



高职高专“十二五”规划教材

统计基础与实用方法

Tongji Jichu Yu
Shiyong Fangfa

马冀 赵养森 罗宏 / 编著



立信会计出版社
LIXIN ACCOUNTING PUBLISHING HOUSE

高职高专十二五规划教材

统计基础与实用方法

马 冀 赵养森 罗 宏 编著



立信会计出版社

LIXIN ACCOUNTING PUBLISHING HOUSE

图书在版编目(CIP)数据

统计基础与实用方法 / 马冀, 赵养森, 罗宏编著.

—上海: 立信会计出版社, 2012. 8

高职高专“十二五”规划教材

ISBN 978-7-5429-3614-1

I. ①统… II. ①马…②赵…③罗… III. ①统计学

—高等职业教育—教材 IV. ①C8

中国版本图书馆 CIP 数据核字(2012)第 182933 号

策划编辑 陈岗伟

责任编辑 余 榕

封面设计 周崇文

统计基础与实用方法

出版发行 立信会计出版社

地 址 上海市中山西路 2230 号 邮政编码 200235

电 话 (021)64411389 传 真 (021)64411325

网 址 www.lixinaph.com 电子邮箱 lxaph@sh163.net

网上书店 www.shlx.net 电 话 (021)64411071

经 销 各地新华书店

印 刷 常熟市梅李印刷有限公司

开 本 787 毫米×1092 毫米 1/16

印 张 19.75

字 数 454 千字

版 次 2012 年 8 月第 1 版

印 次 2012 年 8 月第 1 次

印 数 1—3 100

书 号 ISBN 978-7-5429-3614-1/C

定 价 32.00 元

如有印订差错,请与本社联系调换

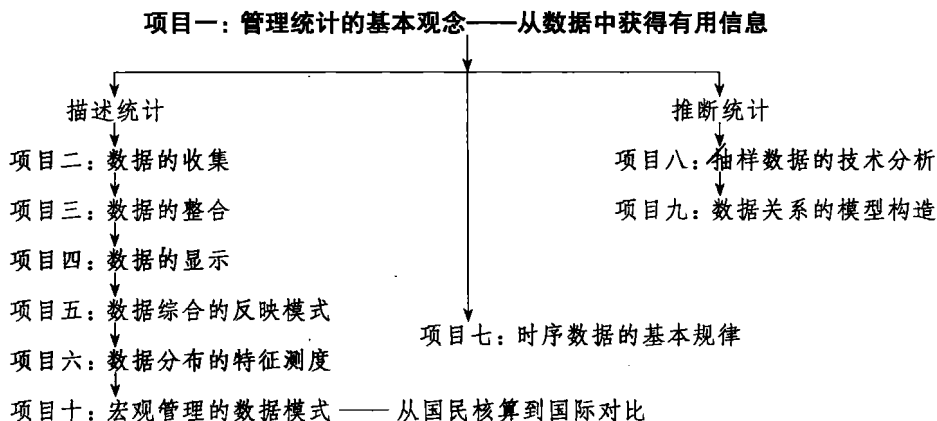
前 言

要做好经济工作和管理工作,就必须具备良好的统计素养。统计学与经济学的合流部分形成了“计量经济学”,使经济学的面目为之一新。而统计学与管理学的合流部分产生了“科学管理学派”,使管理学走上了精确化和科学化的道路。其实,统计学的应用领域不只限于经济学和管理学。无论是在自然科学领域、社会科学领域,还是在国家的宏观管理和企业的生产经营管理领域,甚至是人们的日常生活领域,随着信息需求量日益增多,信息处理的技术就更加复杂,作为信息技术支柱的统计方法也就越来越受到人们的关注。

本书是专门为高职院校管理类各专业编写的一本以非主流方式表达管理统计思想和管理统计方法的教科书。本书从高职学生的培养目标出发,注重能力教育的开发和实践,同时又结合课程本身的特点,将管理统计的职业能力概括为科学认识问题的能力、快速处理数据的能力、准确计算数值的能力、提供管理服务的能力、依据数据建模的能力和培养创新思维的能力。不仅如此,本书在表述方式上也做了重大的改变,在描述基本概念、基本原理和基本方法时,都以故事开篇、核心理念、概念在先、原理解读、案例收官为线索展开叙述的。众所周知,数字是统计的语言,本书的这种表述就是让读者从故事中感悟统计数字的精髓,从而建立起科学的统计理念,再透过概念和原理来实现专业的升华,并在应用中体现统计方法和统计数字的灵魂。管理统计作为处理和分析数据的方法和技术,具体包括数据的收集、整合、显示、分析和解释等内容。本书就是按照这个过程来组织设计内容的。本书共包括10个项目和3个附录。项目一:管理统计的基本观念——从数据中获得有用信息,目的是让读者对统计和统计数据有一个基本的了解。项目二:数据的收集,明确数据从哪里来,通过什么方法和途径获取所需的统计数据。项目三:数据的整合,正确领悟数据处理的专业技术特征,而不是加加减减的数字游戏,掌握数据整合的基本方法。项目四:数据的显示直观明了展示统计数据的工具——图表,把握图表的显示技术。项目五:数据综合的反映模式,探讨数据综合的基础模式、对比关系的一般模式和扩展模式——指数。项目六:数据分布的特征测度,探讨如何使用特征值对数据进行概括。项目七:时序数据的基本规律,探讨时间序列的描述性方法和时间序列的趋势结构模型。项目八:抽样数据的技术分析,探讨概率抽样方法,重点介绍参数的估计方法和假设检验的基本原理。项目九:数据关系的模型构造,探讨

变量之间关系强度的方法、一元线性回归模型、列联关系模型、误差分解模型(方差分析)和可靠性检验方法。项目十:宏观管理的数据模式——从国民核算到国际对比,探讨国民经济核算中的基本总量指标以及国际经济对比过程中的数据模式。附录一:统计分析软件简介,介绍常用的综合统计分析软件和常用的专用统计分析软件,着重介绍 Excel 数据分析方法和社会科学统计软件包 SPSS 两种软件的统计分析功能。附录二:描述统计基本概念、基本原理和基本技能问题训练。附录三:用 Excel 生成概率分布表。

本书从描述统计和推断统计两个层面介绍统计的基础理论和实用方法。描述统计方法只能对统计数据做基本的利用,若要从统计数据中挖掘规律性的东西,那就必须广泛采用推断统计方法,推断统计数据中隐含着丰富的信息。所谓推断统计,就是在收集、整理观察样本数据的基础上,对有关总体作出推断,其特点是根据随机性的观察样本数据、问题的条件和假定,以概率形式对未知事物作出推断。本书的内容就是按二元结构安排的:



本书可供高职院校管理类各专业有针对性地选择讲授(或学习)描述统计和推断统计的部分或全部内容,以满足各专业技术对统计方法的需要。现代统计分析中的数据计算都是依赖计算机来完成的,但具有统计分析功能的软件很多,限于篇幅,本书选用了大多数人所熟悉的 Excel,该软件也具有强大的统计分析功能。本书的图表制作和大多数的统计计算都是由 Excel 完成的,因此,建议读者在使用本书之前,学会使用 Excel 进行一些基本的统计分析。

本书由马冀、赵养森和罗宏编著。各项目的执笔人是:项目一至项目五由马冀编写;项目六至项目九由赵养森编写;项目十和附录由罗宏编写。

我们力图将本书写得通俗易懂,以便读者更容易接受。如果本书出版后,能从读者或授课教师处得到一些使用信息反馈,将会对完善本书的内容有所帮助,在此提前向这些帮助我们改进工作的人致以衷心的感谢。

编 者

2012年8月

目 录

项目一 管理统计的基本观念——从数据中获得有用信息	1
任务一 广泛应用统计的世界	1
任务二 统计的基本含义	6
任务三 统计学中的概念框架	8
任务四 统计数据的计量尺度	14
任务五 统计数据的基本形式	15
项目二 数据的收集	18
任务一 统计数据的来源	18
任务二 统计调查的方法	19
任务三 收集数据的形式	22
任务四 统计调查的设计	24
任务五 调查问卷的设计	26
任务六 敏感问题的随机化设计	31
项目三 数据的整合	36
任务一 统计数据的预处理	37
任务二 统计数据的整合法	39
任务三 统计数据的质量	45
任务四 数的表达与近似计算	48
任务五 数的完整与修约规则	50
项目四 数据的显示	54
任务一 统计图	55
任务二 统计表	69
项目五 数据综合的反映模式	72
任务一 数据综合的基础模式	73

任务二 对比关系的一般模式	73
任务三 对比关系的扩展模式——指数	79
任务四 综合指数的编制	81
任务五 指数体系的用途	92
任务六 常见指数的计算(几种重要的统计指数)	99
项目六 数据分布的特征测度	108
任务一 集中趋势的测度	109
任务二 离散程度的测度	123
任务三 分配态势的测度(偏态与峰态)	133
项目七 时序数据的基本规律	140
任务一 时序数据的基本构成	141
任务二 时序数据的规范分析(水平分析、速度分析)	144
任务三 时序数据的预测分析	157
任务四 时序数据的景气预测	175
项目八 抽样数据的技术分析	179
任务一 概率抽样方法	180
任务二 参数估计的方法	187
任务三 样本容量的确定	192
任务四 样本数据的假设检验	194
项目九 数据关系的模型构造	204
任务一 两种依存关系,两种计算方法	205
任务二 相关关系的描述与测定	206
任务三 一元线性回归	210
任务四 多元线性回归	216
任务五 列联关系模型	218
任务六 方差分析模型	221
项目十 宏观管理的数据模式——从国民核算到国际对比	233
任务一 国民经济核算数据体系	234
任务二 生产总量指标的计算	239
任务三 国民经济的动态指标	246
任务四 国际经济对比数据模式	248
任务五 综合国力的测定方法	254

任务六 国民经济可持续发展指标体系	259
附录一 统计分析软件简介	262
一、统计分析软件简介	262
二、Microsoft Excel 数据分析方法	263
三、社会科学统计软件包 SPSS	265
附录二 描述统计基本概念、基本原理和基本技能百题训练	267
附录三 用 Excel 生成概率分布表	296
参考文献	305

项目一 管理统计的基本观念—— 从数据中获得有用信息

II► 故事开篇

一位幽默的非医学专家人士说过：服用感冒药物的确能在一个星期内治愈感冒，但人们却忽略了另一个事实，即使不服用抗感冒药物，一般而言，感冒也能在七天内痊愈。新闻调查显示，政府的施政满意度是 59%，可是没有人来问过老百姓的意见，这个数字到底是怎么来的？天气预报说，明天下雨的几率是 75%，就是出门只要带 3/4 把伞吗？医学专家说，SARS 患者的死亡率只有 5%，所以大家都应该放心了吗？这些重要的信息都是统计数字。这些数字到底从何而来呢？人们是否应该全盘接受呢？这就得靠人们的统计素养来作出判断了。我国著名经济学家、人口学家马寅初先生曾经说过一段极为精辟的话：“学者不能离开统计而研究，政治家不能离开统计而施政，事业家不能离开统计而执业。”统计就这么简单——用数据说话。

II► 核心理念

没有统计，其他科学可以存在，但是很渺小。医学若无统计就不是科学。一位资深的海外统计学家说：统计就和柴、米、油、盐、酱、醋、茶一样，存在的时候并不是很突出，一旦不见了，人生就是黑白的了。美国著名科幻作家 H·G·威尔斯曾说过：对于追求效率的公民而言，统计思维总有一天会和读写能力一样必要。

任务一 广泛应用统计的世界

统计数据帮助我们认识世界：我们生活的世界是广泛应用着统计数据的世界。不但电视、广播、报纸、书刊和文件中大量使用着统计数据，就是在日常的生活和交谈中，也不时可以用到和听到一些统计数据。统计数据普遍使用于行政、司法、教育、文化、医药、卫生、人口、劳动和经济生活等各个方面。我国经济和社会发展的成就可通过一系列的统计数据来加以描述。例如，对全面建设小康社会，提出了 16 项具体的指标数据，包括经济、社会、环境和制度四

个方面。在经济主题方面,具体有4项指标;在社会主题方面,具体有7项指标;在环境主题方面,具体有3项指标;在制度主题方面,具体有2项指标。这些指标数据的具体表现如下:

经济主题:具体有4项指标。

第一项指标:人均GDP。人均GDP反映的是一定区域内的经济发展水平。全面建设小康社会的目标是,到2020年,人均GDP目标设定为25000元。按照汇率法计算,这一水平相当于现在的3000美元左右。考虑到通货膨胀和美元币值变动等因素,到2020年,人均GDP可能在4000~5000美元之间。

第二项指标:非农产业就业比重。随着经济的发展,劳动力从第一产业向第二、第三产业转移是一条普遍的规律。根据国际经验,在人均GDP达到3000美元左右时,非农就业比重可以达到60%左右。到2020年,我国非农产业就业比重可望达到60%以上。

第三项指标:恩格尔系数。恩格尔系数是反映居民富裕程度和生活水平及质量的重要指标,它是居民用于食物消费的支出与总消费支出之比。一般来说,收入水平越高,其恩格尔系数越小。根据国际经验,人均GDP在3000美元左右时,居民消费的恩格尔系数均在30%上下。到2020年,我国城市居民消费的恩格尔系数可以降到30%以下,农村居民的恩格尔系数可以降到40%以下。全部人口中最低收入1/5人口消费的恩格尔系数至少应降到50%。

第四项指标:城乡居民收入。居民收入增长主要决定于经济增长。考虑到在全面建设小康社会的进程中,提出的新的发展思想,居民收入增长率与GDP增长率之间的差别有望缩小,城乡居民之间的收入相对差距也有望缩小。可以估计,在2020年GDP翻两番之时,城镇居民人均可支配收入将是2000年的3倍左右,按2000年不变价计算为20000元;农村居民人均纯收入将是2000年的3.5倍,按2000年不变价计算为8000元;城乡居民收入之比由2000年的2.8:1下降到2.5:1左右。

社会主题:具体有7项指标。

第五项指标:基尼系数。基尼系数是测定收入分配差异程度的指标。基尼系数的取值在0和1之间,取值为0时,表示收入分配绝对平均;取值为1时,表示收入分配绝对不平均。根据世界银行的研究报告《中国:促进公平的经济增长》提供的数据,1999年,我国的基尼系数为0.437。现在,多数学者认为,我国的基尼系数已在0.45以上,收入分配不甚合理。在未来20年,我国将加大对于收入分配的调节力度。基尼系数将呈下降之势,到2020年,有可能下降到0.4以下。

第六项指标:社会基本保险覆盖率。社会保障是一种社会稳定机制,是保证全体社会成员基本生存需要的手段,是促进经济、社会持续协调发展的有效工具。全面建设小康社会要求为全体国民提供基本的社会保障,包括基本的医疗保险、养老保险和最低生活保险。到2020年,社会基本保险覆盖率应达到100%。

第七项指标:平均受教育年限。提高居民的受教育水平和劳动者科学文化素质是全面建设小康社会的重要内容。为了从总体上综合反映一个国家或地区的人口受教育情况和人口素质,选用平均受教育年限作为衡量全面建设小康社会教育发展水平的指标。据联合国开发计划署《2001人类发展报告》的资料,2000年,我国15岁以上人口平均受教育年限为6.4年。同年,世界高收入国家为10年,美国为12.1年。我国一般选用6岁和6岁以上人口平均受教育

水平作为判断国民受教育水平的依据。根据第五次人口普查资料,我国6岁以上人口平均受教育年限为7.62年。在全面建设小康社会的过程中,我国将全面实现9年义务教育,一些比较发达的地区可能实现更高水平的义务教育,同时,其他各种形式的教育也将得到较快发展,到2020年,我国6岁以上人口平均受教育年限应达到10年。

第八项指标:出生时预期寿命。预期寿命是一个综合性较强的指标,既能反映社会、经济的进步状况和医疗水平的发展状况,也能从一个侧面反映人们的营养状况和生活质量的改善情况。根据世界银行《2003年世界发展报告》,2000年,低收入国家的出生时预期寿命为59岁,中等收入国家为69岁,中低收入国家为64岁,高收入国家为78岁,我国为71.4岁。我国的预期寿命明显地高于同等发展水平的发展中国家。据估计,到2020年,我国的预期寿命可达到75岁。

第九项指标:文教体卫增加值比重。文化产业的发展是社会发展的重要组成部分,文化是影响社会经济运行方式和条件的重要因素。随着经济社会的发展,文化产业增加值占GDP的比重将不断上升。文化产业覆盖众多不同的行业。根据美国、加拿大和墨西哥三国联合建立的北美产业分类方法(NAICS),大致地说,文化产业包括娱乐业与电子传媒业、印刷业与出版业、旅行与旅游业。在我国的统计体系中,还不存在对于这类产业的系统统计。因此,选用文教体卫增加值占国内生产总值的比重作为反映文化产业的代行指标。2000年,我国文教体卫增加值占国内生产总值的比重为3.6%,文化产业比较落后。随着经济的发展和人们对于文化产业重要性认识的提高,我国的文化产业将得到较快的发展。据估算,到2020年,我国文教体卫增加值占国内生产总值的比重可达到10%。

第十项指标:犯罪率。为居民提供安全的生活环境是社会发展的主要内容。国际上,一般选用暴力犯罪率作为反映居民生活安全环境的逆指标。这里,我们选择刑事犯罪率作为反映居民生活安全环境的指标。2000年,我国平均万人中公安机关立案的刑事案件有29起。2020年,该比率应降至15起/万人以下。

第十一项指标:日均消费支出小于5元的人口比重。日均消费支出小于1美元(购买力平价)的人口比重是国际上通用的反映贫困状况和公平状况的重要指标。目前,我国缺乏类似的统计,但有关城乡人口的贫困标准,可以根据这些标准设计这一指标。现阶段,我国城市居民家庭贫困线为2310元/人,农村居民家庭贫困线为627元/人;城市最低救助标准应为每人每日6.3元,农村最低救助标准应为每人每日1.7元。这些标准是保证贫困人口基本生存需要的最低标准。根据全面建设小康社会的要求,到2020年,城乡贫困人口日均消费支出均不应低于按目前价格计算的5元。到2020年,我国日均消费支出小于5元的人口比重应降到0。

环境主题:具体有32项指标。

第十二项指标:能源利用效率。能源是经济增长的动力,能源消耗的总量及其结构直接影响着环境质量。为了反映经济增长对于能源的依赖程度,以及经济增长可能产生的环境影响,我国选择能源利用效率作为环境主题方面的重要指标。国际上,一般选用千克油当量的产出(购买力平价)作为反映能源利用效率的指标。由于对购买力平价方法存在争议,还由于我国统计年鉴上只存在标准煤的数字,根据世界银行《2001年世界发展指标》提供的有关资料的基础上,利用汇率法换算,以及油当量和煤当量的换算关系,估算了相关国家的能源利用效

率:1998年,我国千克煤当量的产出是0.65美元,世界平均水平是2.32美元,低收入国家0.58美元,中等收入国家1.45美元,上中等收入国家1.82美元,日本是5.87美元,美国是2.66美元。2000年,我国千克煤当量的产出是6.8元(相当于0.82美元),比1998年有所提高,但是与世界平均水平相比还有很大差距。随着技术的进步和新型工业化战略的实施,我国的能源利用效率将进一步提高。到2020年,按2000年的价格计算,我国千克煤当量的产出可达到20元(相当于2.4美元),与目前世界的平均水平相当。

第十三项指标:使用经改善水源人口的比重。能够使用上经过改善的水源是人类文明进步的体现。国际上,用能够使用经过改善的水源的人口比重(简称安全卫生水普及率)作为评价居民用水状况的指标。根据世界银行《2001年世界发展指标》,2000年,我国安全卫生水普及率为75%,世界平均水平为81%,上中等收入国家(人均GNI大于3000美元小于9400美元)为87%,高收入国家为100%。根据国际经验,到2020年,城乡居民的安全卫生水普及率可达到90%左右。但是,根据我国全面建设小康社会要求,到2020年,我国城乡居民的安全卫生水普及率应达到100%。

第十四项指标:环境污染综合指数。环境质量是影响居民生活质量的重要因素。国际上,通常用环境污染综合指数作为评价环境质量的逆指标。环境污染综合指数是将大气、水质、噪声等污染状况综合起来对环境质量进行考察的指数。现在,国内有关方面正在研究环境污染综合指数的计算方法和判断标准。

制度主题:具体有2项指标。

制度是用以协调人类行为的规则和组织。制度涉及许多方面,但是,对于全面建设小康社会来说,廉政建设和政府管理能力是最为重要的两个方面。前者反映了政治文明发展的状况;后者则反映了政府在提供公共服务方面的能力。

第十五项指标:廉政建设。国际上广泛使用由设在柏林的世界性反腐败组织——透明国际设计的腐败指标来评定公共机构的腐败状况。这种方法计算复杂,与调查对象的主观判断关系密切,客观性不太强。结合中国国情,同时又考虑到可操作性以及数据的可获得性,用全国检察机关直接立案的贪污贿赂和渎职案件数与国家机关、政党机关和社会团体就业人数之比作为反映廉政状况的代行指标(姑且称之为“廉政指数”)。2000年,我国万名国家机关、政党机关和社会团体就业人员中检察机关立案的件数为41起,到2020年,应降至10起/万人。

第十六项指标:政府管理能力。政府管理能力涉及公共财政管理、国家储备状况、法律法规建设、社会秩序维护等诸多方面。为体现小康指标体系的简洁性和易操作,我国选用非正常死亡率作为全面小康指标体系中评价政府管理能力的代行指标。这一指标能够在很大程度上反映政府的管理能力和应对危机的能力。非正常死亡包括因交通事故、火灾、安全生产事故,以及自然灾害等造成的死亡人数。由于数据的可获得性,用交通事故和火灾死亡人数的比重估计非正常死亡率,2000年为15%,2020年应降至5%。

统计数据不仅告诉人们事物的现状,还能告诉人们事物已经发生了怎样的变化以及还会发生怎样的变化。为了对不确定的社会经济现象的发展前景进行科学预测,就必须依据大量的统计数据资料。

统计方法帮助人们研究问题:人们不但能从统计数据中认识世界,还可以运用统计方法

解决生产、生活中的许多问题。比如,在收集了大量的某地区的气象数据,并经过统计方法处理以后,不只是风云可测,而且可以作出准确的气象预报。如果要判断一种新药的疗效,离开了统计方法,人们也就无法得知一种新药比原来使用的药物更有效。科学史上大量的例子,充分说明了统计方法在研究自然界和人类社会的规律性方面的积极作用。

奥地利生物学家孟德尔在 1865 年发表了一篇文章,提出了所谓基因的学说(“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。这项伟大发现的过程足以说明统计方法在科学研究中所起的作用。孟德尔用豌豆做试验,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是黄色的和绿色的,孟德尔将两个纯系进行杂交,发现这种黄绿杂交品种所结的豆子全部是黄色的,与黄色纯系没有差异。但将这种杂交体再进行一次杂交而产生第二代时,孟德尔发现某些第二代杂交豆子呈黄色,而另一些则呈绿色,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,到此为止,所得到的还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为基因的实体以控制豆子的颜色。这个实体有两个状态 y (黄色)和 g (绿色),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,而第四种使豆子呈绿色。遗传学上称 y 为显性, g 为隐性。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分别为 yy 和 gg 。杂交一代的基因型则只有一个可能性,即 yg 。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 。前三种是黄色的,而后一种是绿色的,这就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。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 20 世纪 50 年代,基因的存在已经在分子水平上获得了证实。统计方法在科学上的重要用途就是客观地评价某项理论上的结论是否与观察结果相符,以作为该理论是否成立的印证。

概率的问题是从赌博实践中提出来的。有人说:“现代统计学,是由充满灰色与乌烟酒气的豪赌中附带而成”。早在 16 世纪中叶的意大利,数学家卡达罗、天文学家伽利略就对赌博中机遇戏法做过数学论证。卡达罗曾写过一本叫《论赌博》的书,由于当时不愿泄露赌技,直到他死后多年才正式出版。据卡达罗自称,他本人就是一个精明的赌博者。因此,我们有理由相信,卡达罗的著作就是数学理论与赌博实践相联系的产物。伽利略也写过一篇同样题目的论文。在这篇没有注明写作日期的论文里,他解答了一个意大利贵族向他请教的赌博难题。它就是:投掷三颗色子,出现 9 点和 10 点各有 6 种不同的组合,但在经验上,为什么出现 10 点的次数多于出现 9 点的次数呢?从现代概率数学的角度来看是比较容易的。但伽利略当时给他的解答如下:出现 9 点的组合数为 126、135、144、225、234、333;出现 10 点的组合数为 136、145、226、235、244、334。因各种组合出现的机会并不相等。如 333 的组合只有一个途径掷出;而 334 这种组合就有三种不同的途径可以掷出。因此 10 点出现的机会比 9 点出现的机会要多。从这个问题分析的结果来看,伽利略已应用了概率数学的内容。此后,法国的两位数学家帕斯卡尔和费马特,对此也产生了浓厚的兴趣。一天,有一个后来被封为侯爵的、名叫梅累

赌博嗜好者向帕斯卡尔提出一个难题请求解答。这个难题是：甲、乙两人相约赌博，共赌若干局，谁先赢 3 局，就算赢，全部赌金就归谁。设两人赌技相同，若已赌了 3 局，现在一个人赢了 2 局($2 < 3$)，而另一个赢了 1 局($1 < 3$)，赌博终止，应如何分配赌金？对这个问题，立时能想到的分发有两种：一种是平分。这对甲不公平，他多赢一局。另一种是全归甲。这对乙不公平，他虽然暂时落后，仍有起死回生的可能。公平的分发只能是按一定的比例。乍一看会觉得甲、乙按 2 比 1 的比例分配赌金公平，进一步分析发现不然。试想甲、乙不中断赌博，则在至多两局内可见高低。两局有四种可能结果：甲甲、甲乙、乙甲、乙乙（甲甲表示甲连胜两局，其余类推）。根据甲、乙赌技相同的假定，这四种结果是具有同等可能的，而前三种结果都使甲得到最后胜利，只有第四种有利于乙。因此，甲得胜的概率为 $3/4$ ，他应当分到全部赌注的 $3/4$ 才算公平。这似乎将警士捉赌也列为赌博方法之一。1654 年 7 月 29 日，帕斯卡尔把这一难题的解答寄给当时数学名人费马特。他们在通信中几经研商之后，最后得出一般化的解法。对赌博中机遇问题的研究，最终奠定了概率论的基础。但概率论的创立者并不是意大利人，而是法国人。此后打破了赌博数学一向保密的惯例。

众所周知，《红楼梦》一书共 120 回，一般认为前 80 回为曹雪芹所写，后 40 回为高鹗所续，长期以来对这个问题一直有争议。此问题能否从统计上作出论证呢？1985—1986 年，复旦大学李贤平教授带领他的学生做了这项有意义的工作，他们创造性的想法是将 120 回看成是 120 个样本，然后确定与情节无关的虚词作为变量（所以要抛开情节，是因为在一般情况下，同一情节大家描述的都差不多，但由于个人写作特点和习惯的不同，所用的虚词是不会一样的），数出每一回里变量出现的次数作为数据，用多元分析中的聚类分析法进行分类，果然将 120 回分成两类，即前 80 回为一类，后 40 回为一类，很形象地证实了《红楼梦》不是出自同一人的手笔。之后又进一步分析前 80 回是否为曹雪芹所写，这时又找了一本曹雪芹的其他著作，做了类似计算，结果证实了用词手法完全相同，断定为曹雪芹一人手笔。而后 40 回是否为高鹗写的呢？论证结果推翻了后 40 回是高鹗一个人所写。这个论证在红学界产生很大的轰动，他们用多元统计分析方法支持了红学界的观点，使红学界大为赞叹。

由于统计在现实世界中产生了如此大的影响，这就要求从事统计工作的人员需要正确掌握统计方法，即每个从业人员都应该懂得统计，以便运用统计方法更好地认识世界和改造世界。

任务二 统计的基本含义

概念在先

统计在现实世界中虽然十分重要，但是对究竟什么是统计、统计有什么用等问题，人们的感受和认识却很不一致。

对统计的种种理解：

其一，从物价指数来认识统计。国家统计局在《关于××××年国民经济和社会发展的公

报》中披露,××××年社会商品零售价格同上年相比提高了 5.6%。这个数字是否如实地反映了客观实际,人们众说纷纭。有人认为这个数字离实际太远了,甚至举例说:“我买某种商品,比上年至少多付出一半以上的钱,能说这个统计数字可靠吗?”不少人以种种“事实”来论证这个数字的不可靠。但是也有人认为这个数字是根据大量数据,用科学的统计方法计算出来的。因此,它基本上反映了社会商品零售价格变动的真实情况。

其二,从企业管理来认识统计。有人认为没有统计将会影响企业的管理工作;但也有人不然。他们认为,“统计统计,七分估计,三分统计”,没有统计,企业的生产经营活动照常进行。

其三,从统计的作用来认识统计。有人认为统计的作用贵在如实反映情况,必须“是一说一,是二说二”;但也有人认为统计无非是一种工具,它应当听命于人,为人服务,“我要你数字大,你就得大,要你小,你就得小。”此外,不会读统计数字的人认为统计数字枯燥乏味;而善于读统计数字的人却从统计数字中得到丰富的知识。许多人为维护统计数字的真实性而忘我地工作;也有一些人却在那里弄虚作假。许多正确的统计结果给人们的生活带来很大的方便,如气象预报、健康状况调查等;相反,也有不少错误的统计使人们上当受骗,如鸡血疗法治愈多种疾病的疗效在 95%以上、人血馒头能医治痼病等。有人说,统计扩展人们的视野;也有人说统计束缚人们的手脚。

不仅在日常生活中人们对统计有种种不同的理解,就是在学术领域里,人们对统计的认识也有差别。早在 1869 年的第七次国际统计会议上,比利时统计学家凯特勒就做过一个统计,当时对统计的定义,就有 180 种之多。虽然后来再没有人做过类似的统计,但有人做过这样的概括:世界上有多少统计学家,就有多少关于统计的定义。

为什么理解不同:为什么对统计会有这么多认识上的分歧呢?原因可以概括为以下两个方面。

其一,在不同的场合,统计一词的具体含义不同。统计有时候是作为动词用的,比如,“统计统计吧”;有时候又作为名词用,比如,“这个统计很好嘛!”统计有时还被当做“统计科学”的同义语,比如,“张三是学统计的”;而“李四是搞统计的”,又作为“统计工作”使用。在工厂里,经常可以听到这样的介绍:“这位是陈统计”。这里统计又被理解为“搞统计工作的人”,也就是统计工作人员。在许多情况下,统计又成了“统计资料”或“统计数据”的简称。

即便在同一具体的语言环境里,每个人的理解也各有不同。对统计资料的作用,有的人持肯定的态度;有的人加以怀疑或否定。有的人对统计工作的理解比较正确、全面;有的人则难免错误、片面。至于在学术领域里,人们对统计学的对象和性质就各有不同的见解,那更是必然的现象。学术上不同观点的争论是件好事,它有利于学术研究的繁荣,从而也推动统计工作向前发展。

其二,在我国,生产力水平还不很高,对信息的要求也就相对较低,加之统计在实际工作中的地位并没有完全确立,统计的重要性还没有引起人们的重视。长期以来,人们受小生产观念的影响,也没有真正懂得统计的意义和作用,不善于也不习惯于运用统计这一武器来做好经济工作。有些人有时喜欢隐瞒或夸大某些东西;也有些人以实用主义的观点来对待统计,要统计数据适合自己的“口味”,他们不是根据统计数据来作出结论,而是事先有了结论再去选择或纂

改统计数字。当然,由于统计工作本身的一些缺点,还不能完全满足经济建设的各方面对统计数字的需要,从而也影响了统计作用的充分发挥,也就影响了人们对统计的正确理解。信息化的社会经济环境将会不断改变这种观念。

统计的定义:对“统计”一词究竟应该怎样正确理解呢?必须从独立的意义和统一的概念上对统计一词做一个基本的解释。所谓“独立”,就是撇开“统计”这个词的具体语言环境和各种应用场合,单就“统计”这个双音词,理解它的确切含义;所谓“统一”,就是不管各种不同门类的统计学和各种不同学派对统计一词的不同理解,找出它们的共同点来做一般性的解释。

在这个意义上的统计,就是人们日常生活中经常接触到的“据统计……”、“据不完全统计……”的含义。这个含义又该如何理解呢?它既不是指在具体语言环境下的统计资料、统计工作、统计人员和统计学科的含义,也不涉及在学术意义上的方法论科学还是实质性科学或是自然科学、社会科学还是通用科学的问题。这个意义上“统计”一词的含义是调查、整理、分析和解释数据活动的总称,也可以把它理解成一种对客观事物进行调查研究的活動。这就是“统计”一词的基本解释。众所周知,理论是实践的高度概括和总结。因此,《大不列颠百科全书》对统计学下了一个定义,其基本含义也包含了这些基本范畴。

定义:统计学(statistics)是收集、分析、表述和解释数据的科学。这一定义揭示了统计是一套处理和分析数据的方法和技术,由此看来,统计离不开数据,在取得数据后,为满足分析的需要,还要对数据进行整合,而后再对数据进行分析 and 解释。因此,统计研究的过程就可以描述为:收集数据—整合数据—分析数据—解释数据。但在实际应用中,统计又根据掌握数据信息的不同程度,会采用不同的数据处理方法。由此得出以下两个推论:

推论 1:描述统计(descriptive statistics)是研究完全信息的数据收集、整合和表述以及分析的统计学分支。

推论 2:推断统计(inferential statistics)是研究不完全信息条件下,如何利用样本数据来推断总体数据特征的统计学分支。

任务三 统计学中的概念框架

基本观念是解决问题和思考问题的定位模式。核心概念是贯穿整个理论和方法体系的概念框架。统计的概念框架就是建立在统计基本观念基础之上的一系列理论和技术范畴。

总体和样本:总体和总体单位是统计学的核心概念,总体是统计认识问题的视角和目的;总体单位是构成总体的元素和统计认识问题的出发点。

定义:包含被研究现象全部个体(或数据)的集合,称为总体(population)。

总体是具有某种共同性质的许多个体(或元素)组成的集合。例如,全国工业企业是一个统计总体。因为,每个工业企业的经济职能是相同的,它们都是进行工业生产活动的基层单位。各个个体在某一点上的共同性是形成统计总体的必要条件,也是统计总体的一个重要特征。在统计活动中,总体范围的确定,有时是比较容易的,例如,要检验一批灯泡的

使用寿命,这一批灯泡构成的集合就是总体,每个灯泡就是构成总体的单位,总体范围很清楚。但有些场合总体范围的确定则比较困难,例如,对于新推出的一种饮料,要想知道消费者是否喜欢,首先必须弄清哪些人是消费的对象,就是要确定构成该饮料的消费者这一总体,事实上这很难确定。当总体的范围难以确定时,可根据研究的目的来定义总体。根据总体所包含的单位数目是否可数,可将总体分为有限总体和无限总体。有限总体是客观存在的总体,其范围和构成单位都是能够明确加以确定的,而且总体包括的单位数目也可以点计清楚。例如,一个国家或一个地区的人口、一个部门所属的单位、一个城市的学校、一个企业在一定时期内生产的产品等都是有限总体。而无限总体,有两种不同的定义:一种是客观现实中不存在的总体,例如,科学试验统计,试验一次,得到一个试验结果,而试验的次数是可以无限增加的。工业生产技术试验、农业田间试验等都属于这一类。另一种是总体虽然是客观现实中存在的,但目前人们还无法确定它的界限,或者总体虽然有限,但构成总体的单位却是无限的。例如,气象统计、天体统计、连续大量生产的某种小件产品的统计就属于这一类。

一般地说,在自然技术统计中,认识的对象大都属于无限总体;而社会经济统计的认识对象大都属于有限总体。对这两类不同的总体,统计的内容、方法和具体过程是有所不同的。例如,对于无限总体,直接不存在总体总量的问题;而对于有限总体,则必须了解它的总量是多少。这就产生了如何确定总体和总体单位的范围及界限问题。这些问题,有的是很复杂的,需要专门研究来认识。另外,要了解无限总体,不可能采用全面调查的方法,只能调查其中一小部分单位,据以推断总体。至于要了解有限总体,不但可能采用全面调查的方法,而且在有些情况下,还必须采用全面调查的方法。当然,也可以采用抽样调查或其他调查方法。如何把多种多样的调查方法结合起来加以很好的运用,是社会经济统计必须认真研究的问题,如人口普查(全面)及人口抽查(部分)。凡是调查总体的一小部分单位时,往往要根据这些局部资料来推算总体资料。

综上所述,可以将总体的基本特征概括如下:① 同质性,就是总体所有单位都必须具有某一共同的品质或数量特征表现。即在一个总体中各个总体单位至少有一个标志是相同的。② 大量性,就是构成总体的单位数要足够多,但并不是单位数很多,而是足以说明问题的单位数。③ 差异性,就是总体单位在具有共同性质的同时,还表现在这个可变的品质或数量特征上的差异,可以说差异性才是进行统计的前提条件。这三个特征缺一不可,才能形成统计总体。有了统计总体,才能进行一系列的统计计算和分析研究。例如,要研究国有企业的生产经营情况(同质性),若从总数 5 000 个企业中抽选出有代表性的 100 个企业(大量性),而这些企业的行业不同、规模不同,资金占用也不同(差异性)。实际上所抽出的部分单位也是一个小规模的整体。

总体和总体单位的概念并不是固定不变的,它随着研究目的的不同是可以相互转换的。例如,某市机械工业局所属的企业是一个总体,每个机械工业企业是这个总体中的一个单位,这个总体是研究该市机械工业企业状况的对象。现在假设把研究领域扩大到该市以局为单位的全部工业企业,则除机械工业局外,还有其他许多工业局,如纺织工业局、轻工业局、化工局、冶金局等。这时,机械工业局又变成全部工业局这个新的总体中的一个单位了。