

University Textbooks in Economics & Management

经济学·管理学学科共同课教材

统计学

STATISTICS

周恒彤 编著



东北财经大学出版社

Dongbei University of Finance & Economics Press

University Textbooks in Economics & Management

C8/168D

2007

统计(910)目次第五版

统计出学大经统北京：北大一，著编组同国，学书统

11.5005

(统计明同其科学地国普，学书统)

经济学·管理学学科共同课教材

统计学

STATISTICS

周恒彤 编著



东北财经大学出版社

Dongbei University of Finance & Economics Press

大连

© 周恒彤 2007

图书在版编目(CIP)数据

统计学 / 周恒彤编著. — 大连 : 东北财经大学出版社,
2007. 11

(经济学·管理学学科共同课教材)

ISBN 978 - 7 - 81122 - 196 - 1

I. 统… II. 周… III. 统计学 - 高等学校 - 教材 IV. C8

中国版本图书馆 CIP 数据核字(2007)第 170985 号

东北财经大学出版社出版

(大连市黑石礁尖山街 217 号 邮政编码 116025)

总编室:(0411)84710523

营销部:(0411)84710711

网 址:<http://www.dufep.cn>

读者信箱:dufep@dufe.edu.cn

大连金华光彩色印刷有限公司印刷 东北财经大学出版社发行

幅面尺寸:170mm × 240mm

字数:352 千字

印张:15 1/4

2007 年 11 月第 1 版

2007 年 11 月第 1 次印刷

责任编辑:孙晓梅 王 莹

责任校对:那 欣 何 群

封面设计:冀贵收

版式设计:钟福建

ISBN 978 - 7 - 81122 - 196 - 1

定价:26.00 元

前 言

本书可用做普通高等学校本科经济学、管理学学科“统计学”课程教材(大约48学时可学完本书)。

客观世界的各种现象都是由许多因素共同作用所形成的结果。人们在实践活动中,常常需要预见一件事情将要发生何种结果或是猜测一件已经发生的事情为何种结果。这时,由于人的认识能力的限制,绝大多数情况下只能控制事情的一部分影响因素,而对于余下的大量的影响因素,则无法加以控制。由于无法掌握那些未被控制的影响因素是如何对事情施加影响的,因而人们无法对事情的结果做出确切的(确定的)预见或猜测,于是,呈现在人们面前的各种事件总是存在着各式各样的不确定性。幸好,这些不确定性之中有一定的规律可循,并且,通过来自这些事情的信息,其中的规律可以被认识。统计学把这些信息叫做数据。统计学(数理统计学)教给人们,怎样通过数据去认识不确定性中的规律性,以便使我们对事情的结果所作的预见或猜测尽量接近实际。

本课程的授课对象是经济学、管理学学科的学生。从经济学、管理学这一研究领域出发,本课程应该教给学生,怎样通过社会经济调查数据去认识社会经济现象不确定性中的规律性,以便帮助其在社会经济活动中做出正确的决策。然而,数理统计学是在自然科学和工程技术科学的土壤中诞生的,将其用于处理社会经济调查数据,尚有一些社会经济领域特有的问题需要回答和解决。在这方面,本书给予了足够的重视。

任何一种统计方法都有其自身的应用场合、应用条件和功能作用。假若在应用统计方法时忽视了这些问题,把统计方法用于不相匹配的数据,用统计方法去做它不能胜任的事情、去回答它不能回答的问题,就会犯错误,就会得出错误的研究结论。我们把这种情况叫做统计陷阱。本书在介绍各种统计方法的时候,十分注意指出统计陷阱的所在以及防止落入陷阱的方法。

在统计学的历史上,先于数理统计学产生的统计学在当时被称做“政治算术”。其基本内容是用数字去说明国家的社会经济状况。后来,在“政治算术”的基础上,逐渐发展为由各国专门的政府机构从事政府统计工作,负责向政府管理部门和公众提供统计数据,描述社会与经济发展的现实状态。作为面向经济学、管理学学科的学生开设的“统计学”课程,无疑应当介绍这方面的有关知识。这类内容仅仅在社会经济领域的“统计学”课程中才有,在自然科学和工程技术科学领域的“统计学”课程中则是没有的。

这样,本书就包含了两个部分的内容:一是属于“政治算术”范畴的内容;二是属于数理统计学范畴的内容。这二者之间有什么联系呢?事实上,社会经济统计调查数据具有双重的用途,可以同时应用于上述两个统计学范畴。本书对这一观点进行了必要的阐述,并基于这一观点构建了全书的内容体系框架。

本书的内容体系是这样安排的:

首先,由于统计学是数据的科学,不论是描述社会与经济发展的现实状态,还是力图准确地把握具有不确定性的社会经济现象,都离不开数据,都要依赖从数据中提取的信息。因

此,本书的第1章,在简要地叙述了什么是统计学之后,分类详细介绍了各种类型的数据。

其次,人们的一切社会活动和经济活动,都不可避免地要与不确定性打交道。为了使自己的决策正确,为了使自己的各种活动多一些成功,少一些失败,就要研究不确定性。人类在研究不确定性的长期实践中,逐渐形成了一个研究不确定性的完整的理论体系(称之为“概率论”)。在这个理论体系中,规定了一套解读不确定性的“语言”。既然统计学意在运用数据去与不确定性打交道,那么,在学习统计学之前,自然应当首先熟悉上述这套解读不确定性的“语言”。因此,作为知识准备,本书的第2章介绍了概率论的若干知识,旨在使学生初步掌握描述和认识不确定性的技能。

再次,本书从第3章开始,具体介绍了统计学的有关内容。

统计活动要从收集数据开始。在社会经济领域中收集数据的工作叫做统计调查。第3章介绍了怎样在弄清调查意图的基础上明确地定义要调查的问题、确定调查单位和调查条目,另外还介绍了抽样调查的基本知识。

对于统计调查数据,一方面可以对其进行描述性整理,用来描述调查对象总体的现实状态;另一方面可以用来推断产生这些数据的母体的统计特征。第4章的内容属于前一个方面;第5、6章的内容属于后一个方面。

第7章的内容与第4章属于同一个范畴,与第4章不同的是:第4章所面对的调查对象总体是简单总体,可以直接用调查对象总体的单位数目来表示它的规模;而第7章所面对的则是由相互之间不能进行单位数目加总的不同的简单总体复合而成的复合总体。

第8章是与第3章至第7章不同的另外一类问题:第3章至第7章所处理的是截面数据,而第8章所处理的是时序数据。

最后,各章所配的思考题和习题,有些用于对教学内容的复习和巩固,有些则是教学内容的进一步延伸。学生应独立思考,亲自动手去完成这些题目。

本书是以我在天津财经大学讲授本科“统计学”课程所用的讲稿为基础写成的。在此,我要感谢我的学生和同事。他们曾经对课程内容提出过许多宝贵的、富有建设性的意见。

另外,我特别要诚挚地感谢东北财经大学出版社编辑一部主任孙晓梅女士,她的热情邀请和积极鼓励对本书的完成和出版起了决定性的作用。

周恒彤谨识
2007年4月

目 录

第1章	导言	1
§ 1.1	什么是统计学	1
§ 1.2	变量和数据	3
§ 1.3	社会经济领域中的统计学	7
	本章要点	8
	思考题和习题	9
	推荐阅读文献	9
第2章	随机现象以及对它的概率描述	10
§ 2.1	关于随机现象的若干概念	10
§ 2.2	事件和概率	13
§ 2.3	随机变量及其分布	23
§ 2.4	几种常用的概率分布	34
§ 2.5	两个随机变量的联合概率分布	40
	本章要点	42
	思考题和习题	43
	推荐阅读文献	46
第3章	统计调查	47
§ 3.1	统计调查的一般问题	47
§ 3.2	统计指标的类型	50
§ 3.3	统计分组	55
§ 3.4	制订调查方案和采集数据	63
§ 3.5	抽样调查	67
	本章要点	80
	思考题和习题	81
	推荐阅读文献	85
第4章	数据的描述性整理	86
§ 4.1	根据事先拟订的调研方案整理数据	86
§ 4.2	数据分布状态的描述	88
§ 4.3	数据的代表值	96
§ 4.4	数据的散布特征数	104
	本章要点	108
	思考题和习题	109
	推荐阅读文献	118
第5章	推断统计概说	119
§ 5.1	可重复随机试验的随机样本	119
§ 5.2	参数估计概说	120
§ 5.3	假设检验概说	123

	本章要点	132
	思考题和习题	133
	推荐阅读文献	136
第6章	推断统计的几个课题	137
	§ 6.1 单总体参数的检验	137
	§ 6.2 两个总体参数的比较	139
	§ 6.3 单因子方差分析	141
	§ 6.4 分布拟合检验	145
	§ 6.5 两个分类变量的独立性检验	148
	§ 6.6 回归分析	151
	本章要点	158
	思考题和习题	159
	推荐阅读文献	163
第7章	统计指数和因素分析	165
	§ 7.1 统计指数的一般问题	165
	§ 7.2 综合指数	167
	§ 7.3 平均指数	171
	§ 7.4 指数的推算、指数数列	176
	§ 7.5 因素分析	180
	本章要点	189
	思考题和习题	190
	推荐阅读文献	195
第8章	时间数列数据研究	196
	§ 8.1 时间数列的描述性指标	196
	§ 8.2 时间数列数据变动规律图示	206
	§ 8.3 水平型时间数列和预测	209
	§ 8.4 趋势型时间数列和预测	212
	§ 8.5 季节型时间数列和预测	214
	§ 8.6 混合型时间数列和预测	216
	本章要点	221
	思考题和习题	222
	推荐阅读文献	225
常用统计数值表		226
	附表1 累积法平均增长速度(递增速度)表(节选)	226
	附表2 标准正态变量密度曲线下的面积	227
	附表3 χ^2 值表	228
	附表4 $F_{m, \infty}$ 值表	229
	附表5 F 值表	230
	附表6 t 值表	234
	附表7 随机数字表	235
参考文献		236

第1章

导言

§ 1.1 什么是统计学

我们先来看几个例子。

例 1—1 想要知道两个天体之间的距离,用同一台天文望远镜,采用同样的观测方式,反复观测了若干次,获得了一系列观测数据。观测者用这些观测数据获得了关于天体之间距离的知识。

例 1—2 有一台机器,在一组给定不变的条件下连续不断地加工出若干件产品。我们关心产品的尺寸。由于在加工过程中存在着许多无法控制的偶然因素的干扰,使得所加工出来的产品的尺寸不会完全相同。不过,我们有理由认为,在给定条件下工作的这台机器所加工产品的尺寸会有一个稳定值(或称做统计规律值)。人们通过加工出来的这些产品的尺寸数据获得了产品尺寸统计规律的知识。

例 1—3 某地区商业银行决定从本年年初起降低储蓄存款利率。现在对该地区全体居民户作调查,收集每个居民户去年(降低利率前)和今年(降低利率后)各年的年平均储蓄存款额^①。研究人员通过这些数据对下列假设是否成立进行了检验。所检验的假设是:“降低利率后比降低利率前具有居民参加银行储蓄的热情下降的统计规律”。

例 1—4 某学生即将参加英语六级考试,想要预测在这次考试中他能否通过。为此,收集了该学生英语水平方面的有关数据。例如,掌握英语单词的数量、阅读英语的速度、以往参加各种类型英语考试的成绩等等。人们根据这些数据对该生通过本次英语六级考试的可能性(概率)做出了预测。

例 1—5 某人胃部不适,请医生诊断他患了何种疾病。为此,医生收集了该患者有关的临床数据。例如,是否有上腹疼痛、是否有黑粪、是否有腹部肿块、是否发热等等。根据这些数据,医生对该患者胃部实际状态的各种可能性(胃癌、胃部良性病变、其他胃部病变、正常)的概率做出了猜测(判断)^②。

上述 5 个例子都属于统计学的研究范畴。这 5 个例子中共同的活动内容是:有一个具有**随机性**的研究对象,我们想要了解这个研究对象的某种或某些种未知特征。为此,收集来自这个研究对象的数据。统计学教给我们,怎样通过数据去认识研究对象的未知特征。

^① 一个居民户的年平均储蓄存款额是指该户在年内在各个时点上储蓄存款额的平均值。

^② 本例见周怀梧、孙伟民:《临床计量医学》,1~3页,上海,上海医科大学出版社,1999。

上面的表述中所说的随机性,指的是下面两种情况之一:事情还没有发生,将要发生的结果呈现不确定性;事情已经发生,但所发生的为何种结果我们不知道,对所发生的结果的猜测呈现不确定性。

在统计学中,把来自研究对象的数据叫做**样本**(或叫做**子样**^①),把产生数据的那个研究对象叫做**总体**(或叫做**母体**)。

用这些术语把上述5个例子共同的活动内容更简单地概括为:**用样本(子样)推断总体(母体)**。解决此类问题的统计学理论与方法,通常被称做**推断统计学**。

还有与上述五例不相同的另外一种统计活动。我们仍然先看例子。

例1—6 在例1—2中,我们只想描述已经加工出来的那些产品。于是,在测量每一件产品尺寸的基础上,计算出整批产品的平均尺寸、尺寸符合技术标准要求的产品件数所占百分比等等(今后把它们叫做统计指标)。

例1—7 在例1—3中,我们只想描述该地区全体居民户在降低利率前后储蓄存款额的实际状态。于是,在收集每个居民户去年(降低利率前)和今年(降低利率后)各年的年平均储蓄存款额数据的基础上,分别计算出该地区全体居民户在降低利率前后年平均储蓄存款额的户均值、降低利率以后与以前这两个户均值之间的相除之商和相减之差。

我们注意到,例1—6与例1—2的不同之处在于,例1—2的研究对象包括加工产品行为所有可能发生的结果,而例1—6的研究对象则仅仅是上述所有可能发生的结果中真正发生的现实的结果。例1—7与例1—3的不同之处在于,例1—3的研究对象包括降低利率前(或后)居民户年平均储蓄存款额所有可能发生的结果,而例1—7的研究对象则仅仅是上述所有可能发生的结果中真正发生的现实的结果。通过对例1—6中的研究对象和例1—7中的研究对象计算统计指标,描述了客观现象的现实状态。

其实我们知道,例1—6所描述的那些产品实际上是例1—2总体的样本。例1—7所描述的全体居民户各自实际发生的年平均储蓄存款额在降低利率前后的两组数据实际上分别是例1—3所考虑的居民户在降低利率前所有可能发生的年平均储蓄存款额总体和居民户在降低利率后所有可能发生的年平均储蓄存款额总体这两个总体的样本。例1—6和例1—7实际发生的这些数据,若回到例1—2和例1—3中,它们将作为样本,担负着推断总体的任务;而在例1—6和例1—7中,它们仅仅是用自己描述自己。

例1—6和例1—7共同的活动内容是:**用样本数据描述样本自身**。讲述此类问题的统计学理论与方法,通常被称做**描述统计学**。

在描述统计学中,如果样本中的元素是一个一个的社会经济活动单元,那么,把对这些社会经济活动单元收集得到的原始数据加以综合所算得的统计指标,描述了社会经济发展的现实状态。

例1—6和例1—7中的研究对象通常被叫做**调查对象总体**,有的文献将它们称为**有限总体**。在这里,一方面,用“调查对象”、“有限”这种限定词把此种称谓与例1—2和例1—3中的“总体”区分开;另一方面,强调现在所说的“总体”(“调查对象总体”、“有限总体”)指的是某一项统计调查活动中我们所关注的全体调查对象的集合体。相对于例1—6和例1—7中的“调查对象总体”、“有限总体”而言,有的文献把例1—2和例1—3中的“总体”叫做**调查的理论总体**,有的文献也把它叫做**超总体**。“调查对象总体”(“有限总体”)则被看做调查的理

^① 在讲述抽样技术的文献中还会见到“子样本”这样一个类似的术语。“子样”与“子样本”意义不同。“子样”所表达的是这样一个意思:总体是“母亲”,样本是来自总体的“孩子”。“子样本”又叫“次级样本”,指的是从所抽取出来的初级样本中进一步抽出的样本。

论总体(超总体)的随机样本^①。

可是,也有些文献(以及本书的后面某些章节)在谈到例1—6和例1—7那样的“调查对象总体”、“有限总体”的时候,并不使用“调查对象”、“有限”等限定词,而一律把它们都叫做“总体”。因此,当我们见到“总体”(或“统计总体”)这样的用语时,须注意区别,它所指的是例1—1~例1—5中的总体,还是例1—6和例1—7中的调查对象总体。

总结上述7个例子,我们这样来表述统计学:统计学是一门关于数据的方法论科学。它提供了一套概念、原则和方法,用来指导人们怎样收集数据、怎样分析数据、怎样由数据得出结论。这种指导是在两个基本方向上进行的:一是,怎样依据数据所提供的信息,了解随机现象中我们所关心的各种未知特征;二是,怎样用数据描述客观现象的现实状态。

§ 1.2 变量和数据

对研究对象所观察的量叫做**变量**,它是研究对象某种特征的名称。如例1—1中“天体之间的距离”;例1—2中“产品的尺寸”;例1—3中“居民户降低利率前年平均储蓄存款额”和“居民户降低利率后年平均储蓄存款额”;例1—4中的学生“掌握英语单词的数量”、“阅读英语的速度”、“以往参加各种类型英语考试的成绩”;例1—5中的患者“是否有上腹疼痛”、“是否有黑粪”、“是否有腹部肿块”、“是否发热”;例1—6中单件产品的“产品尺寸”、“尺寸是否符合技术标准要求”,整批产品的“平均尺寸”、“尺寸符合技术标准要求的产品件数所占百分比”;例1—7中单个居民户的“居民户降低利率前年平均储蓄存款额”、“居民户降低利率后年平均储蓄存款额”,整个地区全体居民户的“降低利率前户均年平均储蓄存款额”、“降低利率后户均年平均储蓄存款额”、“降低利率后与降低利率前相比户均年平均储蓄存款额变动百分比”、“降低利率后与降低利率前相减户均年平均储蓄存款额增加绝对额”等等上述这些用来概括研究对象这种或那种特征的名称都叫做变量。

一个特定的研究对象某种特征的具体表现叫做**变量值**,或叫做**数据**。

可见,第一,数据必须要和变量联系在一起才有意义,也可以说,变量是数据的名称,数据必须要有名称。第二,收集数据之前必须先定义变量,变量指明了我们关心研究对象的何种特征并规定了这个特征的内涵和外延。

数据有原始数据和派生数据之分。观察研究对象所得到的直接结果叫做**原始数据**。对原始数据进一步加工所得到的数据叫做**派生数据**。例1—1~例1—7中所说收集数据,都是指收集原始数据。例1—6和例1—7中用收集到的原始数据所计算的统计指标是派生数据。今后若无特别的说明,所说的数据都是指原始数据。

一、四种不同的计量尺度以及与之相对应的四种变量和四种数据

数据是对客观现象进行观察(或叫做观测、计量)所得到的结果。进行计量,要使用一定

^① 在有限总体概率抽样(从例1—6和例1—7的调查对象总体中抽取概率样本来估计这两例中的统计指标数值,这就是有限总体概率抽样,我们将在第3章介绍这方面的基本知识)理论中,人们通常把有限总体(调查对象总体)的统计指标数值看做固定不变的,用概率样本来估计这个数值,这叫做**随机化推断的理论框架**。除此以外,还有另外一种理论框架,叫做**模型化推断的理论框架**,是由布鲁尔(K. W. R. Brewer)和罗亚尔(R. M. Royall)提出来的。布鲁尔的论文是:《有限总体中的比率估计:在存在随机过程的假定下所推断出的几个结果》,载《澳大利亚统计杂志》,1963(5),93~105页;罗亚尔的论文是:《论在某些线性回归模型条件下的有限总体抽样理论》,载《生物统计》,1970(57),377~387页。这种理论框架把有限总体(调查对象总体)看做一个超总体的随机样本,把有限总体(调查对象总体)的统计指标看做由该超总体的某个适当的数学模型所定义的随机变量,至于我们所面对的那个特定的有限总体(调查对象总体)的统计指标数值则仅仅是上述模型的一个具体实现。以上内容可参见[美]W. G. 科克伦:《抽样技术》,张尧庭、吴辉译,230~231页,北京,中国统计出版社,1985;秦怀振:《抽样调查中若干理论与实践问题的研究》,7~8页,北京,中国统计出版社,2003。

的计量尺度,不同类型的变量须用不同的计量尺度来计量,因而,变量和数据可以按计量尺度来划分类型。

例 1—8 对某从业人员几个变量的计量结果见表 1—1。

表 1—1

某从业人员几个变量的计量结果

变量名称	计量结果
(1) 性别	男
(2) 最终学历	硕士研究生
(3) 计算机能力测试成绩	90 分
(4) 月收入	2 000 元
(5) 身高	1.70 米

在表 1—1 中可以看到,变量 1 和变量 2 的计量结果比较粗略,所记录的仅仅是对对象的属性;变量 4 和变量 5 的计量结果则比较精确,所记录的是对象数量特征的测度值。

计量学中有四种计量尺度:定类尺度、定序尺度、定距尺度、定比尺度。这四种计量尺度的计量精确程度不同,排在前面的尺度计量较粗略(称之为较低级的尺度),排在后面的尺度计量较精确(称之为较高级的尺度)。与计量时所使用的尺度相对应,变量(数据)也就有四种不同的类型,即定类尺度变量(数据)、定序尺度变量(数据)、定距尺度变量(数据)、定比尺度变量(数据)。

(一) 定类尺度变量和定类尺度数据

定类尺度(又叫**列名尺度**、**名义尺度**)是通过确认被测对象所属类别来测度被测对象时所使用的尺度。具体做法是:把变量所有可能发生的结果(所有可能发生的取值)划分成若干个平等并列的、穷尽并且互斥的类别,确认被计量对象属于其中哪一个类别。用这种方法进行观测的变量叫做**定类尺度变量**,对定类尺度变量进行观测得到的结果叫做**定类尺度数据**,这种数据的计量尺度层次最低。

这里,所谓“平等并列”,意思是说各类之间没有等级顺序的差别,各类的排列顺序可以改变;所谓“穷尽并且互斥”,意思是说变量所发生的任意一个结果应当能够属于其中的某一个类别,并且只属于这一个类别。

不同的定类尺度数据反映了对象属性类别的不同。定类尺度数据只能分类,相互之间不能比较大小。

在表 1—1 中,“性别”是定类尺度变量,相应的计量结果是定类尺度数据。

(二) 定序尺度变量和定序尺度数据

定序尺度是通过确认被测对象所属级别来测度被测对象时所使用的尺度。具体做法是:把变量所有可能发生的结果(所有可能发生的取值)划分成若干个穷尽并且互斥的有序级别,确认被计量对象属于其中哪一个级别。用这种方法进行观测的变量叫做**定序尺度变量**,对定序尺度变量进行观测得到的结果叫做**定序尺度数据**,定序尺度数据的计量尺度比定类尺度高一个层次。

定序尺度数据之间不但有属性分类的差异,而且有属性顺序的差异。所以说,定序尺度对事物的计量比定类尺度精确一些。不过,不同的定序尺度数据仅仅反映属性顺序的不同,而不反映属性差别的大小。

定序尺度数据不但能够分类,而且能够在相互之间排序。但是,相互之间不能进行加、

减、乘、除等数学运算。

在表 1—1 中,“最终学历”是定序尺度变量,相应的计量结果是定序尺度数据。

(三) 定距尺度变量和定距尺度数据

定距尺度(也称间隔尺度)是通过确认被测对象所属级别的位置与某一指定位置之间的距离来测度被测对象时所使用的尺度。具体做法是:对变量所有可能发生的结果(所有可能发生的取值)赋以全体或某一区间内的实数值(有时可能只需使用整数,有时可能只需使用自然数),并指定其中的某一个位置为 0,然后确认被计量对象的位置与 0 之间的距离。用这种方法进行观测的变量叫做**定距尺度变量**,对定距尺度变量进行观测得到的结果叫做**定距尺度数据**,定距尺度数据的计量尺度比定序尺度又高了一个层次。

定距尺度的计量结果表现为数值。例如,温度的计量结果表现为以摄氏或华氏的“度”为计量单位的数值,考试成绩的计量结果表现为以百分制的“分”为计量单位的数值等等。

定距尺度所采用的是算术尺度。因此,1 摄氏度与 2 摄氏度之间的间隔等于 5 摄氏度与 6 摄氏度之间的间隔,0.1 摄氏度与 0.2 摄氏度之间的间隔等于 0.5 摄氏度与 0.6 摄氏度之间的间隔等等。这样,我们就可以用给定的计量单位确切地描述两个统计数字之间的距离。例如,给定用摄氏度作计量单位来计量温度,我们便可以说,2.2 摄氏度与 3.5 摄氏度之间的间隔等于 $3.5 - 2.2 = 1.3$ 计量单位长度(即 1.3 摄氏度)。

定距尺度的取值,在实质上与定类尺度、定序尺度相同,也是对客观事物某一种状态所赋予的一个符号。只不过,在定距尺度中,我们可以确切地说出两个“符号”之间的间隔长度,这是它比定类尺度、定序尺度的“高级”之处。

需要注意的是,在定距尺度中,数值“0”与其他数值一样,只不过是一个符号,也只不过标示了客观事物的其中一种状态。在这里,对“0”的使用是主观的,也就是说,究竟哪一种状态用“0”来标示,是人为规定的。人们可以基于不同的研究需要,把不同的状态规定为 0(对温度的计量是一个十分典型的例子)。可见,在定距尺度中,没有绝对的 0 点。“0”并不意味着“没有”,并不意味着“不存在”。

由于定距尺度的上述特点,使得定距尺度数据不但能够分类、能够在相互之间排序,而且能够在相互之间进行加、减的数学运算。但我们规定,在定距尺度数据之间不能进行乘、除的数学运算。

在表 1—1 中,“计算机能力测试成绩”是定距尺度变量,相应的计量结果是定距尺度数据。

(四) 定比尺度变量和定比尺度数据

定比尺度(也称比率尺度)是用被测对象的实际量与某种给定计量单位的比值来测度被测对象时所使用的尺度。例如,我们测得被测对象的长度是 8 厘米,这个结果的计量意义是:被测对象的实际长度除以计量单位 1 厘米的长度,所得比值为 8。用这种方法进行观测的变量叫做**定比尺度变量**,对定比尺度变量进行观测得到的结果叫做**定比尺度数据**。定比尺度数据的计量尺度比定距尺度更高了一个层次。

定比尺度的计量结果和定距尺度一样,也是表现为数值。例如,人口数的计量结果表现为以“人”(实际上是以“个”)为计量单位的数值,居民消费支出的计量结果表现为以人民币“元”为计量单位的数值等等。

定比尺度用于计量事物的数量多少、规模大小。用定比尺度计量事物时,要给“0”一个固定的意义,并以此作为计量的基础。例如,计量人数时,把“没有人”规定为 0;测量物体的

长度时,把测量的起始点规定为0;计量企业的资本金时,把既没有资本金又没有负债规定为0,拥有资本金的数额用正数表示,负债的数额用负数表示。

由于在使用定比尺度时规定了固定的0点,所以定比尺度数据便可以和实数轴上的点一一对应,从而使得定比尺度数据不但能够分类、能够在相互之间排序、能够在相互之间进行加、减的数学运算,而且能够在计算结果有实际意义的前提下进行实数的各种运算。

在表1—1中,“月收入”变量和“身高”变量都是定比尺度变量,相应的计量结果是定比尺度数据。

现在提出这样一个问题:假若对表1—1所观察的这个从业人员同时还观察了月支出和体重,然后派生出两个新的变量:月收入/月支出、身高/体重,那么这两个变量是何种计量尺度的变量呢?

所谓计量,指的是对个体的直接观察。其中包括:观察它的属性,用点数的方法观察(如被调查的从业人员一个月乘坐公交车的次数),用度量的方法观察等等。可见,计量结果都是原始数据,是直接的观察结果。然而,月收入/月支出、身高/体重的计算结果是两个计量结果相除的派生数据,并不是直接观察结果。所以,严格地说,它们是派生数据而不是原始数据,因而也不涉及计量尺度问题。^①

前面说过,定比尺度数据具有可加性。但是,相除得到的派生数据并无可加性(同一个比式定义下的不同比值相加没有实际意义),这一点一定要注意。

(五) 计量尺度的转化

上述四种计量尺度对事物的计量层次由低级到高级、由粗略到精确逐步递进。高层次的计量尺度可以计量低层次计量尺度能够计量的事物,但不能反过来。

在表1—1中,首先,假若作统计分析的时候不考虑各种不同学历的层次高低,不计顺序,则对“最后学历”变量的计量结果便可以被看做定类尺度数据;其次,我们对“计算机能力测试成绩”变量的所有可能的取值(或是对“月收入”变量的所有可能的取值、对“身高”变量的所有可能的取值)作组距式分组设计,在统计分析的时候只关心各个计量对象所属的组别,于是,这三个变量的计量结果(数据)就变成了“这一个组”或是“那一个组”,它们便既可以被看做定序尺度数据,也可以被看做定类尺度数据;最后,假若在计量“月收入”或是在计量“身高”的时候,把所规定的“0”值仅仅看做标示相应状态的一个符号,那么,这两个变量的计量结果便可以被看做定距尺度数据。

二、数值变量和分类变量

通常把定距尺度变量和定比尺度变量统称为**数值变量**(或称做**定量变量**、**数量变量**),把定类尺度变量和定序尺度变量统称为**分类变量**(或称做**定性变量**、**属性变量**)。

假若对数值变量的所有可能的取值作组距式分组设计,在统计分析的时候只关心各个计量对象所属的组别,于是,计量结果便成为各个计量对象所属的组别。这种计量结果是分类数据(定类数据或定序数据),数值变量此时变成了分类变量。

三、绝对数变量和比率变量

直接观察得到的定比尺度数据叫做**绝对数数据**,相应的变量叫做**绝对数变量**(如月收

^① 在实际工作中,人们习惯上把属于被观察个体的所有数据都看做原始数据,其中既包括直接观察结果,也包括相除派生结果。相应地,也总是希望为相除得到的派生数据规定一个计量尺度名称,在这种情况下,人们会把这种派生数据归入定比尺度数据类型,把相应的变量归为定比尺度变量。

入、月支出、身高、体重);用两个绝对数数据相除算得的比值叫做**比率型数据**,相应的变量叫做**比率变量**(如月收入/月支出、身高/体重)。

四、实验数据和调查数据

实验数据是在科学实验过程中观察得到的数据;**调查数据**(或称做**观察数据**)是对被调查单位已经发生的事件进行回顾性登记得到的数据。

实验数据:例如,为检验脊髓灰质炎疫苗的有效性,安排两组儿童参加实验,其中一组使用疫苗,另一组使用安慰剂,要求两组儿童之间其他与脊髓灰质炎发病有关的条件相同。在使用了疫苗或安慰剂以后,观察每一组的每一名儿童是否发生脊髓灰质炎。这些观察结果便是实验数据。将两组儿童的数据进行对比,可以完成我们所需要的检验。

调查数据:例如,为研究发生肺癌与吸烟是否有关,对某一成年人人群作调查,登记每个人“是否吸烟”、“是否患肺癌”两个方面的状况。调查登记得到的这些数据便是调查数据。根据调查结果,我们可以把每个人分别放入下面四个组中的某一个组。这四个组是:不吸烟且未患肺癌、不吸烟但患了肺癌、吸烟但未患肺癌、吸烟且患了肺癌。用这份整理好的资料可以去研究开始时所提出的问题。

调查数据与实验数据之间有一个重要的区别。在作科学实验的时候,可以根据研究目的,有意识地控制实验条件,因此,**实验数据是在某些重要的实验条件被控制不变的情况下发生的**;相反,调查数据是对已经发生的事件进行回顾性登记的结果,在调查数据发生的时候,**任何条件都未被控制**。

在前一个例子中,我们想要检验疫苗抵御疾病的有效性,在收集被试者脊髓灰质炎发病情况数据的时候,就应当要求两组被试者之间除了使用与未使用疫苗的差别之外,不应当再有其他的差别。满足这一要求并不困难:研究人员可以通过科学地设计实验做到这一点。

在后一个例子中,我们想要研究患肺癌是否与吸烟有关,在收集被调查者肺癌发病情况数据的时候,就应当要求吸烟者与不吸烟者之间除了吸烟与不吸烟的差别之外,不应当再有其他的差别。满足这一要求是有困难的:被调查者中哪些人吸烟、哪些人不吸烟,这是早已形成的既成事实,研究人员无法去安排哪个人吸烟、哪个人不吸烟,于是,在吸烟者与不吸烟者之间还可能会在许多其他与肺癌有关的因素上表现不相同(例如,有可能吸烟者多数是老年人而不吸烟者多数是青年人,有可能吸烟者多数从事有毒作业而不吸烟者多数从事环境较好的工作等等)。为了在处理调查数据的时候满足“吸烟者与不吸烟者之间除了吸烟与不吸烟的差别之外,不应当再有其他的差别”这样的要求,我们可以这样做:把被调查的人群按年龄和工作环境划分开,分别对同样年龄同样工作环境的人研究患肺癌是否与吸烟有关。这种处理调查数据的方法将会在本书以后章节多处讲到。

§ 1.3 社会经济领域中的统计学

统计学是一门数据科学。它指导人们应用自然科学领域、工程技术科学领域、社会经济领域的数

据去研究本领域的问题。
本课程是在普通高等学校经济学、管理学学科开设的一门共同课,它自然应当着重阐述怎样应用统计学的理论方法,通过社会经济数据,去研究社会经济领域中的问题。因此,我们现在所开设的“统计学”课程,说得更具体些,应当把它叫做“社会经济领域中的统计学”。

在统计学的历史发展过程中形成了两个基本分支。第一个分支是在17世纪由“政治算

术学派”的代表人英国人威廉·配第(William Petty, 1623—1687)开创的,当时叫做“政治算术”,后来改称“统计学”。这个分支的主要内容是通过编制适当的统计指标体系来描述社会经济发展状况。为此,各国政府都建立了专门的机构,每年收集社会经济数据,计算并公布本年的统计指标数字。第二个分支是19世纪末形成的以概率论为理论基础的统计学,称之为“数理统计学”,并且也叫做“统计学”。这个分支的主要内容是研究怎样通过数据(样本)去推断不确定性现象的未知特征。

我国统计学界曾广泛使用过“社会经济统计学”这一名称。它属于上述“政治算术统计学”的范畴。20世纪80—90年代国内出版的各种版本的《社会经济统计学原理》、各经济部门的统计学(如《经济统计学》、《工业统计学》、《农业统计学》、《商业统计学》等等)、以介绍社会统计指标体系为主要内容的《社会统计学》^①等等这些教材,反映了“社会经济统计学”的基本内容。

我们现在所说的“社会经济领域中的统计学”,除了包括上面所说的“社会经济统计学”之外,还包括将数理统计学应用到社会经济领域,通过社会经济数据来研究社会经济领域的不确定性问题。

用数理统计学理论方法来处理社会经济数据,研究社会经济领域中的事情,需要首先回答两个基本问题:一是,这样做是否合理,也就是说,能不能做;二是,怎样使用社会经济数据,也就是说,怎样做。

之所以提出第一个问题,是因为数理统计学是在自然科学和工程技术科学的土壤中形成和发展起来的,而这种领域中现象的不确定性,通常是由“自然力”的作用形成的;然而,社会经济领域中现象的不确定性,在许多时候是由人的自主行为造成的。于是自然会问:后一种情形下的不确定性是否与前者具有相同的统计学性质?之所以提出第二个问题,是因为社会经济数据几乎都属于调查数据,前面说过,使用调查数据的时候,须采用必要的统计学手段来控制数据发生的条件。换句话说,正确使用实验数据的关键是在数据发生之前科学地设计实验,而正确使用调查数据的关键则是在数据发生之后科学地处理数据。因此,在使用社会经济数据的时候,要讨论专门针对调查数据的一些特有的统计学问题。

关于上述两个基本问题,将在后面的章节中具体讨论。

本章要点

1. 统计学是一门关于数据的方法论科学。它提供了一套概念、原则和方法,用来指导人们怎样收集数据、怎样分析数据、怎样由数据得出结论。这种指导是在两个基本方向上进行的:一是,怎样依据数据所提供的信息,了解随机现象中我们所关心的各种未知特征;二是,怎样用数据描述客观现象的现实状态。

2. 讨论怎样用样本(子样)推断总体(母体)问题的统计学理论与方法,叫做推断统计学。用样本数据描述样本自身的统计学理论与方法,叫做描述统计学。

3. 来自研究对象的数据叫做样本(或叫做子样),产生数据的那个研究对象叫做总体(或叫做母体)。

4. 描述统计学所描述的对象叫做调查对象总体,或叫做有限总体。在调查对象总体(有限总体)中所计算的统计指标描述了客观现象的现实状态(实际发生的状态)。

5. 对研究对象所观察的量叫做变量,它是研究对象某种特征的名称。一个特定的研究

^① 此外尚有以讲授数理统计方法为主要内容的《社会统计学》,它不属此例。

对象某种特征的具体表现叫做变量值,或叫做数据。数据有原始数据和派生数据之分。观察研究对象所得到的直接结果叫做原始数据。对原始数据进一步加工所得到的数据叫做派生数据。若无特别的说明,所说的数据都是指原始数据。

6. 可以用四种不同的计量尺度取得数据,相应地有四种计量尺度变量(数据):定类尺度变量(数据)、定序尺度变量(数据)、定距尺度变量(数据)、定比尺度变量(数据);把定距尺度变量和定比尺度变量统称为数值变量(或称做定量变量、数量变量),把定类尺度变量和定序尺度变量统称为分类变量(或称做定性变量、属性变量);直接观察得到的定比尺度数据叫做绝对数数据,相应的变量叫做绝对数变量,用两个绝对数数据相除算得的比值叫做比率型数据,相应的变量叫做比率变量。

7. 实验数据是在某些重要的实验条件被控制不变的情况下发生的;相反,调查数据(观察数据)是对已经发生的事件进行回顾性登记的结果,在调查数据发生的时候,任何条件都未被控制。统计学中,常常需要在控制数据发生条件的要求下研究数据,对于后一种情况,控制数据发生条件的办法是把需要加以控制的那些条件纳入研究模型。

思考题和习题

1.1 统计学可以用数据解决什么问题?请仿照例1—1至例1—7另外举7个例子来说明。

1.2 有两位教师按照同样的教学大纲、使用同样的教材分别在两个班讲授《统计学》课程,且期末考试使用同样的试卷。教学管理部门想要依据两个班的及格率来评价、比较两位教师的教学质量。这样做是否有什么不妥当之处?

推荐阅读文献

1. [美] Gudmund R. Iversen、Mary Gergen:《统计学》,吴喜之等译,1~43页,北京,高等教育出版社,2000。

2. [美] David Freedman:《统计学》,魏宗舒等译,3~30页,北京,中国统计出版社,1997。

3. 冯士雍:《对高等院校〈统计学〉教材建设的若干建议》,载《统计研究》,2004(6),52~54页。

4. 周恒彤:《对统计学学科体系建设的感悟》,载《统计研究》,2005(2),3~9页。

5. 周恒彤:《论作为一级学科的统计学》,载《统计研究》,1997(2),8~11页。

6. 周恒彤:《统计学的几个基本理论问题——与聂皖生同志商榷》,载《统计研究》,1995(2),44~47页。

第 2 章

随机现象以及对它的概率描述

§ 2.1 关于随机现象的若干概念

一、随机现象的种类

随机现象的种类如图 2—1 所示。

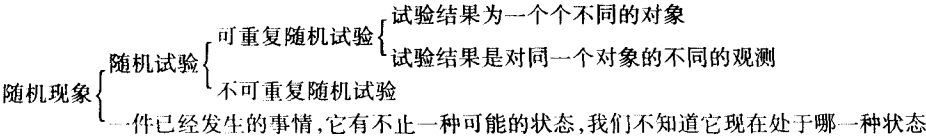


图 2—1 随机现象的种类

在第 1 章的几个例子中, 例 1—1、例 1—2、例 1—3 和例 1—4 是随机试验, 而例 1—5 是一件已经发生的事情。

随机试验的共同特点是: 事情尚未发生, 我们想要预见事情将会发生何种结果。然而, 事情将会发生的结果具有多样性并呈现不确定性, 因而在进行预见时就必须把结局的多样性和不确定性考虑在内。

与随机试验不同, 在例 1—5 中, 事情已经发生, 我们想要猜测事情所发生的是哪一种状态。然而, 事情所发生的状态具有多样性并呈现不确定性, 因而在进行猜测时就必须把事情所处状态的多样性和不确定性考虑在内。

在 4 个关于随机试验的例子中, 例 1—1、例 1—2、例 1—3 可以在一组不变的条件下无限地重复进行同一个随机试验, 称之为**可重复随机试验**; 例 1—4 无法在不变的条件下重复进行同一个随机试验, 称之为**不可重复随机试验**。另外, 我们又注意到, 在 3 个可重复随机试验的例子中, 例 1—2 和例 1—3 的试验结果是一个个不同的对象, 而例 1—1 的试验结果是对同一个对象的不同观测。

有的教科书中仅仅把随机试验所观察的现象叫做随机现象, 并相应地把随机试验表述为“对随机现象所进行的观察”, 不妨把它叫做对“随机现象”的狭义解释。在这里, 我们把“一件已经发生的事情, 它有不只一种可能的状态, 我们不知道它现在处于哪一种状态”这类情形也包括在随机现象之中, 不妨把它叫做对“随机现象”的广义解释。

事实上, 如果人们不知道“一件已经发生的事情”处于何种状态, 会在不止一种可能的状态之间进行猜测, 那么, 当这件事情尚未发生, 人们想要预见它的结果时, 那就是一个随机