映总体或现象的数字资料以及与之相联系的其他资料的总称。统计数据是对总体或现象进行计量的结果,统计研究客观事物的数量方面离不开统计数据。

一、统计测量

统计数据来源于测量过程,对统计数据的属性、特征进行分类、标示和计算,称为统计测量或统计量度。例如,对工业企业经济效益的统计,对居民生活水平的统计,也可以说是对工业企业经济效益的计量,对居民生活水平的计量,如此等等。由于客观事物有的比较简单,有的比较复杂,有的特征和属性是可见的(如人的外貌特征),有的则是不可见的(如人的偏好和信仰),有的表现为数量差异,有的表现为品质差异。因此,统计计量也就有定性计量和定量计量的区别,并且可分不同的层次。美国社会学家、统计学家史蒂文斯(S. S. Stevens) 1968 年按照变量的性质和数学运算的功能特点,将统计计量划分为四个层次或四种计量尺度。

1. 定类尺度 (nominal level of measurement)

将数字作为现象总体中不同类别或不同组别的代码,这是最低层次的尺度。在这种情况下,不同的数字仅表示不同类(组)别的品质差别,而不表示它们之间量的顺序或量的大小。这种尺度的主要数学特征是“=”或“≠”,其计量的结果是定类数据。

例如将国民经济按其经济类型,可以分为国有经济、集体经济、私营经济、个体经济等类,并用(01)代码表示国有经济,(02)表示集体经济,(03)表示私营经济,(04)表示个体经济。并且用(011)代表国有经济中的国有企业,(012)代表国有联营企业;用(021)表示集体经济中集体企业,(022)表示集体联营企业;用(031)表示私营经济中的私营独资企业,(032)表示私人合伙企业,(033)表示私营有限责任公司;用(041)表示个体经济中的个体工商户,(042)表示个人合伙,等等。其中两位代码表示经济大类,而三位代码则表示各类中的构成。不同代码反映同一水平的各类(组)别,并不反映其大小顺序。各类中虽然可以计算它的单位数,但不能反映第一类的一个单位可以相当于第二类的几个单位等。

2. 定序尺度 (ordinal level measurement)

定序尺度不但可以用数表示量的不同类(组)别,而且也反映量的大小顺序关系,从而可以列出各单位、各类(组)的次序。这种尺度的主要数学特征是“ $>$ ”或“ $<$ ”。

例如对合格产品按其性能和好坏,分成优等品、一等品、合格品等。这种尺度虽然也不能表明一个单位一等品等于几个单位二等品,但却明确表示一等品性能高于二等品,而二等品性能又高于三等品等。定序尺度除了用于分类(组)外,在变量数列分析中还可以确定中位数、四分位数、众数等指标的位置。其计量的结果是定序数据。

3. 定距尺度 (interval level of measurement)

定距尺度也称间隔尺度,是对事物类别或次序之间间距的计量,它通常使用自然或度量衡单位作为计量尺度。定距尺度是比定序尺度高一层次的计量尺度。它不仅能将事物区分为不同类型并进行排序,而且可以准确地指出类别之间的差距是多少。其计量的结果是定距数据。

例如,学生某门课程的考分,可以从高到低分类排序,形成90分、80分、70分,直到零分的序列。它们不仅有明确的高低之分,而且可以计算差距,90分比80分高10分,比70分高20分等等。定距尺度的计量结果表现为数值,可以进行加或减的运算,但却不能进行乘或除的运算,其原因是在等级序列中没有固定的、有确定意义的“零”位。例如,学生甲得分90分,学生乙得0分,可以说甲比乙多得90分,却不能说甲的成绩是乙的90倍或无穷大。因为“0”分在这里不是一个绝对的标准,并不意味着乙学生毫无知识。恰如我们不能说 40°C 比 20°C 暖和2倍一样。没有确定的标准的“零”位,但有基本的确定的测量单位,如学生成绩的测量单位是1分,质量价差的测量单位是1元,温度的测量单位是 1°C 等,这是定距尺度的显著特点。

4. 定比尺度 (ratio level of measurement)

定比尺度是在定距尺度的基础上,确定可以作为比较的基数,将两种相关的数加以对比,而形成新的相对数,用以反映现象的构成、比重、速度、密度等数量关系。由于它是在比较基数上形成的尺度,所以能够显示更加深刻的意义。定比尺度的主要数学特征是“ $\div$ ”或“ $\times$ ”。其计量的结果是定比数据。

例如将某地区人口数和土地面积对比计算人口密度指标,说明人口相对的密集程度。甲地区人口可能比乙地区多,但甲地区的土地更广阔,用人口密度指标就可以说明相对说来甲地区人口不是多了,而是少了。又如将一个国家(地区)的国内生产总值与该国(地区)居民对比。计算人均国内生产总值,可以反映国家(地区)的综合经济能力。1998年我国国内生产总值约占世界生产总值的12%,排列世界第七位,堪称世界经济大国,但我国人口占世界总人口的21.2%,如果按人均国内生产总值计算,在世界各国中又居于比较落后的位次,说明我国仍属于发展中国家。

上述四种计量尺度对事物的计量层次是由低级到高级、由粗略到精确逐步递进的。高层次的计量尺度具有低层次计量尺度的全部特性,但不能反过来。显然,我们可以很容易地将高层次计量尺度的测量结果转化为低层次计量尺度的测量结果,比如将考试成绩的百分制转化为五等级分制。在统计分析中,一般要求测量的层次越高越好,因为高层次的计量尺度包含更多的数学特性,所运用的统计分析方法越多,分析时也就越方便,因此应尽可能使用高层次的计量尺度。

二、统计数据类型

1. 按被描述对象和时间的关系划分

统计数据按照被描述的对象和时间的关系不同分为截面数据 (cross-sectional data)、时间序列数据 (time series data) 和面板数据 (panel data)。

(1) 截面数据是不同统计单位在同一时间点上收集的数据。一般是按照统计单位排列的。

(2) 时间序列数据是在不同时间点上收集到的数据。这类数据反映了某一事物、现象等随时间的变化状态或程度。

(3) 面板数据是截面数据与时间序列数据综合起来的一种数据类型。其有时间序列和截面两个维度, 当这类数据按两个维度排列时, 是排在一个平面上, 与只有一个维度的数据排在一条线上有着明显的不同, 整个表格像是一个面板, 故称面板数据。

2. 按统计计量层次划分

统计数据是采用某种计量尺度对事物进行计量的结果, 采用不同的计量尺度会得到不同类型的统计数据。从不同计量尺度计量的结果来看, 可以将统计数据分为以下四种类型:

(1) 定类数据。表现为类别, 但不区分顺序, 是由定类尺度计量形成的。

(2) 定序数据。表现为类别, 但有顺序, 是由定序尺度计量形成的。

(3) 定距数据。表现为数值, 可进行加、减运算, 是由定距尺度计量形成的。

(4) 定比数据。表现为数值, 可进行加、减、乘、除运算, 是由定比尺度计量形成的。

前两类数据说明事物的品质特征, 其结果均表现为类别, 也称为定性数据 (qualitative data) 或品质数据; 后两类数据说明的是现象的数量特征, 能够用数值来表现, 由于定距尺度和定比尺度属于同一测度层次, 可以把后两种数据看作是同一类数据, 统称为定量数据或数值型数据 (quantitative data)。在统计软件 IBM SPSS 中就将数据测量划分为 Nominal、Ordinal、Scale 三个类别, 其中定距和定比数据就合并为 Scale。

区分测量的层次和数据的类型是十分重要的, 因为对不同类型的数据将采用不同的统计方法来处理和分析。比如, 对定类数据, 通常计算出各组的频数或频率, 计算其众数和异众比率, 进行列联表分析和卡方检验等; 对定序数据, 可以计算其中位数和四分位差, 计算等级相关系数等非参数分析; 对定距或定比数据还可以用更多的统计方法进行处理, 如计算各种统计量、进行参数估计和检验等。我们所处理的大多为定量数据。

适用于低层次测量数据的统计方法, 也适用于较高层次的测量数据, 因为后者具有前者的数学特性。比如: 在描述数据的集中趋势时, 对定类数据通常是计算众数, 对定序数据通常是计算中位数, 但对定距和定比数据同样也可以计算众数和中位数。反之, 适用于高层次测量数据的统计方法, 则不能用于较低层次的测量数据, 因为低层次数据不具有高层次测量数据的数学特性。比如, 对于定距和定比数据可以计算平均数, 但对于定类数据和定序数据一般不计算平均数。这对于选择统计分析方法是十分有用的。

三、数据的来源

从事研究和管理, 所使用的数据来源一般有直接和间接两种。直接来源也就是第一手资料, 来源于实地调查。因此通过各种有效的方式如访谈、观察、实验等获取数据信息就成为

关键；间接来源是获得二手资料，往往是通过公开出版物、第三方机构或个人已经取得的数据的使用取得，这些数据的发布形式有传统纸张版、数据库或者网络版。

间接数据来源常用的渠道有综合性强的如统计年鉴等；企事业单位发布的各种统计报表、报告数据等；行业出版物和学术期刊。随着网络科技的普及和发展，各种政府网站（www.stats.gov.cn，www.cei.gov.cn，www.mofcom.gov.cn）、专业网站如知网（www.cnki.com）、维普（www.cqvip.com）、万方（www.wanfangdata.com.cn，www.mycos.com.cn）应运而生，各种搜索引擎如百度、搜狗、Google 等也是一个重要数据来源。

四、统计数据的质量

统计数据质量是一个具有丰富内涵的综合性概念，而不仅仅是把准确性作为唯一的衡量标准。具体来说，它包括统计数据的内容质量、表述质量及约束标准这三大方面。参见表 1-1。

表 1-1 不同国家、地区和国际组织的统计数据质量维度

国家或国际组织	统计数据质量的维度
加拿大统计局	适用性、准确性、及时性、可取得性、衔接性、可解释性
英国政府统计	准确性、及时性、有效性、客观性
荷兰统计局	适用性、准确性、及时性、有效性、减轻调查负担
韩 国	适用性、准确性、及时性、可取得性、可比性、有效性
美国分析局	可比性、准确性、适用性
澳大利亚统计局	准确、及时、适用、可取得性、方法科学性或健全性
欧洲统计局	适用性、准确性、及时性、可取得性、衔接性、可比性、方法专业性
香港人口统计署	相关性、准确性、及时性、可比性、一致性、可获得性
瑞典统计局	准确性、及时性、可比性和一致性、可获得性和一致性
新西兰统计局	相关性、准确性、及时性、一致性、可解读性
国际货币基金组织—数据公布通用系统（GDDS）	数据特征、数据质量、诚信度、公众可获得性
国际货币基金组织—数据质量评估框架（DQAF）	前提条件、诚信、方法的健全性、准确性和可靠性、适用性、可获得性

来源：向蓉美，国家统计数据质量研究述评，载于政府统计数据质量研讨会论文集，2010 年。

1. 统计数据的内容质量

统计数据的内容质量是统计数据最基本的特征，它包括相关性、准确性与及时性。一旦缺少了其中任何一个，统计数据就失去了转化为信息的性质和基本作用。因此，这三个特征也可称为统计数据质量的主要特征。

（1）准确性。准确性指观测值或估计值与未知的真值之间的距离（接近程度），通常用统计误差来衡量。它是统计数据质量的基础和核心内容，也是传统的“统计数据质量”概念所考虑的主要问题。一般来讲，误差分为系统误差和随机误差两部分，有时也用引起不准确性的主要潜在原因（如抽样误差、无回答误差等）来分类描述。完全准确的测量经常受到成本的限制，有时甚至是不可能的。所以关键的是误差是否已降低到用户可以接受的地步。

(2) 相关性。相关性是指统计机构所生产的数据是否正是用户感兴趣的统计数据。统计数据的相关性反映了它满足用户需求的程度,它与所提供的可利用数据是否关注了对用户来说最重要的主题有关。由于对相关性的评价是主观的,会随用户需求目标的改变而改变,所以统计机构所要做的是平衡不同用户的互相矛盾的需求目标,在给定的资源条件限制下,尽可能的满足大部分用户的大部分需求。

(3) 及时性。与用户需求相关的准确的统计数据如果没有在用户做出决策之前传递给用户,那么该数据对用户来说,是没有用的。所以,及时性也是统计数据能否满足用户需求的重要特征。如果该现象本身变化比较迅速,则对该类统计数据的及时性要求高;如果该现象本身变化比较缓慢,则对及时性要求不高。

2. 统计数据的表述质量

对统计数据质量来说,仅考虑其内容方面的质量是不够的。统计数据需要表述,特别是将某个统计数据同其他相关统计数据相互联系地加以表述时,要考虑表述的质量问题。比如单个数据的内容是正确的,但表述不清晰、不充分,就会影响整套数据的质量,甚至引起误解。所以,统计数据质量必须考虑其表述质量。统计数据的表述质量包括可比性、可衔接性和可理解性。

(1) 可比性。可比性是指同一项目的统计数据在时间上和空间上的可比程度。这要求统计的概念和方法在时间上保持相对稳定,在不同地区使用统一的统计制度方法和分类标准,保持统计数据的口径范围、计算方法在时间上一致衔接,在地区之间可比。

(2) 可衔接性。可衔接性是指同一机构内部不同统计调查项目之间、不同机构之间以及与国际组织之间统计数据的衔接程度。这要求全国范围内所有专业统计项目在统一的统计框架体系、分类标准下,按统一的方法编制统计数据,在统计调查和数据加工整理中使用统一的方法和程序,同时采用国际统计标准,如联合国 1993 年 SNA 的框架体系等。

(3) 可理解性。可理解性是指统计数据便于用户正确理解并使用的程度。统计数据是提供给用户使用的,如果某些用户不能理解,看不懂统计数据和统计分析报告,当然也就谈不上使用数据。为了恰如其分地使用从统计机构那里得到的统计数据,用户必须了解他们所获得数据的性质。这就要求统计机构在提供统计数据的同时附带提供对数据的补充说明,如提供隐含在有关概念下面的说明、已使用的分类法、数据收集和加工过程中所使用的方法以及统计机构自身对数据质量的评价。

3. 统计数据的约束标准

在实现统计数据目标的过程中,除了注意统计数据的内容质量和表述质量这两方面外,还必须注意以下两项具有普遍意义的约束标准,这也体现了统计数据的质量特征。

(1) 可取得性。可取得性是指用户从统计部门取得统计数据的便利程度。对于有用的统计数据,用户必然要考虑:能得到哪些数据,如何得到这些数据。因此,统计数据必须以一种用户能够使用(搜寻方便)而且能够负担的形式提供给用户。这要求提供统计数据时,必须列明用户从统计机构可以取得的统计数据内容,同时要应用先进便捷的统计数据服务方式,使用户取得数据更为便利。

(2) 有效性。有效性是指统计数据的利用所产生的效益要大于提供该数据的成本。如果情况相反,则提供这种数据对提供方和使用方来说都是不值得的。虽然目前统计数据的效益和成本特别是前者不容易准确地计量,但是,保持这样一种基本的指导思想是十分必要的。

这要求在统计数据的其他质量不受大的影响的前提下,尽可能降低统计数据的生产费用,提高效率。

第四节 统计学的基本概念

一、标志和指标

(一) 标志

标志也称为标识,是说明总体单位所具有特征的名称。按表示特征不同分为品质标志和数量标志。

1. 品质标志

品质标志是表明总体单位品质和属性特征的标志。它一般用文字来表示。如工人的性别、工种。又如企业的经济类型、规模大小等这类标志只是从性质的差别上来表现总体特征和属性的,如果说某公司下属各企业为总体单位,则各企业生产经营方向、所有制性质是企业质的属性,这些品质标志用文字说明,一般不用数值表示。实践中也有以量来反映总体单位质的属性的,如反映产品质量的“一等品”,劳动效率为“人均年销售额 200 万元”等。

2. 数量标志

数量标志是表明总体单位数量特征的标志,即变量。它是用数值表示的。如工人的工龄、年龄、工资等。如作为总体单位的各工业企业的资本总额、资本综合成本、工业总产值、净产值、商品产值、成本总额、期间费用、实现利税总额等,这些都是以具体的绝对数值来表示的。又如企业的工人数、总产值、设备原值等,这类标志只是从数量上的多少来表现总体的特征所以称为数量标志,数量标志的具体表现值就称为标志值。

(二) 指标

1. 统计指标的含义

统计指标是在具体条件下,反映客观现象总体的数量特征的概念和具体数值。它是根据总体单位某一标志值汇总或综合而成,是用来说明总体某一数量特征的,也有质的特征的量。如企业作为研究总体时,它的工业生产总值 500 万元,年产量 1 000 台,这些指标是由该企业每个职工的产值、产量等汇总取得的。例如我国第四次人口普查结果,1990 年 7 月 1 日 0 时中华人民共和国人口为 1 160 017 381 人,其中大陆 30 个省、直辖市、自治区人口为 1 133 682 501 人,共有 276 947 962 户,平均每户 3.96 人。又如我国 1993 年国民经济社会发展统计公报公布:全国 GDP 为 31 380 亿元,其中农业、工业、建筑业和第三产业分别为 6 650 亿元、14 140 亿元、2 105 亿元和 8 485 亿元,分别比上年增长 4%、21.1%、15%和 9.3%;社会商品零售总额 12 237 亿元,比上年增长 26.1%;城镇居民人均生活费 2 337 元,农村居民人均收入 921 元,扣除物价上涨因素后,分别比上年增长 10.2%,3.2%。这类统计数值就是统计指标。从以上可以看出,统计指标构成的要素包括:①指标名称(GDP);②计量单位(亿元);③指标数值(31 380);④时间(1993 年);⑤地点(全国)等五个要素。此外还隐含指标含义与计算方法两个要素。