

第一章

统计学基础

学习目标

了解统计学的产生、发展及其应用；
理解统计的含义、过程和职能；
掌握统计学的研究对象和研究方法；
掌握统计学的基本名词。

引例

生活中的统计学

在大数据时代,人们的工作和日常生活都离不开各种各样的数据。例如,人们早上起来会关心当天气温的高低和空气质量情况,球类比赛的解说员总要统计竞赛双方的进攻次数和成功率,学生考试后非常关心自己的考试成绩和名次,企业管理人员要掌握生产销售情况和利润额,报刊和电视中常会提到国内生产总值、消费者价格指数和经济增长率,等等。

很多测评机构在比较各个高校的实力或比较不同高校培养的大学生受社会欢迎的程度时,不仅会用到就业率等指标,还会用到大学毕业生的工资水平。要反映一个学校所培养学生的工资水平显然不宜用该校全部学生工资的最高水平或最低水平,否则容易受个别极端值(统计上也称为异常值)的影响。通常,测评机构会采用该校全部学生工资的平均水平、工资达到一定水平的毕业生所占比重等统计指标来衡量。但要对毕业生进行全面调查既不太可能也没必要。为此,测评机构可进行抽样推断,即随机抽取一定数量的大学毕业生构成样本进行推断统计。

企业要想知道新产品与原来的产品相比是否有很大的改善,就需要收集大量的数据对新产品进行检验。

医疗部门要对某种疾病进行调查,需要对发病环境的饮用水、土壤、空气质量、生活习惯等因素进行分析,分析这些因素和该种疾病是否相关,相关程度有多大,病因是由一个因素引起的还是由多个因素引起的,等等。

人们要想正确理解这些现象及其原因,就需要具备一定的统计学知识。

第一节 统计学及其应用

一、统计的含义

统计是处理数据的一门科学。从发展过程来看,统计有着丰富的内涵和外延。通常,“统计”一词有三方面的含义,即统计活动、统计资料和统计学。

(一) 统计活动

统计活动又称统计工作,是指收集、整理和分析统计数据,并探索数据的内在数量规律性的活动过程。

统计工作在人类历史上出现得比较早。早在原始社会时期,我国就有结绳记事、结绳计量的方法,不仅产生了简单的总量指标,还应用了简单的统计分组。这是统计的萌芽阶段。在夏朝时期,我国就有了人口、土地等方面的统计。当时为了治国治水的需要,人们进行初步的国势调查,将全国分为九州,并按土质的优劣将九州的田、赋分成上、中、下三等,又将每等分为上、中、下三级,由此形成“三等九级”的复合分组。西周时期,我国已设有专门负责国势调查的官员——职方氏。春秋时期的管仲提出了系统、周密的国情调查提纲与国情研究纲要,指出“举事必成,不知计数不可”“不明于计数,而欲举大事,犹无舟楫而欲经于水险也”,要求“明法审数”。战国时期的商鞅提出:“强国知十三数:竟内仓、口之数,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、刍蒿藁之数。”在秦汉时期,我国就有了地方田亩和户口的记录。在唐宋时期,我国有了统计资料的汇编——《国计簿》《会计录》,同时在统计图表的理论和方法方面也有了较大的发展。自宋朝开始,我国逐渐建立经常人口登记和保甲制度,特别是明初的户贴制度被称为“世界上最早实行全面人口普查的历史记录”。

在西方,古埃及为建造金字塔和农业灌溉系统进行了全国的人口及财产调查;而古罗马帝国将国势调查作为治理国家的有效手段,规定每5年进行一次人口、土地、牲畜、家奴的调查。到了封建社会,国家往往采取编制财产目录的形式进行统计调查,后来统计调查成为表明各国国情的工具。

(二) 统计资料

统计资料是统计活动进行收集、整理、分析,能反映事物现象的数据资料和其他相关资料的总称,是人类活动各数量方面客观情况的记录。它包括数据资料和文字资料,以数据资料为主。统计数字、统计分析报告、统计台账、统计表、统计图等都是数据资料。当今是大量

据时代,人们的工作、学习和日常生活都离不开各种各样的数据。例如,人们外出旅游要利用天气预报和电子地图关注目的地的天气情况和路线情况,还要利用相应的网站查询住宿情况;企业要进行采购、生产、销售和利润核算,就需要对相关的采购资料、生产资料、销售资料等进行收集、整理、分析和研究;国家要进行经济建设和促进社会发展,就要对各个方面都进行普查,以了解国情和民情。

(三) 统计学

统计学是一门关于如何收集、整理和分析数据并进行推断的方法论科学。其目的是探索数据内在的数量规律性,以科学认识客观事物。统计学以现象总体的数量方面为研究对象,用以阐明统计设计、统计调查、统计整理和统计分析的理论与方法。

《不列颠百科全书》将统计学定义为收集、分析、表述和解释数据的科学。

统计活动、统计资料和统计学之间有着密切联系。统计活动与统计资料之间是过程与成果的关系,统计资料是统计活动的直接成果。就统计活动和统计学的关系来说,统计活动属于实践的范畴,统计学属于理论的范畴。统计学是统计活动实践的理论概括和科学总结。它来源于统计实践,又高于统计实践,反过来又指导统计实践。统计活动的现代化与统计学研究的支持是分不开的。离开了统计数据,统计学就失去了存在的意义;而没有使用统计方法整理的数据也是没有顺序的无用数据。

统计活动、统计资料和统计学相互依存、相互联系,共同构成一个完整的整体,这就是人们所说的统计。

二、统计学的产生和发展

进入资本主义社会以后,随着社会生产力的发展,人们对统计数据资料的需求增多,专业的统计机构和研究组织逐渐出现,统计初步发展为社会分工中的一个独立部门。到了17世纪中叶,统计学应运而生。

(一) 古典统计学(17世纪中叶至18世纪中叶)

17世纪中叶至18世纪中叶是统计学的创立时期。在这一时期,统计学理论界初步形成了一定的学术派别,主要有国势学派、政治算术学派和古典概率论。

1. 国势学派

国势学派产生于17世纪的德国,其代表人物是康令、阿亨瓦尔等。康令将国势学从国法学、地理学、历史学混合在一起的杂学中分离出来,并从1660年11月20日起,在不伦瑞克城西尔姆斯特大学以“国势学”为题,讲授政治活动家应具备的知识,主要是记述和比较各国的领土、人口、财政、军事、政治和法律等方面的情况,在研究目的、研究对象和研究方法方面初步形成了社会经济统计的体系。阿亨瓦尔是这一学派最有名的人物。1749年他的《近代欧洲各国国势学纲要》一书出版。在该书的序言中作者使用了“统计学”一词。他收集了大量的实际资料,分门别类地记述了有关国情、国力的系统知识,进一步发展了国势学的知识体系,使其逐步形成了国势学派。国势学派的特点是用文字的方法来记述社会经济现象和用哲学的观点来论证社会经济现象。

国势派对统计学的创立和发展的主要贡献如下:为统计学这门学科起了至今仍为世界公认的名称“统计学”,并提出了仍为统计学者所采用的“统计数字资料”“数字对比”等一

些术语;认为“对比乃统计方法之母”,即只有在对比中才能更清楚地看出事物的规模、结构及其发展变动等。但是,该学派多用文字上的比较级和最高级对各国的社会经济情况进行对比分析,始终没有把数量对比分析作为这门科学的基本特征,没有以研究总体的数量特征为己任,主要使用文字来记述国家的显著事项。正因为如此,国势学派被认为“有统计学之名,而无统计学之实”。

后来,在政治算术学派的影响下,表式学派从国势学派中分化出来,改用列表的方法研究社会经济现象的数量方面,开始体现统计学的实质特点。

2. 政治算术学派

政治算术学派产生于17世纪中叶的英国,其代表人物是威廉·配第和约翰·格朗特。该学派主张以数字、重量和尺度来说话,用图表形式来概括数字资料。

1662年,英国学者约翰·格朗特发表了论文《关于死亡率的自然观察和政治观察》。他收集整理了自1603年以来公布的死亡表,对其加以观察比较,论证了出生、死亡、死因、男女性别比率等人口动态存在一定的规律,并使用新颖的方法编制出世界上第一张生命表。所以,他被公认为人口统计学的创始人。

威廉·配第的代表作是《政治算术》。在该书中,他独创地应用算术方法来研究经济问题,运用大量的数字资料对比英国、法国、荷兰的经济力量,为英国的未来发展指明了方向。他在书中所阐明的统计学的研究目的、研究任务和研究对象为社会经济统计学的形成奠定了基础。他所创造的统计方法如计量方法、图表方法、分组方法、推算方法等都成为社会经济统计学的基本原理。可以说,威廉·配第是政府统计的创始人。因此,马克思认为:“威廉·配第是政治经济学之父,在某种程度上也可以说是统计学的创始人。”但是,政治算术学派并没有使用统计学的名字,被认为“有统计学之实,而无统计学之名”。

综上所述,政治算术学派的出现标志着古典统计学的诞生。政治算术学派以数量分析为特征,研究客观现象的数量关系。就其内容和方法来看,应该算是统计学的真正渊源。不过,无论是国势学派还是政治算术学派,他们的著作和内容还属于实质性的社会科学。

3. 古典概率论

古典概率论的研究虽始于16世纪的意大利,但到17世纪中叶才得到重视和研究,并在18世纪的法国、瑞士等国家得到广泛发展。19世纪初,法国数学家、统计学家拉普拉斯在总结前人成果的基础上出版了《概率的分析理论》一书,从而形成完整的应用理论体系。古典概率论对统计学的贡献可归纳为以下几点:

- (1) 总结了古典概率论的研究成果,初步奠定了数理统计学的理论基础。
- (2) 把大数定律作为概率论与政治算术的桥梁。
- (3) 提出应以自然科学的方法研究社会现象,为数理统计的产生提供了必要的理论依据。

(二) 近代统计学(18世纪末至19世纪末)

18世纪末至19世纪末是统计学的发展时期。在这一时期,各种学派的学术观点已经形成,并且形成两个主要学派,即数理统计学派和社会统计学派。这两个学派既是关于统计学原理的两个相对独立的学派,也是关于统计学的学科性质相互争论的学派。近代统计学也可以称为现代统计学的奠基时期,因为其内容体系和发展线索是现代统计学的基础。

1. 数理统计学派

18 世纪末,概率论的发展和数学方法在社会经济统计方面的运用促进了概率论和政治算术的结合,从而形成数理统计学派。

19 世纪中叶,比利时物理学家、数学家、天文学家和统计学家阿道夫·凯特勒在统计学发展中的最大贡献是将概率论引入了统计学,从而使统计学发生了质的飞跃。凯特勒的研究成果在自然科学、经济学、生物学等科学中得到不断应用,并逐渐形成一门独立的学科。其代表作是《社会物理学》。他认为概率论是适用于政治及道德科学中以观察与计数为基础的方法。他以此方法对自然现象和社会现象的规律性进行观察,并认为要促进科学的发展,就必须更多地应用数学。他的统计学著作有几十部之多,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。同时,他还是第一届国际统计会议的召集人。因此,凯特勒被称为数理统计学的奠基人、近代统计学之父。

后来经过英国的高尔登和皮尔森等人的发展,统计学研究的数理统计学派逐渐形成。1867 年,德国数学家维特斯坦发表了《关于数理统计学及其在政治经济学和保险学中的应用》一文,首次提出了“数理统计学”的说法。

2. 社会统计学派

19 世纪后半叶,在德国兴起了社会统计学派,由克尼斯首创,主要代表人物为恩格尔和梅尔。他们以大量观察法寻求社会现象规律,因此被称为社会统计学派。强调研究对象的总体性和采用大量观察法构成社会统计学派研究方式的两大特点,这也是现代统计研究的两大特点。

英国人皮尔森提出了经验分布函数、相关分析、动差法、卡方检验和大样本的抽样理论,并且构建了描述统计学的体系。因此,有人认为,皮尔森是近代统计学的创始人。

(三) 现代统计学(20 世纪初至今)

20 世纪初以来,随着科学技术的发展,社会发生了巨大变化,统计学也进入了快速发展时期,归纳起来有以下几个方面:

(1) 数理统计学在随机抽样的基础上建立起推断统计的理论和方法。英国统计学家威廉·戈塞被认为是推断统计学的奠基人。他提出了用于小样本理论的 t 分布。该分布的形式是由罗纳德·艾尔默·费希尔发展起来的。戈塞和费希尔曾相互通信并讨论问题。在费希尔、奈曼等人的推动下,假设检验理论得以逐渐建立起来。美国统计学家瓦尔德对统计学中的估计和假设检验理论加以归纳,提出了统计决策、序贯分析等理论。同时,费希尔还是实验设计统计学的开创者。

(2) 社会经济统计学在 19 世纪中叶以后也逐渐成熟和发展起来。“十月革命”胜利后,苏联的大多数统计学家受社会统计学派的影响,主张统计学是一门实质性的社会科学。1954 年 3 月,由苏联科学院、中央统计局、高等教育部联合召开了统计科学讨论会,并把统计学定义为“在质与量的密切联系中研究大量社会现象的数量方面,研究社会发展规律在具体地点及时间条件下的数量表现的社会科学”。这一定义对我国及东欧的社会主义国家的影响都很大,促使这些国家形成了以马克思政治经济学为理论基础的社会经济统计学派。该学派是在反虚无论、反消亡论、反万能科学论和反数学形式主义中形成的,并以物质产品为核算范围,建立了物质产品平衡表核算体系。直到 1993 年,该核算体系才逐渐被改变,但该学派的影响至今仍很大。这方面的现代统计工作包括社会调查、经济计量学和投入产出核

算,以及更完备的国民经济核算体系。

(四) 统计学的发展趋势

从古典统计学到现代统计学,统计学几经演变,已有 300 多年的历史。自 20 世纪 50 年代以来,统计理论、方法和应用进入了一个全面发展的新阶段。一方面,统计受计算机科学、信息论、人工智能等现代科学技术的影响,新的研究领域层出不穷。特别是近 20 年来,海量数据的存储,互联网、云计算等科技的发展对统计学科的发展更有着重大影响。另一方面,统计方法的应用领域不断扩展,统计学与其他新科学、新理论的结合,不断产生新的边缘科学或新的统计分支。现在就连法律、历史、语言、新闻等学科都越来越重视对统计数据进行分析,国外的人文与社会学科也普遍开设统计学的课程。由此可见,统计方法与数学、哲学一样成为所有学科的基础。

(五) 统计学的分科

统计的内容非常丰富,其方法应用于自然科学和社会科学的众多领域。统计学有两种基本的分类:从统计方法的功能来看,统计学可分为描述统计学和推断统计学;从统计方法研究的重点来看,统计学可分为理论统计学和应用统计学。

1. 描述统计学和推断统计学

描述统计学研究如何取得反映客观现象的数据,并通过图表形式对所收集的数据进行加工处理和显示,进而通过综合、概括与分析得出反映客观现象的规律性数量特征。描述统计学的内容包括统计数据的收集方法、数据的加工处理方法、数据的显示方法、数据分布特征的概括与分析方法等。

推断统计学研究如何根据样本数据去推断总体数量特征。它是在对样本数据进行描述的基础上对统计总体的未知数量特征做出以概率形式表述的推断。

描述统计学与推断统计学的划分,一方面反映了统计方法发展的前、后两个阶段,另一方面反映了统计方法探索客观事物数量规律性的先、后两个过程。如图 1-1 所示,统计研究过程的起点是统计数据,终点是探索出客观现象内在的数量规律性。在这一过程中,如果收集到的是总体数据(如普查数据),那么运用描述统计学就可以达到认识总体数量规律性的目的;如果获得的只是研究总体的一部分数据(样本数据),那么要找到总体的数量规律性,就要运用概率论并根据样本信息对总体进行科学的推断。显然,描述统计和推断统计是统计方法的两个组成部分。描述统计是整个统计学的第一步,包括数据的收集、整理和显示,分析和提取数据中的有用信息以最终推断总体。推断统计则是现代统计学的主要内容,在现代统计学中的地位和作用越来越重要,已成为统计学的核心内容。这是因为在对现实问题的研究中,所获得的数据主要是样本数据,选用何种统计量、哪种推断方法及如何进行推断是探索事物内在规律的关键。从描述统计学发展到推断统计学,既是统计学发展的巨大成就,也是统计学发展成熟的重要标志。

2. 理论统计学和应用统计学

理论统计学是指统计学的数学原理,即数理统计学,主要探讨统计学的数学原理和统计公式的来源。由于现代统计学几乎用到了所有方面的数学知识,因而从事统计理论和方法研究的人员需要有深厚的数学基础和经过严格的数学训练。同时,由于概率论是统计推断的数学基础和理论基础,因此,广义的统计学也应包括概率论。理论统计学包括的主要内容

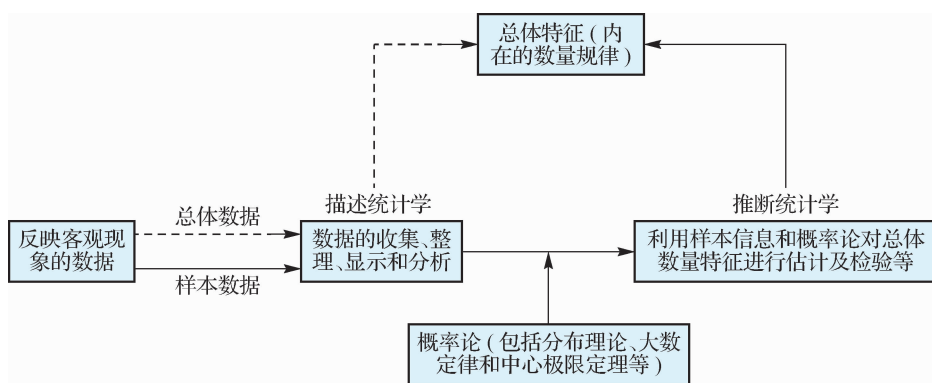


图 1-1 描述统计学和推断统计学在统计分析过程中的关系

有概率理论、抽样理论、实验设计、估计理论、假设检验理论、决策理论、非参数统计、贝叶斯统计、时间序列分析和随机过程等。理论统计学是统计方法的理论基础。

应用统计学探讨如何运用统计方法解决实际问题。将理论统计学的原理应用于各个学科领域就形成了各种各样的应用统计学。例如,统计方法在物理研究中的应用可形成统计物理,统计方法在生物学中的应用可形成生物统计学,统计方法在医学中的应用可形成医疗卫生统计学,统计方法在农业试验、育种等方面的应用可形成农业统计学,统计方法在风险管理及保险中的应用可形成保险精算学,统计方法在管理领域的应用可形成管理统计学,统计方法在人口学中的应用可形成人口统计学,等等。以上这些应用统计学的不同分支所应用的基本方法都是描述统计学和应用统计学的主要方法,但由于各个领域都有其特殊性,因此统计方法在实际应用中具有不同的特点。例如,正态分布在教育测量和分数转化问题的研究中得到了发展。

三、统计学的应用

统计学的应用十分广泛。首先,统计学作为应用性很强的科学,其生命力和发展动力在于其与实用学科的密切联系。如果割断这种联系,统计学就会变成无源之水、无本之木,产生不出有意义的问题和方法。因此,统计学与其他学科和领域所形成的边缘和交叉性质的学科特别多,如工业统计学、农业统计学、生物统计学、医药统计学、人口统计学、空间统计学等。其次,统计学所研究的处理数据的方法在当今的信息化时代,无论是在自然科学领域、社会科学领域,还是在工程技术领域都发挥着重要的作用。如果没有统计方法,大量杂乱无章的数据就无法显示其内含的有用信息。

(一) 统计学在生活领域的应用

统计从收集资料到分析、处理资料,再到将其转换为有用的信息,成为知识经济时代最重要的生产要素,从理论上讲是一种方法,从应用上讲是一种技术。统计技术是推动社会进步的先进生产力。

统计的力量在于其无处不在的应用性。无论人们从事什么工作,都有可能遇到下列问题:如何收集有价值的资料,如何组织、解释所收集的资料,如何分析并给出适当的推论,等等。这些都是统计学应用的主要范畴。在生活领域中,诸如市场调查、工业产品质量控制、

农产品品质的改良试验、医学的临床试验等都是必须运用统计学理论的常见例子。

（二）统计学在经济领域的应用

统计学最初产生于对经济现象的研究。至今,经济领域仍然是统计学最重要的研究领域。统计学在经济领域的应用形成了经济统计学。经济学在研究经济现象及其发展变化的规律性时,除要进行规范性的理论分析外,还离不开对现实经济活动的实证研究。经济学家只有通过对现实经济活动的运行条件、运行过程和运行结果的数量分析,才能得出真正符合客观实际的规律性结论。经济现象是人类参与的活动,其影响因素异常复杂。对社会经济现象规律性的认识,不可能与研究物理、化学现象一样,在严格控制的条件下到实验室中去反复实验,而只能被动地对实际的经济关系和经济活动的运行情况进行观测。在这一点上,经济学家面对的问题类似于气象学家和地震学家面对的问题。因此,无论是宏观经济学研究还是微观经济学分析,都需要大量地运用统计方法,通过各种调查方法来收集实际的经济统计数据,并分析其数量规律性。统计学是经济学研究的一种基本手段。

众所周知,量变引起质变。数量关系的背后牵扯着市场的稳定与发展。金融业的现代化推动了统计与数理方法的应用研究。反过来,当今世界的金融管理特别是防范金融风险也越来越需要量化研究。早在1995年9月,美国斯坦福大学经济学教授刘遵义就通过实证比较、数量分析和模糊评价等方法预测出菲律宾、韩国、泰国、印尼和马来西亚有可能发生金融危机。后来,事实果然如此。这从一个侧面提醒人们,没有完整、科学的分析预测工具,就可能在国际金融竞争中蒙受重大损失。因此,只有加强对作为金融信息的各种变量进行研究,才能提高对金融运行规律的认识,才能把握市场发展的动向。

投资顾问使用各种统计数据进行分析,并用于指导他们的投资建议。对于股票投资来说,投资顾问通过审查各种财务数据如市盈率、股息率等,将个股信息与股票市场加以比较,就可以得出个股价格是被高估还是被低估的结论。例如,在2005年1月16日的上证指数中,30只股票的平均市盈率为21.08,东方电子集团有限公司的市盈率为17.92。此时市盈率方面的统计信息显示:与上证指数股票的平均收入相比,东方电子集团有限公司的股票价格较低。因此,投资顾问可以得出结论:东方电子集团有限公司的现行价格被低估了。这一结论及其他有关东方电子集团有限公司的信息有助于投资顾问给出买入、卖出或持股的建议。

在价格分析中,经济学家要测量和探索物价变动的数量规律性,从而在比率和平均数的基础上产生了物价指数法。

金融领域需要更先进的统计技术,如证券市场投资及风险的分析、外汇市场投资及风险分析、保险经营中的精算模型和计算分析等。当今,经济统计面对的是越来越多的数据和解决现实问题的复杂方法,因而统计测度与系统分析同等重要。

（三）统计学在管理领域的应用

在经济全球化的现代市场经济条件下,准确的统计资料与统计方法的正确运用是政府实现有效管理和国家实现现代化的关键。要对国民经济这样庞大的系统进行有效的管理和控制,就必须全面掌握社会再生产的条件、过程和结果的数量信息,对国民经济系统进行全面的统计研究。为此,从世界各国政府到联合国都建立了相当规模的统计机构。为了从总量上反映社会再生产的条件、过程和成果,世界各国还建立了国民经济核算体系,并将其作

为宏观经济统计的基本框架。为满足宏观经济管理的需要建立的统计体系的内容范围十分广泛。其中,国内生产总值统计、投入产出统计、资金流量统计、资产负债统计和国际收支统计是最核心的部分。此外,对社会资源、经济增长、经济结构、经济平衡、经济效益的统计分析,对价格指数、通货膨胀的统计分析,对综合国力的国际对比,对国民经济的宏观监测与预警分析都是宏观经济统计的重要内容。为了满足国民经济各个部门和各个专业管理的需要,各国还建立了部门的或专业的统计体系,如农业统计、房地产统计、劳动统计、货币银行统计等。

统计在企业管理中也被广泛应用。企业管理是指企业对生产经营活动进行策划、组织、营销、激励、调节、控制,以获得最大的经济效益。企业的生产经营活动离不开统计。例如,企业运用统计方法开展市场调查、市场预测和 market 分析,以调整企业经营策略,对产品质量进行监测与控制,对企业的经济效益进行评价。因此,统计学应是企业经营管理者们的必备知识。

需要特别指出的是,很多管理活动都会运用统计工具。相关人员将这一有用的工具与自身专业领域相结合,能够提高管理的效率。比如,会计是一项专业性很强的管理活动,会计师事务所对其客户进行审计时常使用统计抽样的方法。假设一家会计师事务所想确定客户的资产负债表中所显示的应收账款的数量是否公正地体现了实际应收账款的数量。通常,应收账款的账目数量很大,要审查和验证每个账户既耗时又费钱。在这种情况下,审计人员通常的做法是选择应收账款的一个子集,即样本。在审查过抽样账款的正确性后,审计人员就可以得出客户的资产负债表中所显示的应收账款的数量是否可以接受的结论。又如,零售付账柜台使用电子扫描仪收集数据。市场调查公司或营销专家需要的各种数据都可以从零售商店销售点的购买扫描记录中获得。市场调查公司或营销专家会对对这些数据进行处理后的统计汇总信息提供给制造商。产品公司经理们可以通过检查并评论扫描得到的数据及促销活动的相关数字更好地理解销售与促销活动之间的关系。这些分析可为未来制定营销策略提供依据。

21 世纪是质量的世纪,各行各业都面临着各种质量问题。1924 年,美国的休哈特提出了统计质量控制理论,阐述了控制和预防缺陷的概念,成功地创造了“控制图”,把数量统计方法引入质量控制,大大推进了质量管理水平。休哈特被称为“统计质量控制之父”。质量大师戴明和朱兰在 20 世纪四五十年代发展了这些原理,并在实践中对其进行了证明。

(四) 统计学在其他领域的应用

随着计算机的发展,统计在各行各业中的应用更广,在人文、社会科学方面也有应用。1987 年,复旦大学数学系副教授李贤平的工作引人注目。他在美国威斯康星大学的计算机前工作了数百小时,绘制了 300 多张图纸,运用计算机技术中的模式识别法和统计学家使用的探索性数据分析法对《红楼梦》进行了统计分析与风格分析。他翻阅了大量的红学研究论文和资料,利用过去红学家们发掘的资料进行考证。他将《红楼梦》的一百二十回作为一个整体,以 47 个虚字为识别特征,对它们在书中各回的出现频率进行统计分析,并使用频率绘成图纸,根据图纸反映出的表明不同创作风格的星云状和阶梯状的图形提出了又一次震惊红学界的新观点:《红楼梦》各回的写作风格具有不同的类别,各回实际上是由不同作者在不同时期完成的。李贤平认为,《红楼梦》前八十回是曹雪芹根据《石头记》增删而成的,其中插入了他早年著的《风月宝鉴》,并增写了具有深刻内涵的许多内容。《红楼梦》后四十回是曹

家亲友在曹雪芹全书尚未完成就突然去世之后,收集整理原稿并加工补写而成的。程伟元将全稿以活字版印刷刊行。高鹗校勘异文补遗订讹。他的这一看法否定了被红学界一直视为“曹雪芹作前八十回,高鹗续后四十回”的定论,使这一传统观点受到了严峻挑战。有的红学家认为这是一个新突破、新开拓,有的则认为还应做进一步的探讨。这一情况表明,利用计算机等现代科学技术来研究《红楼梦》这样伟大的作品,确实别有一番天地,拓展了视野,也产生了值得重视的成果。

我国与市场经济发达国家的差别非常大,在法律事务中应用统计学几乎是空白。因此,统计在法律事务中具有广阔的应用空间。以美国为例,当案件所面临的问题变得越来越复杂时,统计就显得越来越重要了。统计学家分析数据,帮助法官或陪审团决定某人是否有罪,或是否必须对所造成的伤害进行赔偿。律师需要聘请统计学家或具有统计专业知识的人进行咨询,或对大量事实进行验证。这样的统计学家不但要懂统计,还要能够向不熟悉统计的法官、陪审团和律师有效地解释自己的工作。法庭需要运用各种统计方法处理案件。因此,法庭统计是极具挑战性的一个应用统计的重要分支。

第二节 统计学的研究对象和研究方法

一、统计学的研究对象

统计学的研究对象是现象总体的数量方面,即现象总体的数量特征和数量关系。

统计学研究的目的在于认识现象总体的规律性。由于组成总体的个体受偶然因素的影响,在数量上的表现千差万别,因此分析人员只有通过大量观察与收集充分的资料,才能真正地认识和把握事物的本质特征。

人们在研究现象时需要事物的总体进行研究。构成总体的各单位除在某一方面必须是同质的外,在其他方面要有差异。只有对总体足够多的单位进行研究和分析,才能消除个别单位偶然性或特殊性的不正常影响。统计研究的不是抽象的量而是一种社会调查活动,具有明确的现实意义,是非纯数量的研究。抽象的量能使人们认识现象的质,能使人们注意决定物质的数量界限,还能表现现象在一定历史条件下发展变化的规律性。因此,数量的多少、现象之间的数量关系与质量互变的数量界限都是人们要研究的内容。

二、统计学的基本研究方法

(一) 大量观察法

大量观察法是统计学特有的方法,常用于统计调查阶段。大量观察法是指在对事物了解的基础上,对总体的全部或足够多的单位进行统计观察和登记并掌握与问题有关的全部事实的方法。社会现象和自然现象都受各种社会规律或自然规律相互交错的影响。在现象总体中,个别单位往往受偶然因素的影响。如果任选其中之一进行观察,其结果不足以代表总体的一般特征。只有观察全部或足够多的单位并加以综合,影响个别单位的偶然因素才会相互抵消,现象的一般特征才能显示出来。大量观察的意义在于使个体与总体之间在数量上的偏差相互抵消,以显示现象的必然状态和本质数量特征。

（二）统计分组法

统计分组法用于统计整理阶段,是指根据一定的研究目的和现象的总体特征,将总体各单位按一定的标志分组,把社会经济现象划分为不同性质或类型的组别,再利用统计表、统计图等描述研究总体现象的数量特征的方法。

（三）统计描述法

统计描述法又称综合分析法,是指把大量观察得到的资料加工成各种综合指标(如绝对数、相对数、平均数等),以反映总体的一般数量特征,再利用对比分析法、动态趋势分析法、因素分析法、相关与回归分析法等对综合指标进行分解和对比分析,以研究总体的差异和数量关系。

（四）统计推断法

统计推断法是指在一定的置信标准要求下,由样本统计量来判断总体参数的归纳推理方法。统计推断法包括参数估计、假设检验、方差分析等。统计推断是现代统计学的基本方法,在统计研究中得到了极为广泛的应用。从这种意义上说,统计学是在不确定性条件下做出决策或推断的一种方法。

（五）统计模型分析法

统计模型分析法是根据统计资料,运用统计方法,对研究现象的结构或过程建立一种统计表达式,并进行有关分析的一种研究方法。例如,相关分析法、回归分析法和统计预测法都是统计模型分析法。其实质是综合分析法的一种。

以上介绍的五种统计学的基本研究方法并不是统计学研究的所有方法。分析人员在运用统计学的研究方法时还应注意将多种方法结合起来。在调查方法上,分析人员要注意将大量观察和典型调查结合起来。

第三节 统计的过程和职能

一、统计的过程

统计作为对社会经济现象总体数量特征的一种直接认识活动,是一个不断深化、循环往复的过程。但是,就某一项统计工作而言,它一般包括统计设计、统计数据的收集、统计数据的整理和统计数据的分析,如图 1-2 所示。



图 1-2 统计的过程

（一）统计设计

统计设计是整个统计研究的前期工程。其完成质量直接关系到整个统计研究的质量。统计设计是根据统计任务与统计对象本身的特点和统计研究的目的,对统计工作的各个方

面和各个环节进行的通盘考虑和全面安排;统计设计是根据所要研究问题的性质,在有关学科理论的指导下,制定统计指标、指标体系和进行统计分类,并给出统一的定义、标准,同时提出收集、整理和分析数据的方案和工作进度,以及经费的运用等。搞好统计设计不仅要以统计学的一般理论和方法为指导,还要对所要研究的问题具备深刻的认识和相关的学科知识。

(二) 统计数据的收集

统计数据的收集有两种基本方法。大多数自然科学和工程技术研究有可能通过有控制的科学实验去取得数据,这时可以采用实验法。在统计学中有专门一个分支——实验设计,就是研究如何科学地设计实验方案,从而使通过实验采集的数据符合分析的目的和要求。而对于社会经济现象来说,一般无法进行重复实验,研究人员要取得有关数据就必须到社会总体中去选取足够多的单位进行调查观察,并加以综合研究。如何科学地进行调查是统计学研究的重要内容。

(三) 统计数据的整理

统计数据的整理是根据统计研究的目的,将调查所获取的大量统计资料进行分类和汇总,为统计分析准备系统综合资料的工作过程。通过统计调查收集的统计资料是分散的、不系统的,只能说明事物的表象、事物的某一侧面或外部联系。分析人员需要将其条理化、系统化,即需要对其进行整理。

(四) 统计数据的分析

统计数据的分析是根据统计研究的目的,综合运用各种分析方法和统计指标对取得的数字资料 and 具体情况进行定量和定性分析,以揭示现象发展变化的过程及规律性。这是统计工作的决定性环节,是关系到能否充分发挥统计职能的一项重要工作。常用的分析方法有综合指标法和归纳推断法等。

二、统计的职能

统计具有信息、咨询和监督三大职能。

(一) 信息职能

信息职能是指用科学的方法收集经济、科技、社会等各方面信息,并向全社会提供各方面所需信息的功效和作用。信息职能是统计的主体职能和基础职能。其他职能是在此职能的基础上衍生出来的,并且直接受其质量的影响。

在社会经济活动中,凡经统计调查、统计整理和统计分析能反映客观事物总体数量特征的信息都是统计信息。统计信息是社会经济信息的主体。统计部门会把有关的统计信息提供给决策部门和公众。统计信息与其他信息相比,具有数量性和整体性的特征。统计信息的数量性表现在,统计是从数量方面认识和反映事物现象的。

(二) 咨询职能

咨询职能是指各级统计机构利用统计部门已经掌握的丰富的统计信息资源,运用科学的分析方法和先进的技术手段,对国民经济和社会、科技发展情况深入开展综合分析和专题研究,为管理和科学决策提供各种可供选择的咨询建议与对策方案。

统计是一项服务性很强的工作。统计的咨询职能最能体现社会经济统计工作的方向和性质。我国自改革开放以来,统计部门逐步成为重要咨询机构之一。统计信息咨询除为党政领导加强宏观调控提供具有量化特点的咨询意见和对策建议外,还必须面向研究机构、社会团体、基层企业、个人和国外多类用户大力发展统计信息咨询服务,建立公共统计信息咨询服务机构,培育统计信息市场。只有真正做到服务形式多样,将服务落到实处,统计才能充分发挥作用。

(三) 监督职能

监督职能是指通过统计调查和统计分析及时、准确地从总体上反映经济、社会、科技运行状态,并对其实行全面、系统的定量检查、监督和预警,以促进国民经济按客观规律的要求持续、稳定、协调地发展。统计的监督职能是把统计部门建设成为多功能智力型机构的战略目标的需要。

统计的监督职能是社会经济统计的固有职能,也是社会经济统计的一大特色。在社会经济的监督体系中,统计监督是一项政策性强、不可或缺的工作。

统计的信息、咨询和监督三大职能是相互作用、相辅相成的,共同构成一个有机整体。信息职能是保证咨询职能和监督职能有效发挥的基础;咨询职能是信息职能的深化;监督职能是对信息职能和咨询职能的进一步拓展,并促进信息职能和咨询职能的优化。

第四节 统计学的基本名词

一、总体与样本

(一) 总体

总体是统计总体的简称,是指客观存在的具有某种共同性质的许多个别事物构成的整体。当把某一客观现象作为统计研究对象时,所要研究现象的整体就称为总体。例如,我国要进行人口普查,我国的全部人口就组成一个总体;我国要研究某地区独立核算工业企业的情况,该地区全部独立核算工业企业就是一个总体。

统计的总体既可以由人组成,也可以由物组成。总体根据它所包含的总体单位的数目是否可数,分为有限总体和无限总体。对于有限总体,既可以进行全面调查,也可以进行非全面调查;对于无限总体,则只能进行非全面调查。在统计研究过程中,总体取决于统计研究的目的和任务,有什么样的研究目的就要求有什么样的总体与之相适应。

总体具有以下几个基本特征:

1. 大量性

总体的形成要有一个相对规模的量。总体是由许多个别单位组成的。少数单位或个别事物是不能构成总体的。统计研究的目的是揭示自然和社会经济现象的规律性。人们只有通过大量事物的观察、分析和研究,才能发现其规律性。总体的大量性与研究目的的要求有关,精确度要求越高,总体中的单位数目就越大。

2. 同质性

构成总体的各个单位必须在某一方面且至少在一个方面具有某种共同的性质,这是形成统计总体的前提条件。如果两个个体没有同质性特征,就不能成为同一个总体的个体。换言之,总体是由有某一共同性质的基本单位组成的。例如,将某省所有高校作为总体,是因为每所学校都属于该省管辖且具有高校的性质。

3. 差异性

构成总体的各单位除必须具有某一共同标志表现作为它们形成总体的客观依据外,还必须在研究标志上存在差异;否则,就没有必要进行统计分析了。也就是说,总体的单位除在某一方面必须是同质的外,在其他方面还要有差异。例如,工业企业生产的产品的质量会有差异。这些差异是统计研究的基础,如果总体各单位之间不存在任何差异,就没有必要进行统计分析。

(二) 样本

统计研究最终要确定总体的数量特征,但是有时总体的单位数量很多,甚至无限多,那么就不可能或没必要对每个总体单位都进行调查。这时就要借助样本来研究总体了。样本就是按一定的概率从总体中抽取并作为总体代表的一部分单位的集合。

样本是统计学中非常重要的概念。人们对这一概念的理解要注意以下三个方面的问题:

(1) 构成某一样本的每个单位都必须取自某一特定的统计总体,不允许该总体之外的单位进入该总体的样本。

(2) 样本单位的抽取是按一定的概率进行的,而具体样本的产生应是随机的,因此必须排除人的主观因素对样本单位抽取和样本生成的干扰。

(3) 样本是总体的代表,带有总体的信息,因而能够推断总体。然而,样本只是总体的一个子集,而且具有随机性,故由样本去推断总体会产生代表性误差。其实,如何从总体中抽取样本,怎样控制样本对总体的代表性误差,是推断统计学研究的主要问题。

二、总体单位和标志

(一) 总体单位

总体单位是指构成统计总体的个别事物。例如,以某市所有工业企业的某种生产设备为总体,则每个该种类型的设备就是一个总体单位。根据研究目的的不同,总体单位既可以是人、物,也可以是企业、机构、地域,甚至可以是状况、长度或时间等。

总体与总体单位存在以下关系:

(1) 总体和总体单位是整体与部分的关系。总体是整体,总体单位是个体。总体的特征是通过总体单位归纳综合加以体现的。

(2) 总体和总体单位的划分不是绝对的,而是随着统计研究目的的变化而变化的。总体和总体单位不是固定不变的,而是随着研究目的和研究任务的变化而变化的。也就是说,总体和总体单位是根据研究目的来确定的,范围有大有小,原来的总体有可能转变为总体单位,原来的总体单位也有可能成为总体。例如,当把一个班级作为总体时,每个学生就是这一总体中的总体单位;而把其中的一个学生当作调查对象,研究该学生各方面的情况时,该

学生就成了研究的总体。

(3) 总体和总体单位互为依存。没有总体单位, 总体就不存在; 没有总体, 总体单位也不复存在。

(二) 标志

标志是用来说明总体单位特征的名称。只要是能够说明总体单位某一方面属性特征的名称都可称为标志。例如, 对于人口普查中的某人来说, 有性别、年龄、政治面貌等方面的特征, 这里的“性别”“年龄”“政治面貌”等名称就是标志。可见, 同一个总体单位可以有无数个标志。

标志表现是指标志在总体单位上的具体表现, 用来表明总体单位的具体属性或数值。例如, 人口普查中“某人”的性别为“女”, 年龄为“19”, 政治面貌为“团员”, 此处的“女”“19”“团员”就是标志表现, 即标志值。

为了便于统计分析, 人们可以从不同的角度对标志进行分类。

1. 数量标志和品质标志

按说明现象的性质不同, 标志可分为数量标志和品质标志。一个总体单位常常是品质标志和数量标志兼而有之, 这就要求人们用不同的方法来研究某一总体单位的各个方面。

数量标志表示总体单位数量的特征, 用数值来表示。例如, 某班的每个学生的年龄、分数是说明该班总体单位数量特征的, 为数量标志。数量标志需要用数字来表示。

品质标志说明总体单位质的特征, 不能用数值来表示。例如, 某人的性别、民族、政治面貌、籍贯是说明总体单位属性特征的, 为品质标志。品质标志只能用文字来说明。

2. 可变标志和不变标志

按在总体中各单位的具体表现是否相同, 标志可分为可变标志和不变标志。

当一个标志在各单位的具体表现有可能不同时, 这个标志称为可变标志。可变标志的属性或特征的具体表现是由一种状态变为另一种状态, 统计上称为变异。因此, 可变标志也称为变异标志。例如, 在某班学生总体中, “年龄”“身高”“政治面貌”这些标志在各单位的表现可能不同。它们就是可变标志。在一个总体中, 如果不存在可变标志, 或者说所研究的现象总体中的各单位之间无任何差异, 那么就没必要进行统计分析研究了。

当一个标志在各单位的具体表现不不同时, 这个标志称为不变标志。它是构成总体的必要条件和确定总体范围的标准。例如, 在某班学生总体中, “班级”这一标志在各单位的表现都是相同的。为此, “班级”就是不变标志。在一个总体中至少要有一个不变标志才能把各单位结合成为一个总体。如果没有不变标志, 就无法形成总体。由此可见, 不变标志是组成同一总体的基础。

三、指标

(一) 指标的含义与特点

指标是反映总体数量特征的社会经济范畴。例如, 某企业 2015 年的生产总值为 5 000 万元。它是采用一定的统计方法对总体各单位的标志表现进行登记、核算、汇总而成的指标, 用以说明该企业生产总值这一数量特征。这一数量指标的名称是“生产总值”, 数值是“5 000 万元”。因而, 指标具有以下三大特点:

- (1) 数量性。统计指标都能用数量来表示。
- (2) 综合性。统计指标都是用来说明总体的。
- (3) 质的规定性。统计指标都具有一定的经济内容。

(二) 指标的分类

1. 数量指标和质量指标

按所说明总体现象内容的特征的不同,指标可分为数量指标和质量指标。

(1) 数量指标。数量指标反映总体某一特征的绝对数量。这类指标主要说明客观现象的总规模、总水平或工作总量,一般用绝对数表示。例如,某一地区的总人口、工业企业总数、国内生产总值等。

(2) 质量指标。质量指标反映总体的强度、密度、效果、结构和工作质量等。例如,人口密度、劳动生产率和资金利润率等。这类指标一般用平均数与相对数表示。质量指标的数值并不随总体范围的大小而增减。例如,甲城市有 100 万人,其人均收入为 2 500 元/月;乙城市有 30 万人,其人均收入为 3 000 元/月。

2. 总量指标、相对指标和平均指标

按具体内容和作用的不同,指标分为总量指标、相对指标和平均指标。

(1) 总量指标。总量指标是反映总体现象规模的指标,表明总体现象发展的结果,如城市总人口、国民生产总值等。

(2) 相对指标。相对指标是两个有联系的总量指标和平均指标相比较的结果,如资金利润率。

(3) 平均指标。平均指标是按某一数量标志说明总体单位一般水平的指标,如平均工资、平均成本等。

(三) 指标和标志的关系

1. 指标和标志的联系

标志是指标的基础。许多指标的数值都是由总体单位的数量标志的标志值汇总而来的。例如,2015 年某省大学生招生总量是由该省每个大学的招生量汇总而来的。没有各个高校的招生人数,就难以得到该省大学生总数的指标值。

当研究目的发生变化时,许多指标与数量标志之间存在着变换关系。例如,要研究某所大学的情况,该大学的在校生人数、总面积、总职工人数等就是指标;如果改为将该大学所在省的所有大学作为研究总体,那么该大学的在校生人数、总面积、总职工人数等就是数量标志,而该省所有大学的相应指标为统计指标。

2. 指标和标志的区别

指标说明总体的特征,标志则说明总体单位的特征。指标只反映总体的数量特征,所有指标的值都是数值性的,没有用文字说明的指标。标志则既有反映总体单位数量特征的,也有反映总体单位品质特征的。只是数量标志的值是数值性的,而品质标志的值是表示属性的,可以用文字、字母等来表示。

四、变量和变量值

在统计中,一般把表示现象某种特征的概念称为变量。标志和指标都可以称为变量。

广义的变量为可变的量,包括品质变量和数值变量。狭义的变量只包括数值变量,是指各种可变的数量标志和所有的统计指标,都是用数值表示的。变量值是变量的具体表现:品质变量值表现为具体的属性,数值变量值表现为一系列数值。

(一) 品质变量和数值变量

按属性不同,变量可分为品质变量和数值变量。如果一个变量的变量值由品质数据来记录,那么该变量就是品质变量。例如,“性别”就是一个品质变量,其变量值就是“男”或“女”;“产品等级”也是一个品质变量,其变量值是“一等品”“二等品”“三等品”等。如果一个变量的数值由数量数据记录,那么该变量就是数量变量,或称数值变量,如“产品产量”“商品销售额”等。

按取值是否连续,数值变量可分为离散型变量和连续型变量。离散型变量是指变量值只能以整数出现的变量,即在任意两个相邻的取值之间不会有其他取值。离散型变量的取值是有限的。例如,企业数、人数、医院数和商店数等。连续型变量是指变量值可进行无限分割的变量。也就是说,任意两个相邻的取值之间都可以有一系列数值存在,这些变量就属于连续型变量。例如,某人的身高、体重,零件的尺寸误差,等等。在某些特殊场合,连续型变量也可作为离散型变量处理。例如,某班大学生按年龄分组时,可分为18,19,20,21...,其含义是“18”表示满18岁而不足19岁,“19”表示满19岁而不足20岁,依此类推。

(二) 自变量和因变量

自身变化会引起其他变量变化的量称为自变量;受其他变量影响而变化的量称为因变量。比如,产品产量的高低会影响生产费用的大小。此处的产品产量是自变量,生产费用是因变量。

(三) 确定性变量和随机性变量

确定性变量是指受确定性因素影响的变量。受某些确定性因素的影响,现象的量会沿着某一方向持续变化。这样的量就是确定性变量。例如,由于科学技术的不断提高和医疗卫生条件的不断改善,人类的死亡率在逐步降低,人类的平均寿命在不断延长。这些就是确定性变量。

随机性变量是指受随机性因素影响的变量。例如,从一个总体中抽取5个个体作为样本,每个个体被抽中的概率就是随机变量。

五、统计数据的计量尺度与类型

(一) 统计数据的计量尺度

统计数据是对客观现象进行计量的结果。人们在收集数据之前会先对现象进行计量或测度,这就涉及计量尺度的问题。由于现象信息的性质不同,采用的计量或测度的程度也就有区别,因此,只能对有些事物的属性进行分类。例如,人们对一些事物的看法,人口的性别、籍贯和文化程度,商品的品牌、型号和质量等级等只能按其属性进行分类;而人的身高、体重、脉搏和血压,商品的价格、价值等,则可以用比较精确的数字来计量。从事物现象计量的精确程度来看,采用数字来计量要比分类计量更精确一些。因此,统计计量或测度就有定性计量和定量计量之分。

根据计量学的一般分类方法,按对事物计量的精确程度,可以将采用的计量尺度由低级到

高级、由粗略到精确分为四个层次,即定类尺度、定序尺度、定距尺度和定比尺度。在统计中,人们应对不同类型现象的数据资料采用不同的计量尺度,进而应用不同的统计分析方法。

1. 定类尺度

定类尺度是按事物的某种属性对其进行平行的分类或分组。定类尺度计量的结果只表现为某种类型。它是最粗略、计量层次最低的计量尺度。例如,人口按性别、民族、政治面貌或职业等划分。这种划分只是将研究现象划分为不同的类。这种定类尺度只是测定了事物之间在类别上的差异。各组或各类之间是并列的和平等的,因而可以改变类别之间的顺序,而不会使事物的性质发生变化。但是,各组或各类之间必须是互相排斥的,即某一个人只能属于其中的一类。定类尺度是对事物的最基本测度,是其他计量尺度的基础。

为了便于统计处理,特别是为了便于计算机的识别而对不同的类别用自然数字表示或用编码表示,这并不说明这些数字可以区分大小或进行任何数学运算。例如,在按性别划分时,“1”表示男,“2”表示女;在按籍贯进行分类时,“1”表示北京,“2”表示上海,“3”表示广州等。这些数字不能进行数学运算和大小比较。根据定类尺度的计量结果,人们可以计算每一类别中各元素或个体出现的频数。

2. 定序尺度

定序尺度是对事物之间等级差别或顺序差别的一种测度,是测度序列变量的尺度。它不仅具有定类尺度的特点,将所有单位按互斥和穷尽原则进行分类,还可以确定这些类别的优劣、顺序或等级。例如,将产品按等级分为优等品、合格品、次等品,学生成绩按五级制分为优、良、中、及格、不及格等,这些就是用定序尺度计量。虽然定序尺度与定类尺度的计量一样表现为类别,但这些类别之间是可以比较顺序的,这种尺度比定类尺度的精确性要高一些。不过,定序尺度也只能表示一个类别优于另一个类别,而不能进行具体的数学运算。

3. 定距尺度

定距尺度是比定序尺度更高层次的定量尺度,是对事物类别或次序之间间距的测度。它不仅能够将事物区分为不同类型并进行排序,还可以准确地指出类别之间的差距,是一种较精密的计量尺度。定距尺度通常使用自然的或物理的单位作为计量的尺度。例如,温度按摄氏度计量,学习成绩用百分制计量等。定距尺度的计量结果表现为具体的数值,但它只能进行加减运算而不能进行乘除运算。例如,学习成绩 100 分与 60 分之间的差距是 40 分;甲地某日最高气温为 20°C ,最低气温为 15°C ,则甲地该日最高气温比最低气温高 5°C 。

另外,定距尺度可以转化为定序尺度。例如,百分制的学习成绩可以很容易地转化为 A、B、C、D、E 等级制的学习成绩。

4. 定比尺度

定比尺度又称比率尺度,与定距尺度属于同一层次,也是可以用数值对事物进行计量的较精密的计量尺度。它除具有上述三种计量尺度的全部特性外,还能进行乘除运算。定比尺度与定距尺度的区别在于对“0”的理解不同:在定距尺度中,“0”表示某一个数值,不表示“没有”或“无”;而在定比尺度中,“0”则表示“没有”或“无”。例如,在摄氏度体系中,“0”表示在海平面气压下水结冰的温度,不表示“没有”或“无”,因而温度是定距尺度;而人的体重为 0 千克,则表示这个人不存在,这里的“0”表示“没有”或“无”。类似地,像事物的长度、重量,商品的数量、价格、增加值等,这些数据的“0”都表示该事物不存在。

在现实生活中,大多数情况下使用的是定比尺度。由于在定比尺度中“0”表示不存在,其数值不仅可以比较大小、计算差值,还可以计算数值增加的比值。例如,张强的身高是180 cm,李明的身高是90 cm,则可以说张强的身高是李明的身高的2倍,或者说张强的身高超出李明的身高1倍。对于定距尺度,由于它不存在绝对零点,因此只能比较数值的大小,而不能计算其比值。例如,我们不能说20℃是15℃的几倍,只能说20℃比15℃高5℃。

采用上述四种计量尺度对事物进行计量时,必须符合穷尽和互斥的原则。穷尽原则是指在所做的全部分类中,必须保证每一个元素或个体都能归属于某一类别,不能有所遗漏。互斥原则是指每一个元素或个体只能归属于某一类别,而不能在其他类别中重复出现。一般来说,高层次计量尺度可以计量低层次计量尺度能够计量的事物,反之则不能成立。高层次计量尺度的计量结果可以很容易地转化为低层次计量尺度的计量结果,反之则没那么容易。例如,学习成绩由百分制容易转化成五等级制。

在统计分析中,一般情况下要求测量的层次越高越好,因为高层次计量尺度包含更多的数学特性,如表1-1所示。所运用的统计方法越多,分析时也就越方便。因此,人们在统计中应尽可能地使用高层次的计量尺度。

表 1-1 计量尺度的比较

数学特性	定类尺度	定序尺度	定距尺度	定比尺度
分类(=,≠)	√	√	√	√
排序(<,>)		√	√	√
间距(+,-)			√	√
比值(×,÷)				√

(二) 统计数据类型

人们在统计数据时采用的计量尺度不同,得到的数据也不同。统计数据大体上可以分为两种类型:定性数据和定量数据。

定性数据又称为品质型数据,说明事物的品质特征表现的具体类别,不能用数值表示,其结果通常表现为类别。这类数据是由定类尺度和定序尺度计量得到的,如人的性别、民族、籍贯等数据资料,一般表现为文字。但是,人们在整理统计数据时,为了方便计算机识别,通常给予其相应的编号。

定量数据又称为数值型数据,说明现象的数量特征表现,能够且必须用数值表示。这类数据是由定距尺度和定比尺度计量形成的,如人的年龄、身高、体重等数据资料。

对于定性数据,主要采用分组法计算分析各组的频数、频率和众数等;定量数据既可以用于分析频数、频率等,又可以用于进行频数累计、频率累计等分析处理。



思考练习 >>>>

一、单项选择题

1. 社会统计学派是在()的基础上发展起来的。
 - A. 政治算术学派
 - B. 数理统计学派
 - C. 推断统计学派
 - D. 国势学派

第二章

统计数据的收集、整理和显示

学习目标

掌握统计数据的调查方式、收集方法,以及调查方案和调查问卷的设计;
了解统计数据整理的内容及意义;
掌握统计分组的类型,以及频数分布表的编制方法;
结合统计数据,掌握绘制统计表和统计图的方法。

引例

北京市 2015 年暨“十二五”时期国民经济和社会发展统计公报(节选)

2015 年,全市人民在党中央、国务院和市委、市政府的坚强领导下,正确把握新时期首都城市战略定位,积极推进京津冀协同发展,有序疏解非首都功能,加快培育“高精尖”经济结构,大力统筹稳增长、促改革、调结构、惠民生、防风险等各项工作,经济社会保持了平稳健康发展,“十二五”规划顺利完成。

一、人口

年末全市常住人口为 2 170.5 万人,比上年末增加 18.9 万人。其中,常住外来人口为 822.6 万人,占常住人口的比重为 37.9%。常住人口中,城镇人口为 1 877.7 万人,占常住人口的比重为 86.5%。常住人口出生率 7.96‰,死亡率 4.95‰,自然增长率 3.01‰。常住人口密度为每平方公里 1 323 人,比上年末增加 12 人。年末全市户籍人口为 1 345.2 万人,比上年末增加 11.8 万人。

“十二五”时期,全市人口增速、增量“双下降”,常住人口年均增长 2.0%,比“十一五”时期低 3 个百分点;累计增加 208.6 万人,增量比“十一五”时期减少 215.3 万

人。常住人口增速呈放缓态势,常住人口增量逐年减少。

二、综合

经济增长:初步核算,全年实现地区生产总值增长 22 968.6 亿元,比上年增长 6.9%。其中,第一产业增加值 140.2 亿元,下降 9.6%;第二产业增加值 4 526.4 亿元,增长 3.3%;第三产业增加值 18 302 亿元,增长 8.1%。

按常住人口计算,全市人均地区生产总值达到 106 284 元(按年平均汇率折合 17 064 美元)。

“十二五”时期,全市地区生产总值年均增长 7.5%,低于“十一五”时期平均增速 3.9 个百分点,其中第一产业年均下降 0.6%,第二产业和第三产业年均分别增长 6.5%和 8.0%。三次产业结构由 2010 年的 0.9 : 23.6 : 75.5,调整为 2015 年的 0.6 : 19.6 : 79.8。

全年文化创意产业实现增加值 3 072.3 亿元,比上年增长 8.7%;占地区生产总值的比重为 13.4%,比上年提高 0.2 个百分点。高技术产业实现增加值 5 180.8 亿元,增长 9.3%;占地区生产总值的比重为 22.6%,比上年提高 0.4 个百分点。信息产业实现增加值 3 508 亿元,增长 10.6%;占地区生产总值的比重为 15.3%,比上年提高 0.4 个百分点。生产性服务业实现增加值 12 160.3 亿元,增长 8.6%;占地区生产总值的比重为 52.9%,比上年提高 0.4 个百分点。

财政:全市完成一般公共预算收入 4 723.9 亿元,比上年同口径增长 12.3%。其中,增值税 716.1 亿元,增长 10.7%;营业税 1 186.1 亿元,增长 11.0%;企业所得税和个人所得税分别为 1 024.7 亿元和 478.1 亿元,分别增长 11.9%和 24.7%。一般公共预算支出 5 751.4 亿元,增长 27.1%。其中,用于城乡社区、节能环保、交通运输、社会保障和就业的支出分别增长 77.9%、42.1%、37.8%和 37.6%。

“十二五”时期,一般公共预算收入和一般公共预算支出累计分别达到 18 733.4 亿元和 21 380.3 亿元,分别是“十一五”时期的 2.1 倍和 2.2 倍。

该公报给出了很多数字。这些数字是如何收集和整理获得的?统计数据的收集是统计工作的第二阶段,是人们利用统计方法进行分析的基础。离开了统计数据,统计方法就成了“无米之炊”,就失去了用武之地。因而,获取的数据资料应该具有准确性、及时性、全面性和系统性。所谓统计数据的收集,就是按统计研究的目的和任务,运用各种科学有效的方式和方法,有针对性地收集反映客观现实数据的活动过程。统计数据的整理与显示是统计工作的第三阶段。本章主要介绍如何采用科学的统计调查方式、方法来获取统计数据,以及对这些统计数据整理与显示。

第一节 统计数据的收集

统计数据的收集是整个统计工作的一个有机组成部分,是统计工作的基础环节。那么,数据资料的来源是关键。通常情况下,数据资料的来源有两个:一是间接来源,通过出版物

等收集;二是直接来源,通过调查或科学实验来采集。

一、统计数据的间接来源

间接来源的数据也称二手数据。数据的间接来源是人们获得二手数据的重要来源。二手数据是源于他人调查或试验取得的数据,也称为次级资料,其来源十分广泛。对于数据使用者来说,即使是专业的调查机构也不会亲自去调查所有的数据,况且有些一手的统计资料也是无法直接获取的。因此,人们所使用的数据大多是通过别人的调查或科学实验得来的。

二手数据主要是公开出版或公开报道的数据,主要来自国家和地方的统计部门及各种报刊媒介。例如,中国统计出版社出版的《中国统计年鉴》《中国统计摘要》《中国工业经济统计年鉴》《中国人口统计年鉴》等。提供世界各国社会和经济数据的出版物有中国社会科学出版社出版的《世界经济年鉴》、世界银行发布的各年度《世界发展报告》等。另外,数据使用者还可以通过其他渠道使用一些尚未公开的统计数据,以及广泛分布在各种报纸、杂志、图书、电视传媒中的各种数据资料。现在,随着计算机网络技术的发展,数据使用者也可以在网络上获取所需的各种数据资料。表 2-1 给出了国内提供统计数据的部分网站。

表 2-1 国内提供统计数据的部分网站

网 站	网 址	数据内容
国家统计局网站	www. stats. gov. cn	统计年鉴、统计月报等
中国经济信息网	www. cei. gov. cn	经济信息
中国气象网	www. cma. gov. cn	国家气象数据
中国决策信息网	www. jcxx. cc	决策知识及案例
中国统计信息网	www. tjc. org	中国统计信息
中国证券监督管理委员会网站	www. csrc. gov. cn	上市公司有关信息及有关法规

二手数据数量巨大、用途多样、信息丰富、来源广泛,具有及时、物美和价廉的优势,但在使用时应注意统计数据的含义、计算口径和计算方法,以免误用或滥用;同时,一定要注明数据的来源,以体现对他人劳动成果的尊重。

二、统计数据的直接来源

一般来说,人们进行一项统计分析,尤其是进行有特殊目的的分析,是很难从现有条件所提供的资料中获取全部数据的。因此,人们需要采用专门的方法来收集数据,以达到特殊分析的目的。统计数据的直接来源主要有两个:一是进行科学实验,这是自然科学进行研究获取数据的主要手段;二是开展统计调查,这是取得社会经济数据的主要手段。统计调查包括统计部门进行的统计调查,以及其他部门或机构为特定目的而进行的调查,如某企业为获得其产品目前在市场中的竞争情况的资料而进行的市场调查。下面主要介绍统计调查的相关内容。

(一) 统计调查的方式

比较典型的统计调查方式有统计报表和专门调查。其中,专门调查包括普查、抽样调查、重点调查与典型调查。

1. 统计报表

统计报表是我国统计调查体系中的一种重要的组织方式。它是根据国家统计报表制度规定的统一表式、统一指标内容、统一报送程序和报送时间,由填报单位自下而上地逐级提供统计数据的一种统计调查方式。统计报表按报送范围可以分为全面统计报表和非全面统计报表;按报送时间可以分为日统计报表、月统计报表、季统计报表和年统计报表;按报送受体可以分为国家统计报表、部门统计报表和地方统计报表。

统计报表的作用主要表现在以下几个方面:

(1) 统计报表是国家获得统计资料的重要途径。报表资料是国家制订计划和检查计划执行情况的主要依据。

(2) 通过完整地积累报表资料,可以满足各种分析的需要。

(3) 各级领导部门可以通过统计报表资料经常了解本部门、本地区的经济和社会发展规划,以便做出有利的决策。

2. 专门调查

(1) 普查。普查是专门组织的一次性全面调查,主要用于收集处于一定时点上的社会经济现象的数量,具有特定的目的、特定的对象。普查旨在收集有关国情国力的基础统计数据,为国家制定重大方针政策、编制国民经济和社会发展规划提供依据。普查作为一种特殊的调查组织方式有以下几个特点:

① 普查需要规定统一的标准调查时间,以避免调查数据的重复或遗漏,保证调查结果的准确性。我国历年人口普查及其调查时点如表 2-2 所示。

表 2-2 我国历年人口普查及其调查时点

普查次数	普查时点
第一次全国人口普查	1953 年 6 月 30 日 24 时
第二次全国人口普查	1964 年 6 月 30 日 24 时
第三次全国人口普查	1982 年 7 月 1 日零时
第四次全国人口普查	1990 年 7 月 1 日零时
第五次全国人口普查	2000 年 11 月 1 日零时
第六次全国人口普查	2010 年 11 月 1 日零时

② 普查通常是一次性的或周期性的。由于普查涉及面广、调查单位多,需要花费大量的人、财、物和时间,加上要便于分析研究调查对象发展变化的趋势和规律,一般需要间隔较长的时间进行一次。

③ 普查得到的数据资料一般都比较准确,规范化程度也比较高,可作为抽样调查和其他调查的依据。

④ 普查适宜的对象比较狭窄,适宜调查一些最基本的和最一般的现象。

(2) 抽样调查。抽样调查是一种非全面调查,是按一定的概率从总体中抽取一部分单位构成样本,并依据所获得的样本数据对研究对象总体的数量特征做出具有一定可靠程度的估计和推算。它既是收集统计资料的方法,又是对现象总体做出具有一定可靠性的估计推断的方法,在统计调查和统计分析中应用广泛。

抽样调查包括概率抽样与不等概率抽样。概率抽样即随机抽样,是以概率论为基础,按随机原则抽取样本的抽样方法。不等概率抽样是调查者根据自己的主观判断来抽取样本的方法。不等概率抽样不是抽样调查的重点,此处不再介绍。常用的概率抽样可以分为简单随机抽样、等距抽样、分层抽样、整群抽样和多阶段抽样等。

抽样调查在统计调查中发挥着重大作用,其主要有以下几个特点:

① 节省经费。由于抽样调查的调查单位少,不需要对总体中的每一个个体进行调查,因此大大减少了开支。同时,由于抽样调查的范围较集中,调查工作量可以减轻,调查、登记、汇总等工作都可以专业化,因而节省了人力、物力和财力。

② 时效性强。抽样调查所要收集、处理的数据远远少于全面调查,并且它由专门队伍深入实际、直接取样,因而提高了调查的时效性。

③ 精确度较高。由于抽样调查是自上而下进行的组织调查,工作人员训练有素,而且按随机原则抽取样本,抽样误差可以加以控制,统计计算精确,使得样本有较高的代表性,调查结果比较准确可靠。

④ 适应面广。抽样调查的组织方便灵活,调查项目可多可少,调查范围可大可小,既适用于专题研究,也适用于经常性调查。人们可以根据自己的需要组织抽样调查。

(3) 重点调查。重点调查是专门组织的非全面调查,是从全部单位总体中选择少数重点单位进行的调查。重点调查的关键是正确选择重点单位。重点单位的选择是根据研究的任务和目的有意识地决定的。重点单位需要具备两个标准:一是被选择的重点单位数在总体单位数中所占的比重较小;二是这些重点单位被调查的标志值之和占全部总体指标数值的比重较大。这两个标准必须同时具备,缺一不可。例如,调查人员要了解全国的钢铁生产总量,只需要对产量很大的少数几个钢铁企业,如宝钢、首钢、鞍钢等的钢铁生产情况进行了解就可以了,因为它们的钢铁产量在全国钢铁产量中占很大的比重。重点调查的优点是由于调查单位较少,调查项目的指标可以多一些,了解的情况可以更加详细,花费的资源相对较少。但是,其调查结果一般不能推断总体。

(4) 典型调查。典型调查是根据调查研究的目的和要求,在对被研究的现象总体进行初步分析的基础上有意识地选择一些典型单位进行深入的、细致的、周密的调查,深入了解情况,以认识事物本质和一般规律的一种科学的调查方法。因此,典型调查所选择的典型单位应能反映所研究问题的本质属性或特征。例如,调查人员要调查某类企业的经济效益问题,就可以从某类企业中选择一个或几个经济效益突出的单位展开深入调查,从中总结经济效益好的经验。典型调查主要用于定性研究,其调查结果一般不能推断总体。

通常,以上统计调查方式是结合运用的。在实际工作中,以周期性普查为基础,以经常性抽样调查为主体,以必要的统计报表、重点调查等为补充收集、整理基本统计资料。

(二) 统计数据的收集方法

调查人员在确定调查方式之后都要运用一定的数据收集方法。收集统计数据的方法有询问调查和观察实验两种。

1. 询问调查

询问调查具体包括访问调查法、直接观察法、电话法、座谈会法、网络调查法等。

(1) 访问调查法。访问调查法是调查者与被调查者通过面对面交谈从而得到所需资料

的调查方法。访问调查法又分为标准式访问和非标准式访问。标准式访问是按调查者事先设计好的有固定格式的标准化问卷和表格有顺序地依次提问,并由被调查者做出回答。非标准式访问事先不制作固定格式的标准化问卷和表格,调查者无顺序地提问,与被调查者自由交谈,从而获得所需资料。

(2) 直接观察法。直接观察法是调查者亲自到现场对调查对象进行观察和计量以取得数据资料的一种数据采集方法。

(3) 电话法。电话法是调查者通过电话与被调查者进行语言交流,从被调查者回答的信息中收集统计数据资料的一种数据采集方法。

(4) 座谈会法。座谈会也称为集体访谈法,是将一组被调查者集中在调查现场,让他们对调查的主题发表意见,从而获取资料的方法。参加座谈会的被调查者应是所调查问题的专家或有经验者,通常有6~10人。调查者向被调查者说明调查的目的、要求和具体内容,被调查者可以对不清楚的地方提出疑问,由调查者解答。由于在彼此交流的环境里,被调查者相互影响、启发,不断补充、修正各自的观点,有利于调查者从中获得更多的想法和意见。

(5) 网络调查法。网络调查法是利用网络进行调查的方法。其有两种形式:一种是利用网络直接采用问卷方式收集一手资料;二是利用网络的媒体功能,从网络上收集二手资料。

(6) 个别深度访问法。个别深度访问法是一次只有一名被调查者参加,针对特殊问题的调查。个别深度访问法适合于较隐秘的问题(如个人隐私问题)或较敏感的问题(如政治问题),侧重于定性研究。

(7) 计算机辅助调查法。计算机辅助调查法又称计算机辅助电话调查法,是由计算机与电话相结合完成调查的全过程,一般需借助专门的软件进行,对硬件设备的要求较高。

2. 观察实验

观察实验是调查者通过直接的观察或实验获得数据的方法。观察实验具体包括观察法和实验法。

(1) 观察法。观察法是指针对调查对象的行动和意识或存量变化,由调查者边观察边记录的收集信息的方法。观察者利用感官或借助各种现代化的仪器和手段,如照相机、录像机、显微录像机等间接地从侧面观察和记录。采用观察法,由于调查者不强行介入,因而常常能在被观察者不察觉的情况下获得信息资料。例如,销售人员在售卖点进行观察。

(2) 实验法。实验法是指在所设定的特殊实验场所、特殊状态下,对调查对象进行实验以取得所需资料的一种调查方法。实验法可分为在室内进行的室内实验法和在市场或外部进行的市场实验法。室内实验法可用于广告认知的实验。例如,调查人员如让实验者对同一产品观看新制作的不同画面的广告效果。市场实验法可用于消费者需求的调查等。例如,企业让消费者在一段时间内免费使用一种新产品,以得到消费者对新产品看法的资料。

(三) 统计调查方案的设计

统计调查需要投入一定的人力、物力和财力,特别是对大型调查而言,调查工作就是一项庞大的系统工程。例如,全国人口普查。因此,调查者应该有计划地、科学地组织这项活动,要根据需要与可能设计出切实可行、周密完整的统计调查方案。统计调查方案是指在开

始调查之前根据统计调查的目的和任务,按调查对象的特点,对统计调查工作各方面和各环节做出的全面部署和安排。其是保证调查工作顺利开展的指导性文件。完整的统计调查方案主要包括以下内容:

1. 确定调查目的

统计调查目的是进行统计调查所要达到的目标和需要解决的主要矛盾和问题。调查目的必须十分明确,中心突出,才能有的放矢。它回答的是为什么实施调查,要解决什么样的问题,调查具有什么样的社会经济意义。目的不明确,就无法确定向谁调查,调查什么,以及用什么方式、方法进行调查,其后果必然是调查活动的盲目进行。调查人员要避免那种企图在一次调查中解决所有问题的倾向。例如,我国第六次全国人口普查的目的是这样表述的:“人口普查是一项重大的国情国力调查。2000年第五次全国人口普查以来,我国的人口状况发生了很大变化。组织开展第六次全国人口普查,将查我国人口在数量、结构、分布和居住环境等方面的变化情况,为科学制定国民经济和社会发展规划,统筹安排人民的物质和文化生活,实现可持续发展战略,构建社会主义和谐社会,提供科学准确的统计信息支持。”

2. 确定调查对象

调查对象就是调查单位,是为了解决向谁调查、由谁具体提供统计资料的问题。调查对象是调查项目的承担者,是提供统计数据的基本单位。例如,我国第六次全国人口普查的对象是在中华人民共和国境内居住的自然人,即人口普查的调查对象是每一个人。再如,若要取得某地区商业单位的利润统计数据,就可以将该地区的商业单位确定为调查对象。

调查人员还需要对调查单位和填报单位加以区分。填报单位又称报告单位,是统计调查中负责向上级报告调查内容、提交统计资料的单位。在调查中,调查单位和填报单位可能一致,也可能不一致。总之,确定调查单位可使人知道从哪里取得有关标志的具体资料,其可以是人、企事业单位和物;确定填报单位可使人们知道由谁提供具体资料,其只能是某个人或单位。

3. 确定调查项目和调查表

确定调查项目是要回答调查什么问题。具体来说,调查项目就是所要调查的总体单位的数量标志和品质标志。在多数调查中,调查项目通常以调查表的形式出现。

将调查项目按一定的顺序排列在表格上就形成了调查表。调查表便于调查登记资料的规范化与标准化。调查表有单一表和一览表两种形式。单一表是在一张调查表上只登记一个调查单位的项目,适用于调查项目较多、比较详细的调查,便于进行分类整理。一览表是在一张调查表上登记若干调查单位的项目,容纳的调查项目较少,便于比较各个调查单位的情况,以及核对和汇总计算。

调查表的格式一般由表头、表体和表脚三部分组成。表头在调查表的上方,包括调查表的名称及填报单位的名称、地址、性质、隶属关系等。它们是核实和复查各调查单位资料时必不可少的内容。表体是调查表的主要部分,包括调查项目和项目的具体表现、栏号、计算单位等,便于在整理调查资料时或编写填表说明时使用。表脚通常由填表人签名、填报日期、填表说明、项目解释等内容组成。填表说明用来说明填表时应注意的事项;项目解释则是对调查表所列各调查项目、统计指标等进行统一的规范说明,保证调查资料的准确性和标准化。填表说明和项目解释必须符合国家制定的统一标准。

4. 确定调查时间和时限

调查时间是调查资料所属的时间,即所谓的客观时间。如果所要调查的是时期现象,调查时间就是资料所反映的起止日期;如果调查的是时点现象,调查时间就是规定的统一标准时间。

调查时限是进行调查工作的期限,包括收集资料和报送资料的整个工作所需要的时间,即所谓的主观时间。为保证调查资料的及时性,任何调查都应尽可能缩短调查时限。

5. 确定方案设计中的其他内容

在调查方案中,调查人员还必须研究确定调查所采用的方式和方法,以及调查工作的组织安排等,使调查工作在组织上、措施上有保证。组织工作计划包括明确调查机构与调查地点,选择调查方法等内容。此外,在调查组织工作计划中,对于调查前的准备工作,包括宣传教育、调查员培训、文件印刷、调查资料报送办法、调查经费的预算和开支办法、提供或公布调查成果的时间等均应做出具体规定。

对于大规模统计调查,所制订的调查方案往往需要进行试点调查。通过试点调查,调查人员可以检验调查方案是否可行,以便加以修改和补充。调查人员还要积累实施调查方案的经验,提高业务技能,以便圆满地完成调查任务。

(四) 统计调查问卷的设计

1. 统计调查问卷的含义及特点

统计调查问卷又称调查表,是指调查者根据调查目的和要求设计出的,由一系列问题,备选答案、说明及代码表组成的,用来收集调查数据资料的一种书面问卷。

设计完成的问卷应具有以下几个特点:

(1) 规范性。问卷设计应能够达到满足调查客户需求的目的,具体体现在完整性、准确性和可行性上。

(2) 稳健性。问卷设计应尽量减少由于调查者、被调查者和调查环境的不同可能引起的数据波动。

(3) 效率性。问卷设计应在保证获得同样信息的条件下,选择最简洁的询问方式,使问卷的题量和难度最小,以节约调查成本。

(4) 可维护性。问卷应当结构清晰,便于修改。各项目之间要有明确的界限,保证一个项目的调整不会影响到其他项目。

此外,问卷设计还要注意可比性和连续性,这也是使问卷数据利用价值提高的重要方面。

2. 问卷的基本结构

不同的调查问卷在具体结构、题型、措辞、版式等设计上会有所不同,但在结构上一般都由开头部分、甄别部分、主体部分和背景部分组成。

(1) 开头部分。开头部分一般包括问候语、填表说明和问卷编号等内容。开头部分主要是对调查目的、调查意义、调查组织者的身份和调查的主要内容等进行说明,文字须简明易懂,力求引起被调查者的重视与兴趣。开头部分的措辞要态度诚恳,口吻亲切,并要对被调查者表示真诚的感谢。有时,开头部分还要向被调查者说明问卷填写的方法和要求,以及

需要注意的有关事项。

(2) 甄别部分。甄别又称过滤,目的是筛选掉不需要、不合适的部分,只针对特定的对象进行调查。甄别部分主要用于筛选与调查项目有直接关系的人,达到避嫌的目的,确保被调查者能够作为该调查项目的合格代表,满足调查研究的需要。

(3) 主体部分。问卷主体是调查所要收集的主要信息,由一个个问题及相应的答案组成。通过主体部分问题的设计和被调查者的答复,调查者可以对被调查者的个人基本情况和对某一特定事物的态度、意见倾向及行为有较充分的了解。这一部分设计得如何直接关系到本次问卷调查能否取得有价值的资料。

(4) 背景部分。背景部分主要表示对被调查者合作的感谢,或用以征询其对问卷设计和问卷调查的意见和感受,以及记录下调查者的姓名、调查时间、调查地点等。背景部分要简短明了。有的问卷也可以省略背景部分。

3. 问卷的设计程序

问卷的设计是由一系列相关的工作过程构成的。为使问卷具有科学性、规范性和可行性,调查人员一般可以参照以下程序对问卷进行设计:

(1) 确定调查目的、资料 and 数据的收集方法。调查人员要根据调查目的确定所需要的资料、调查所采用的方式和方法,以及资料的汇总和分析方法。

(2) 设计提问项目。一份问卷的内容不宜过多,不必要的问题不要列入。调查人员在设计问卷时应注意问题的措辞,即用词必须清楚;一个问题只包含一个内容;避免诱导性的用语;避免否定形式的问题;考虑应答者回答问题的能力;考虑应答者回答问题的意愿。

调查人员应站在被调查者的角度,顺应被调查者的思维习惯,使问题的安排具有逻辑性,把能引起被调查者兴趣的问题放在前面,把开放性问题放在后面。

(3) 确定问题的回答形式。由于问卷中的问题有不同类型,所以设计的答题类型和对被调查者的回答要求也是不同的。问卷中的问题类型主要有开放式问题、封闭式问题和量表应答式问题。开放式问题是一种应答者可以自由地用自己的语言来回答和解释有关想法的问题类型。也就是说,在开放式问题中,调查者没有对应答者的选择进行任何限制。封闭式问题是一种需要应答者从一系列应答项中做出选择的问题。量表应答式问题则是以量表形式设置的问题。

(4) 评估问卷。为强调问卷的针对性、经济性、有效性和完美性,调查人员有必要在初步设计好问卷后对满足目的方面,问卷问题的必要性方面,问卷的长度、编码、项目等方面进行检查评估,以使正式的调查问卷科学和适用。

(5) 对问卷的模拟试验。在问卷设计过程中,经常的修改是必需的。调查人员在没有进行预先测试前,不应当进行正式的询问调查。在正式调查之前,调查人员应通过问卷的模拟试验寻找问卷中存在的错误解释、不连贯的地方、不正确的跳跃模型,并进行适当的修改。

4. 问卷的问题设计

问卷设计必须注意问题的设计。问卷中的问题合理、排列科学,可以提高问卷的回收率和信息的质量。

理想的问卷设计应既能使调查者获得所需的信息,又能使被调查者轻松、方便地回答问题。这就要求调查者能依据具体调查内容及要求设计选用合适的问句进行调查。问句的类

型有很多。按问题的内容,问句可以分为事实问句、意见问句与阐述问句;按回答形式,问句可以分为自由回答式问句、多项选择式问句、顺位式问句、程度评等式问句和过渡式问句等。

三、统计数据的质量

(一) 统计数据的质量要求

统计数据的质量要求通常有以下几点:

- (1) 精度,即最低的抽样误差或随机误差。
- (2) 准确性,即最小的非抽样误差或偏差。
- (3) 关联性,即满足用户决策、管理和研究的需要。
- (4) 及时性,即在最短的时间里取得并公布数据。
- (5) 一致性,即保持时间序列的可比性。
- (6) 最低成本,即在满足以上标准的前提下以最经济的方式取得数据。

因此,调查人员在取得统计数据时要做到统计数据来源要有依据,所依据的数据要得到证实,得到证实的数据要定期进行检查。调查人员要通过检查、证实、统计,实时对统计过程进行补充和修改,以使统计数据的质量得到不断改进。

(二) 统计误差

统计的整个工作过程就是对数据的加工过程,从原始数据的收集开始,经过整理、显示、样本信息的提取到总体数量规律性的科学推断都有一个减少误差、提高数据质量的问题。也就是说,统计数据的质量控制问题是贯穿统计研究全过程的重要问题。但是,在不同的统计工作阶段,统计数据误差产生的原因是不同的,误差的程度也不同。

统计调查阶段是收集统计数据的阶段,因而这一阶段统计数据的质量会直接影响整个统计工作的顺利开展。这一阶段的统计误差可分为非抽样误差和抽样误差。

1. 非抽样误差

非抽样误差是指调查过程中由于调查者违反随机原则,抽样不准确,问卷设计不科学,观察、测量、登记、计算上的差错等使调查获得的原始资料不准确而引起的误差。所有统计调查都可能存在非抽样误差,且对其难以计算和控制。非抽样误差也称登记性误差。登记性误差是指由于错误判断事实或错误登记事实而发生的误差。不管是全面调查还是非全面调查都会有登记性误差。登记性误差又可分为偶然登记误差和系统登记误差。

① 偶然登记误差。形成偶然登记误差的原因很多,如调查者注意力不集中、技能低下所发生的遗忘、笔误、错填或者被调查者回答不当等。偶然登记误差的特点是不具有倾向性,即在数量上不偏向某一方。这类误差既可能被夸大,也可能被缩小。调查人员在大量调查资料进行整理时,这类误差通常会相互抵消。

② 系统登记误差。系统登记误差的特点是具有明显的倾向性、一贯性,而在数量上偏向某一方,因此又称偏差。例如,测量工具没有校正好、调查方案中的项目不明确、被调查者有意歪曲事实等。系统登记误差不管是有意的还是无意的,危害性都比较大,因为它对整理综合结果的指标影响很大。

2. 抽样误差

抽样误差也称随机抽样误差或实际抽样误差,是指由于随机抽样的偶然因素使样本不

足以代表总体而引起的样本指标与总体指标之间的绝对离差。抽样误差是利用样本推断总体时产生的误差。由于样本只是总体的一部分,用样本的信息去推断总体,或多或少会存在误差,因而抽样误差对任何一个随机样本来讲都是不可避免的。但是,它是可以计量的,并且是可以控制的。抽样误差也称代表性误差。

以上说明,登记性误差多由主观因素引起。人们可以通过完善调查方案,加强调查过程的检查与监督,提高调查者政治素质和业务技能水平,采取现代化收集信息的手段来提高调查资料的质量,把登记性误差减少到最低限度。通常,代表性误差是针对抽样调查而言的。

第二节 统计数据的整理和显示

在通过统计调查、科学实验或者从网络中获取了统计数据后,调查人员接下来要做的就是对数据进行整理,提取有用的信息。统计整理是统计工作的重要环节,也是统计学研究的重要内容之一。

一、统计数据的整理

统计数据的整理是按统计目的和任务的要求,根据统计对象的特点,对调查的统计资料进行分组、汇总,使其集中化、条理化、系统化的活动或过程。广义统计数据的整理既包括对原始资料的加工整理,也包括对已经加工过的次级资料的再整理。而作为一个相对独立的统计工作阶段来说,统计数据的整理主要是指对原始资料的整理。

1. 整理统计数据的意义

(1) 统计数据的整理是实现从单位到总体、从标志到指标的重要环节和手段。通过统计数据的整理,调查人员可以把反映各单位特征的标志汇总成为指标,获取对总体的基本认识;在分组整理的基础上,通过计算对比得到分析指标,以便通过综合指标对总体做出概括性的说明。

(2) 统计数据的整理全面实现了统计的数量化特征。通过统计数据整理,不仅把调查的数量标志汇总成了指标,还把调查的非数量化内容(品质标志和其他意向性选择等)汇总变成了反映总体数量特征的指标,从而为统计分析的科学化奠定了基础。

(3) 统计数据的整理是积累历史资料的必要手段。在统计研究中,调查人员经常会用到动态分析,这就需要有长期累积的历史资料。根据积累资料的要求,对已有的统计资料进行筛选,以及按历史口径对现有的统计资料进行重新调整、分类和汇总等都必须通过统计整理来完成。

随着电子计算机技术和网络技术的应用和普及,统计信息的来源和使用范围不断扩大,统计整理的概念和作用也极大地得到了扩展。统计数据的整理已不再局限于为了单一的统计研究目的而进行综合加工汇总,而是将大量的原始数据和再加工数据进行有序化、系统化并建立统计数据库,以满足统计研究的多方面、多层次需要。这不但推动了统计数据整理技术的发展,极大地提高了统计数据的整理速度,而且极大地提高了统计信息资源的开发利用程度。

2. 统计数据的整理步骤

从完整的工作程序来看,统计数据的整理步骤如下:

(1) 统计数据整理的方案设计。根据统计研究的目的和任务对整理工作做出统一的安排和布置,这就是设计统计数据整理方案。统计数据整理方案的主要内容包括统一的分组方法、统一的汇总内容、统一的工作程序安排、整理结果的统一表达方式等。

(2) 统计资料的审核与筛选。为了保证统计数据整理的正确性,在统计汇总之前,首先要对统计调查资料进行甄别和审核,以检查资料的准确性和完整性。准确性检查主要是通过计算检查和逻辑检查等方法来检查调查资料的真实可靠性;完整性检查主要是审核各项调查资料是否及时取得,调查项目的内容是否全面完整、是否有空缺和遗漏。审核及时性主要是检查统计调查的主观时间是否符合规定,包括调查期限、报送时间等。调查人员对在审核中发现的问题应及时进行纠正和补充。检查统计数据的准确性方法一般有两种:一种是逻辑检查,即检查原始资料是否合理,相互之间是否存在矛盾;另一种是计算检查,即从指标要素的角度来检查数字资料有无错误,如计算方法、计量单位、指标之间的关系是否成立等。

(3) 对原始资料进行排序、分组、汇总和计算。数据排序是指按一定顺序排列数据。排序借助计算机可以很容易完成。对于数字型数据,排序有递增和递减之分;对于字母型数据,排序有升序和降序之分;对于汉字型数据,排序既可以按拼音首字母排序进行,也可以按笔画的多少进行。通过排序,可以发现一些明显的特征趋势或解决问题的线索。统计分组和统计汇总就是按统计分组的要求,将全部调查资料划分成若干组成部分,根据汇总的要求和工作条件选择适当的汇总组织形式和汇总的具体方法,将个体资料加以归类,并加总计算各组的数值和总体数值,以获得统计整理的成果。可见,统计分组和统计汇总是统计整理的核心内容。

(4) 编制统计图表。统计数据的整理成果通常是采用统计表或统计图的形式表现的。统计表和统计图也是整理统计资料的重要工具。统计数据的整理是统计工作的重要环节,是一项经常性的长期任务。为了系统地收集和运用统计资料,调查人员必须注意采用各种科学有效的存储手段对统计资料进行积累,建立完善的统计资料的保管和提供制度。

二、统计分组

(一) 统计分组的概念

统计分组是按统计研究目的和任务的要求,根据统计对象的特点,采用一定的标志把总体划分为不同部分的一种统计分析方法。统计分组的对象是总体,使组内的差异尽可能小,组间的差异尽可能大,从而使大量数据变为有序的、层次分明的、显示总体数量特征的资料。

数据分组对现象总体而言是“分”,即把总体分为性质相异的若干部分;而对总体单位而言又是“合”,即把性质相同的许多总体单位合为一组。

(二) 统计分组的原则

统计分组除在内容方面必须反映各单位、各组之间的性质差异外,还要在方法上保证资料的完整性和真实性。这是对统计分组的最基本要求。为此,统计分组在方法上必须符合穷尽性原则和互斥性原则。

1. 穷尽性原则

穷尽性原则又称不遗漏原则,即统计分组必须保证总体的每一个单位都能归入其中的一个组,各组的单位数之和等于总体单位总量,总体的指标必须是各单位相应标志的综合。例如,从业人员按文化程度分组,可分为小学、中学(含中专)和大学三组,显然那些文盲及半文盲、大学以上的人员就无法被包括进来。

2. 互斥性原则

互斥性原则又称不重复原则,即统计分组必须保证总体的每一个单位只能属于其中的一个组,不能出现重复统计的现象。例如,某商场把服装分为男装、女装、童装三类,这不符合互斥性原则,因为童装也有男女之分。

(三) 统计分组的种类

各种统计分组的方法归纳起来,可以划分为以下几种基本类型:

1. 简单分组和复合分组

按分组标志的个数不同,统计分组可分为简单分组和复合分组。

(1) 简单分组。简单分组是指按一个标志进行分组。例如,将学生按性别划分为男、女两组就属于简单分组。这种分组比较简单,只能反映现象总体在某一标志特征方面的差异情况。

(2) 复合分组。复合分组是指对同一个总体在同一张分组表上,把两个或两个以上标志层叠起来进行分组。具体地说,复合分组是先按一个标志对总体进行分组,然后在分好的各组中再按另外的标志进行分组。例如,将某企业员工先分为男、女两组,再分别按文盲及半文盲、小学、初中、高中、大专、本科、硕士及以上分组。

2. 品质标志分组和变量标志分组

按分组标志的性质不同,统计分组可分为品质标志分组和变量标志分组。

(1) 品质标志分组。品质标志分组相对来讲比较简单,只有个别特殊情况需要在整理设计时统一做出规定。例如,调查人员按性别对某班学生进行分组,如表 2-3 所示。

表 2-3 某班学生按性别分组

性 别	人数/人	百分比/%
男生	60	75
女生	20	25
合计	80	100

(2) 变量标志分组。变量标志分组即按数量标志分组,如按身高、考试成绩、收入等进行分组,如表 2-4 所示。这种分组方法相对来讲比较复杂,也是人们常用的分类方法。其与品质标志分组最大的不同在于主观性较强,如分组组数的确定与调查者有直接关系。

表 2-4 某班学生按成绩分组

成 绩	人数/人	百分比/%
优秀	10	12.5
良好	20	25
中等	30	37.5
及格	10	12.5
不及格	10	12.5
合计	80	100

三、数值数据的整理与显示

按数量标志分组得到的统计数据均属于数值型数据。调查人员根据分析数据的目的和数据的特征对数值型数据进行分组的方法有单变量值分组和组距分组两种。

(一) 单变量值分组

单变量值分组又称单项式分组,是指在变量分组时将一个变量值设为一个组。例如,某班学生按年龄分组的结果如表 2-5 所示。

表 2-5 某班学生按年龄分组的结果

年龄/岁	人数/人
17	10
18	20
19	20
20	20
21	10
合计	80

单项式分组多用于离散型变量,且在变量值不多、变动范围有限的条件下采用。若离散型变量的变动范围较大,单位数又很多,则采用单项式分组必然会使分组的组数过多,各组频数过于分散,因而不能反映出总体内部真正的结构状况,从而使得统计分组失去意义。

(二) 组距分组

在变量值的变动范围内,将其划分为若干个数值区间,将每一个数值区间内的所有变量值归为一组,组内变量值的性质相同,但组与组之间的性质相异,这样的分组即组距分组。与单项式分组相比,组距分组的各组变量值不是某一具体的点值,而是一个数值区间。例如,某班学生按成绩分组的结果如表 2-6 所示。

表 2-6 某班学生按成绩分组的结果

成绩/分	人数/人
$[0,60)$	3
$[60,70)$	15

续表

成绩/分	人数/人
[70,80)	30
[80,90)	22
[90,100)	10
合计	80

【例 2-1】 2015 年 5 月,某电器公司的 50 名营销员的业绩(单位:万元)如下:

50 73 86 51 73 87 52 73 87 54
 74 87 55 75 88 55 75 90 56 75
 94 60 76 94 61 76 95 64 78 95
 65 79 97 65 80 98 65 80 67 80
 68 81 69 82 70 84 70 85 72 85

要求:编制组距数列并说明数据的分布特征。

解:第一步,对各变量值进行排序,确定最大值、最小值,并计算全距。全距是总体中某个数量标志的最大值与最小值之差,用符号 R 表示。全距的计算公式如下:

$$R = \max x - \min x$$

式中, x 为某个数量标志。

从原始资料来看,各变量值比较分散零乱,不能直接看出其分布特征。为此,下面先将它们按从小到大的顺序排列:

50 51 52 54 55 55 56 60 61 64
 65 65 65 67 68 69 70 70 72 73
 73 73 74 75 75 75 76 76 78 79
 80 80 80 81 82 84 85 85 86 87
 87 87 88 90 94 94 95 95 97 98

经过排序可以看出,变量的最大值是 98 万元,最小值是 50 万元,即该电器公司 50 名营销员业绩的变动幅度为 50 万~98 万元,由此可计算出全距:

$$R = 98 - 50 = 48(\text{万元})$$

计算全距是为了确保在编制变量数列时,总体中的每一个单位都有一个组可以分配而不被遗漏。

第二步,分析数据的特点,确定分组形式。从营销员业绩这一标志的性质来看,它属于连续型变量。由于变量值很多,若编制单项数列,则不便于观察数据的分布特征及变动规律。因此,这一标志应当采用组距分组,编制组距数列。

在组距数列中,是采用等距分组还是采用异距分组,要视变量的分布特征而定。如果变量的分布反映出的现象性质差异的变动没有较明显的跳跃性,可以考虑采用等距分组。等距分组在编制变量数列和进行统计分析时有许多优点,它便于直接对比各组的单位数,显示出某一标志值在总体中的分配情况,也便于计算各项综合指标,简化计算过程。如果变量的分布反映出的现象性质差异的变动并不均衡,很难用等距的办法近似地反映出性质不同的组,但具有较明显的跳跃性,表现出忽大忽小的差异性,就可以考虑采用异距分组,以充分反

映变量的分布特征,保证资料的真实性。从本例变量的排序情况来看,它基本上是均匀分布的,无较大的跳跃性,所以应该用等距分组来编制组距数列。

第三步,确定组数和组距。编制组距数列的一个明显特点就是在各组中将各个组内变量的差异抽象化。这一方面扩大了组与组之间的差异,突出了变量的集中趋势和结构特点;另一方面却掩盖了各组内各单位变量值的具体表现,对各组内变量只能假设其均匀分布。因此,调查人员在编制组距数列时,不能随意地扩大组距,以避免把性质不同的单位归在一组;同时,不能分组过细,这样不易看出总体的分布规律。组距越大,所分的组数越少;组数越多,组距就越小。在确定组距和组数时,没有一定的规律可循,必须根据统计研究的实际需要和数据的变化特点灵活地加以掌握。但是,在一般情况下,组数不应少于5组,也不应多于15组。

组距和组数的确定没有顺序规定,既可以先根据数据的变化特征确定组距,也可以在事先对研究对象的性质比较了解的情况下先确定组数,然后确定组距,但组数必须是整数。

(1) 确定组数。组数即分组的个数。确定组数需要考虑的因素很多,其中最根本的就是要体现出各组间质的不同,除此以外还要考虑总体单位数的多少、全距的大小,但最通用的一般方法是借助美国学者斯特吉斯提出的经验公式确定组数。斯特吉斯认为当总体单位按某标志分组接近正态分布时,可根据总体单位数 N 确定组数 K 。相关计算公式如下:

$$K=1+\frac{\lg N}{\lg 2}$$

在此,我们不妨采用斯特吉斯的经验公式来确定组数,即 $K=1+5.6439=6.6439$,组数取整数为7。

(2) 确定组距。

① 组距。每一个数值区间的区间长度即该组的组距。其计算公式如下:

$$\text{组距}=\text{上限}-\text{下限}$$

确定组距需要考虑的因素很多,其中最根本的就是要体现出各组内的同质性,除此以外还要考虑研究目的、数列分布的偏斜程度等,但最通用的一般方法仍是在美国学者斯特吉斯提出的经验公式的基础上进行测算。由于在一个数列中有关系式 $R=KI$,故有如下公式:

$$I=\frac{R}{K}$$

式中, I 为组距。

对于本例, $I=\frac{48}{7}=6.8571$ 。

调整组距为整数7,有时为了方便起见,组距 I 一般只取5或10的整数倍。因此,考虑到分数分析的习惯,此处可以取组距为10。但是,无论怎样确定都要满足“ $IK \geq R$ ”这一条件。此时,组数则相应调整如下:

$$K=\frac{R}{I}=\frac{48}{10}=4.8$$

由上可知,组数调整为5组。

② 组中值。组中值是指每组的分组变量的中间值。在一定的条件下,组中值用来代表每组各单位在分组变量上的一般水平(平均数)。其计算公式如下:

$$\text{组中值}=\frac{\text{下限}+\text{上限}}{2}$$

第四步,确定组限。组限是组距分组中各组变量值变化的界限,即每组两端的界限。其中,最小的变量值称为下限,最大的变量值称为上限。每一组的上限和下限决定了组中变量的变化范围。组限的合理确定至少要考虑以下几个因素:

(1) 根据极端值的存在与否确定首、末组是采用开口组还是闭口组。闭口组是指上限、下限齐全的组,如“ $[50,60), [60,70), [70,80), [80,90), [90,100)$ ”这样的营销员业绩的分组。开口组是指最小组下限或最大组上限不能完全确定,通常用“ $\times\times$ 以上”或“ $\times\times$ 以下”表示。如果数据分布较集中,不存在极端值,那么首、末组适宜采用闭口组;如果数据分布较分散,并且变量值有特大或特小的极端值,那么首、末组适宜采用开口组。

对于开口组,有时由于计算的需要,还要将其所缺的上、下限补齐。其处理方法为若数列按从小到大的顺序排列,则开口组假设组限的确定公式如下:

$$\text{首组的假设下限} = \text{首组上限} - \text{邻组组距}$$

$$\text{末组的假设上限} = \text{末组下限} + \text{邻组组距}$$

假设营销员业绩分组为“ $[0,60), [60,70), [70,80), [80,90), [90,+\infty)$ ”,则有:

$$\text{首组的假设下限} = 60 - 10 = 50(\text{万元})$$

$$\text{末组的假设上限} = 90 + 10 = 100(\text{万元})$$

但是,有时这种方法也会出现问题。例如,西瓜按重量分为“ $[0,2), [2,6), [6,+\infty)$ ”三组,则有:

$$\text{首组的假设下限} = 2 - 4 = -2(\text{千克})$$

$$\text{末组的假设上限} = 6 + 4 = 10(\text{千克})$$

显然,首组的假设下限为负值,这与事实不符。这时应将“-2”改成“0”,因为西瓜的重量不可能为负数。一般地,当计算出的假设上、下限超过实际可能的极限时,应当将其固定在极限值上。

对于此例,由于不存在极端值,因此首、末组不必采用开口组,而应采用闭口组。

(2) 根据变量的性质确定重叠组限与不重叠组限。离散型变量与连续型变量在分组时,组限的确定方法是不同的。离散型变量可以一一列举,而且相邻两个数值之间没有中间数值。因此,对离散型变量分组,各组的上、下限都可以用确定的数值(整数)表示,相邻两组的组限是不重叠的,即所谓的不重叠组限。不重叠组限组的组中值的计算公式如下:

$$\text{不重叠组限组的组中值} = \frac{\text{本组下限} + \text{后一组下限}}{2}$$

连续型变量在两变量值之间可能有无限多个中间值,不可能一一列举,分组时为了遵循穷尽性原则,相邻组的上限与下限不能用两个确定的值分别表示。这样,只能采用相邻组的组限重叠的方式,使上一组的上限同时也是下一组的下限,即所谓的重叠组限。

在实际工作中为方便起见,在离散型变量分组中也可以采用重叠组限,这样可使分组简明且方便计算组中值。对于本例而言,由于营销员业绩是连续型变量,故必须采用重叠组限。

(3) 根据能否反映数据的真实分布特征确定初始值。在确定组限时,初始值的确定有两种基本方法:一种是以小于或等于最小变量值作为最小组的下限,另一种是以最小变量值作为最小组的组中值。但是,这两种方法一般都要做到最小组的下限低于或等于最小的变量值,最大组的上限高于或等于最大的变量值,以满足穷尽性的分组原则。一般来说,分析

人员可同时得出几种结果并从中选出最能反映数列的真实分布特征的初始值。为此,分析人员在等距分组时,在先确定组数,再确定组距的近似计算中要注意不能舍去后面的小数,而应将小数进位。组距最初取 7 就是这个道理。

本例中,不妨选取 50 作为最小组的下限,则组限确定的情形为“ $[50, 60)$, $[60, 70)$, $[70, 80)$, $[80, 90)$, $[90 \sim 100)$ ”。

第五步,计算各组的频数和频率。在组数已经确定,各组组限划分完毕后,就可以根据实际资料计算各组应包含的变量值的个数,即频数或次数,并据此计算各组的频率。

统计分组后各组对应的单位数称为频数,也称次数;各组单位数占总体单位数的比重称为频率。各组的频率大于 0,所有组的频率总和等于 1。

在将各变量值划归各组计算频数时,如果连续性变量中的变量值正好是某一组的界限,由于采用重叠组限,那么为了满足互斥性原则,通常将其划为作为下限的那一组内,即所谓的“上限不在内”原则。例如,变量值“60”就应划归“ $[60, 70)$ ”组内、“90”就应划归“ $[90, 100)$ ”组内。但是,有的变量是越小越好,这时还可以按“下限不在内”的原则归组。

在变量分配数列中,频数(频率)表明对应组标志值的作用程度。频数(频率)越大,表明该标志值对总体水平所起的作用越大;反之,频数(频率)越小,表明该组标志值对总体水平所起的作用越小。

根据计算出的本例各组的频数和频率,编制该电器公司营销员业绩的分布表,如表 2-7 所示。

表 2-7 某电器公司 50 名营销员的业绩分布表

业绩/万元	频数/人	频率/%
$[50, 60)$	7	14
$[60, 70)$	9	18
$[70, 80)$	14	28
$[80, 90)$	13	26
$[90, 100)$	7	14
合计	50	100

从表 2-7 中可以看出,业绩大于等于 70 万元且小于 80 万元的营销员居多。可见,营销员的业绩分布有明显的集中趋势。

第六步,计算累计频数和累计频率。为了更详细地认识变量的分布特征,分析人员还可以计算累计频数和累计频率,编制累计频数数列和累计频率数列。累计频数和累计频率均有向上累计和向下累计两种。

以变量值为依据,由变量值小的组向变量值大的组累计频数(频率)称为向上累计频数(频率)。向上累计数的意义是小于该组上限的各组的频数(频率)之和。相反,由变量值大的组向变量值小的组累计各组的频数(频率)称为向下累计频数(频率)。向下累计数的意义是大于及等于该组下限的各组的频数(频率)之和。

本例中,根据营销员的业绩编制的向上累计频数(频率)和向下累计频数(频率)分布表如表 2-8 所示。

表 2-8 某电器公司营销员业绩累计频数(频率)分布表

业绩/万元	频数/人	频率/%	向上累计		向下累计	
			频数/人	频率/%	频数/人	频率/%
[50,60)	7	14	7	14	50	100
[60,70)	9	18	16	32	43	86
[70,80)	14	28	30	60	34	68
[80,90)	13	26	43	86	20	40
[90,100)	7	14	50	100	7	14
合计	50	100				

在表 2-9 中,向上累计频数(频率)数据表明,营销员的业绩在 80 万元以下的有 30 人,占全部营销员的 60%;向下累计频数(频率)数据表明,业绩在 70 万元以上的营销员共有 34 人,占全部营销员的 68%。

经过以上步骤,一个完整的组距数列就形成了,由此可以看出该公司 50 名营销员的业绩分布呈“两头小、中间大”的钟形分布状态,这也是统计分布的一种常态。

四、统计图

(一) 条形图

条形图是用宽度相同的条形的高低或长短来表示数据大小或变动的图形。条形图可以横置或纵置。纵置的条形图也称为柱形图。条形图有单式条形图和复式条形图之分。

1. 单式条形图

单式条形图是以若干间距相等的单一条形的高低或长短来表示数据大小或变动的一种条形图。

2. 复式条形图

复式条形图是以两个或两个以上的条形为一组来进行相互比较的一种条形图。

(二) 圆形图

圆形图也称饼图,是用圆形及圆内扇形的面积来表示统计数据的一种图形。圆形图主要用于表示总体各组成部分在总体中所占的比例,以反映现象总体的内部结构及其变化,对于研究结构问题十分有用。

在绘制圆形图时,每个圆形的面积代表 100%,各组成部分在总体中所占的比例用圆内的各个扇形的面积表示,按各组成部分在总体中所占的比例分别乘以 360° 来确定这些扇形的中心角度,据以分割圆内总面积。

(三) 环形图

环形图与圆形图有一些类似的地方,但又有区别。例如,环形图中间有一个“空洞”,总体中每一个组成部分的数据是用环中的一段表示的。另外,圆形图只能表示一个总体各组成部分在总体中所占的比例,而环形图则可以同时绘制多个总体的数据系列,每一个总体的数据系列为一个环。因此,环形图可以显示多个总体的各个组成部分在总体中所占的相应比例,从而有利于进行比较研究。

(四) 直方图

直方图是用矩形的宽度和高度来表示频数(频率)分布的图形。在平面直角坐标系中,通常用横轴表示数据分组,纵轴表示频数(频率),这样各组与相应的频数(频率)就形成了一个矩形,即直方图。

例如,根据【例 2-1】中某电器公司 50 名营销员的业绩分布表(见表 2-8)绘制的直方图如图 2-1 所示。

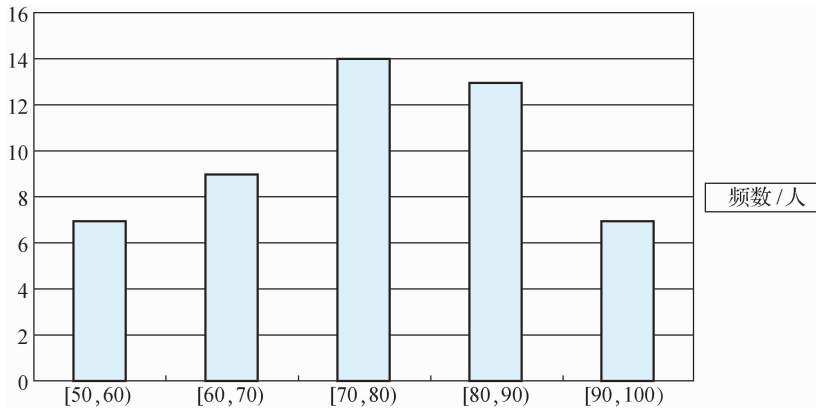


图 2-1 某电器公司营销员业绩分布直方图

当遇到异距数列时,由于各组次数的分布不仅受标志值大小的影响,还与各组组距大小有关。那么,用上述频数(频率)分布直方图就不能客观、准确地反映出频数(频率)分布的特征,这时就应以次数密度作为条形的高度来绘制直方图以表示其分布。次数密度又称频数密度,是指单位组距内分布的次数。次数密度的计算公式如下:

$$\text{次数密度} = \text{频数} \div \text{组距} \times 100\%$$

分析人员以标志值(变量)为横轴,以次数密度为纵轴,可以绘制出次数分布直方图。例如,根据某地区大中型项目投资额分布表(如表 2-9 所示)计算各组次数密度,绘制出频数密度分布直方图(如图 2-2 所示)。

表 2-9 某地区大中型项目投资额分布

投资额/亿元	组 距	次数/个	次数密度/%
[0,1)	1	120	120.00
[1,5)	4	387	96.75
[5,10)	5	60	12.00
[10,+∞)	15	52	3.47

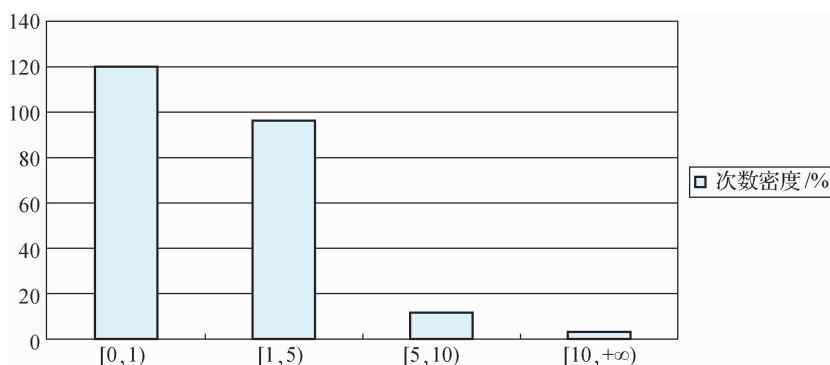


图 2-2 某地区大中型项目投资额分布直方图

(五) 折线图

折线图也称频数多边形图,是指在直方图的基础上把直方图顶部的中点(组中值)用直线连接起来,再把原来的直方图抹掉而形成的图形,如图 2-3 所示。观察的资料越多,组距越小且组数越多,折线图越光滑。

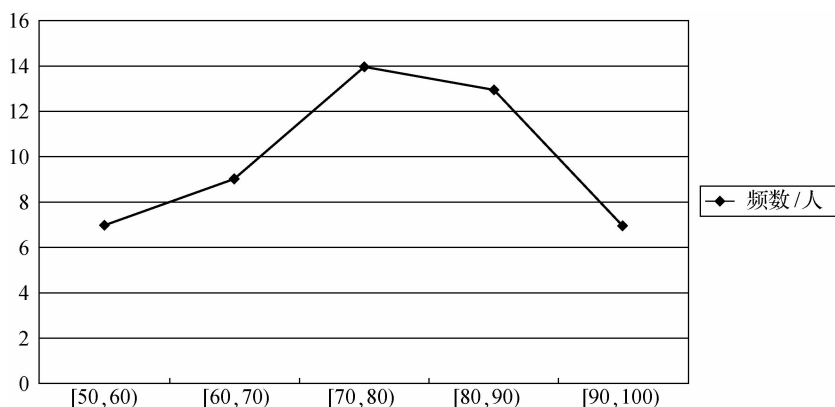


图 2-3 某电器公司营销员业绩的折线图

(六) 茎叶图

分析人员通过直方图可以看出一组数据的分布状况,但直方图没有给出具体的数值。而茎叶图既能给出数据的分布状况,又能给出每一个原始数值。茎叶图由“茎”和“叶”两部分构成,其图形是由数字组成的。分析人员通过茎叶图可以看出数据的分布形状及数据的离散状况,如分布是否对称、数据是否集中、是否有极端值等。

绘制茎叶图的关键是设计好树茎,通常是以该组数据中基本不变或变化不大的位数上的数值作为树茎。树茎一经确定,树叶就自然地长在相应的树茎上了。

下面以【例 2-1】的数据来绘制茎叶图,如图 2-4 所示。当数据很多时,可以将茎叶图进行扩展,如图 2-5 所示。

树茎	树叶	数据个数
5	0124556	7
6	014555789	9
7	00233345556689	14
8	0001245567778	13
9	0445578	7

图 2-4 某电器公司 50 名营销员业绩数据的茎叶图

树茎	树叶	数据个数
5*	0124	4
5.	556	3
6*	014	3
6.	555789	6
7*	0023334	7
7.	5556689	7
8*	000124	6
8.	5567778	7
9*	044	3
9.	5578	4

图 2-5 某电器公司 50 名营销员业绩数据的扩展茎叶图

茎叶图可以直观、形象、全面地反映出总体数据的次数分布状况,但是制作起来不是很方便,尤其是当总体的数据量比较大时更为不便。所以,在实际统计工作中,茎叶图的运用不太广泛。

(七) 曲线图

曲线图是指利用曲线的升降变化来显示所研究现象发展变化趋势的一种图形。它主要用于研究和分析社会经济现象的发展变化与社会经济现象之间的依存关系等方面内容。

由于统计图在表达数据资料方面更加形象、直观,因而被经常使用。但是,绘制统计图要以相应的统计表为基础。与统计表相比,统计图的种类较多。怎样选择合适的统计图是需要分析人员掌握的。具体来说,一方面要对每种统计图的特点有所了解,更要与具体的研究对象与目的要求相结合;另一方面要对所使用的统计数据类型有一定了解,这个工作甚至在统计设计阶段就应该考虑。不同的数据类型可以采用不同的统计图。条形图主要适用于分类数据,如按性别、民族等品质标志分组,而曲线图主要适用于时间序列数据。

在日常生活和经济管理中,常见的频数分布曲线主要有正态分布、偏态分布、J形分布和U形分布等类型,如图 2-6 所示。

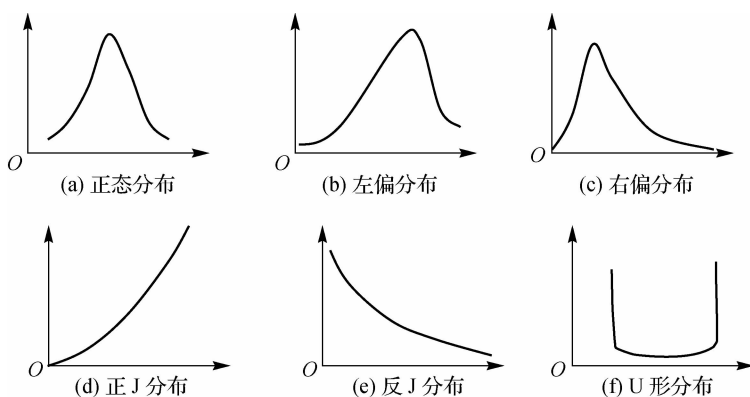


图 2-6 常见的频数分布曲线

1. 正态分布

正态分布是一种对称的钟形分布。在日常生活中,有很多现象服从这种分布,如人的身高、体重、智商,纤维长度,细纱强度,钢的含碳量,粮食作物的产量,橡胶的抗张力,等等。所有的试验数据、测量数据和观测误差数据等都服从正态分布。

2. 偏态分布

偏态分布根据“尾巴”拖向哪一方分为正偏(右偏)分布和负偏(左偏)分布两种。例如,人均收入分配的分布就为右偏分布,即低收入的人数较多,因而左边形成高峰,而高收入的人数较少,且收入越高的人越少,在右边形成一个细长的“尾巴”。

3. J形分布

J形分布有正J分布和反J分布两种。例如,经济学中的供给曲线,随着价格的提高,供给量以更快的速度增加,呈现正J形;而需求曲线则表现为随着价格的提高,需求量以较快的速度减少,呈现反J形。

4. U形分布

U形分布的特征是两端的频数分布多,中间的频数分布少。比如,人的死亡率分布就近似服从U形分布,因为婴幼儿和老年人的死亡率较高,而中青年的死亡率则较低。另外,产品的故障率也有类似的分布。

在上述统计图中,前五种图用 Excel 中的插入功能即可完成操作。

五、统计表

统计表是统计数据整理结果最常用的表示方法。从广义上来讲,统计表包括一切编制和表现统计资料的表格,是最常用、最规范的统计资料表达方式。

(一) 统计表的分类

按主词的分组程度划分,统计表可分为以下三类:

- (1) 简单表。它是指主词不进行任何分组,按习惯或自然顺序排列所形成的统计表。
- (2) 简单分组表。它是指将主词按一个分组标志进行简单分组的统计表。
- (3) 复合分组表。它是指将主词按两个或两个以上标志分组的统计表。

(二) 统计表的构成

1. 结构构成

从统计表的结构来看,统计表由以下五个部分构成:

(1) 总标题。它是统计表的名称。由于统计表中的汇总内容一般都是横向排列,所以总标题一般位于表的上端中央。总标题是对统计总体和指标内容的高度概括。

(2) 横行标题。它用于说明横行内容的名称或各个变量组,位于统计表的最左边。

(3) 纵栏标题。它是说明纵栏内容的名称,位于表的上方。

(4) 统计(数字)资料。它是统计表的主要内容,是纵栏标题和横行标题交叉形成的一个区域。

(5) 说明。规范的统计表还包括表头和表尾两部分说明性的内容。表头一般是对统计总体的性质、稳定性特征及统计表体系的程序化规范,一般位于统计表的左、右两肩上方。表尾主要是对统计资料的注释性补充,一般位于表的最下方。

统计表的结构如表 2-10 所示。

表 2-10 2016 年 1~4 月份规模以上工业企业主要财务指标

分 组	主营业务收入		利润总额	
	1~4 月 /亿元	同比增长 /%	1~4 月 /亿元	同比增长 /%
总计	335 613.8	2.3	18 442.2	6.5
其中:采矿业	13 910.9	-13.1	-40.3	-104.8
制造业	302 110.2	3.4	16 783.7	13.3
电力、热力、燃气及水生产和供应业	19 592.7	-1.6	1 698.8	1.7
国有控股企业	69 817.1	-5.5	3 265.5	-7.8
集体企业	2 041.1	-2.1	137.3	0.4
股份制企业	234 751.6	3.5	12 316.4	7.4
外商及港澳台商投资企业	76 138.1	-0.5	4 761.6	7.3
私营企业	117 489.3	5.9	6 626.1	8.4

注:经济类型分组之间存在交叉,故各经济类型企业数据之和大于总计。

2. 反映的内容构成

从统计表反映的内容来看,统计表由以下两部分构成:

(1) 主词。它是统计资料所属的统计总体、总体的各个组或各个单位的名称,一般位于表的左边。

(2) 宾词。它说明主词的各个指标名称及指标值,一般在主词的右边按一定顺序横向排列。

统计表通常按上述格式编制,当与版面不协调时,也可以将主词与宾词的位置互换。

(三) 统计表的设计

统计表是表现、承载统计资料的最广泛形式。其基本要求是科学、合理、清晰、方便。一般来说,统计表的设计包括以下几个方面内容:

1. 结构设计

统计表的各种标题、数字资料的位置应该合理安排。在版面允许的情况下,尽可能考虑视觉习惯,保持美观的外形。当然,这要根据总体大小与指标多少而定。有时,由于强调问题不同,行标题和列标题可以互换。

2. 表头设计

表头一般应包括表号、总标题和表中的数据单位等内容。总标题应简明确切地概括出统计表的内容,一般需要表明统计数据的时间、地点及何种数据。当表中数据资料的计量单位完全一样时,计量单位写在表的右肩上方。如果计量单位不完全一样,那么横行的计量单位可以专设一个计量单位栏目,纵栏的计量单位一般与纵栏标题(指标)写在一起。

3. 表格设计

在统计表中,上、下两条横线一般用粗线,中间的其他线用细线;左、右两侧应是开口式的,即左、右两端不画封线。

在主词分组和复合分组情况下,一般应先部分后总体,即先排列总体各个单位或各个组,最后一行是总(小)计。如果不是总体各个单位或各个组都要列出,而只是列出总体中的若干个单位或若干个组,那么一般是先写总计,后排列出各部分。

在宾词存在结构性分解的情况下,处理方法与主词的处理方法基本一致。当宾词的数字形式不同时,一般先列绝对数,后列相对数或平均数。

统计表中的数字资料部分不允许空格,当数字为“0”时,“0”要写出;当缺乏某项资料或因数据太小可以忽略不计时,一般用省略号“……”表示;不存在数据的表格画“—”表示。表中的数字资料应填写整齐、规范。

4. 其他设计

当统计表的栏目较多时,可以加以编号。主词部分一般是总体、各个单位、部分的文字名称,可以用甲、乙、丙等文字排序区分;宾词部分一般是指标名称,对应的数字资料可以用①、②、③等数字排序区分;各指标之间如果有对应的计算关系,还可以用数字序号把计算关系表示出来,如“④=②×③”。

若统计表的内容存在特殊情况,如个别指标的统计口径与整个栏目不一致、个别指标具体计算方法与整个栏目不一致、个别资料不完整等,则应在表后分别加以注释和说明。



思考练习 >>>>

一、单项选择题

1. 我国于2010年11月1日零时开始进行的第六次全国人口普查是()。
 - A. 一次性调查和非全面调查
 - B. 经常性调查和非全面调查
 - C. 一次性调查和全面调查

E. 揭示现象之间的依存关系

三、简答题

1. 统计调查方式有哪些? 它们有什么不同的特点和作用?
2. 如何选择统计分组标志?

四、计算题

1. 一家广告公司为研究广告市场的状况,在某城市随机抽取 200 人就广告问题进行了邮寄问卷调查。其中一个问题和备选项分别为“您比较关心下列哪一类广告?”“商品广告、服务广告、金融广告、房地产广告、招生招聘广告、其他广告”。经整理,对这些广告关注的人数分别为 112 人、51 人、9 人、16 人、10 人、2 人。

要求:(1) 编制关注不同类型广告的人数的次数分布数列。

(2) 使用 Excel 绘制该问题的直方图与圆形图。

2. 某班学生年龄(单位:岁)分布如下:

18	17	19	18	19	20	18	19	19	18
19	18	19	18	17	21	18	19	19	18
17	18	19	18	19	20	18	20	19	19
19	18	19	18	17	21	18	19	20	18
18	17	20	17	19	20	19	18	19	18

要求:(1) 以学生年龄为分组标志编制分布数列。

(2) 分别按向上和向下两种形式进行频数累计并列表说明。

3. 某班学生统计学考试成绩(单位:分)如下:

93	50	78	85	66	71	63	83	52	95	93	50
72	85	77	70	90	70	76	69	58	89	72	85
98	62	81	24	76	86	73	83	85	81	98	62
78	72	85	78	82	90	80	55	95	67	78	72
80	61	67	99	89	63	78	74	82	88	80	61

要求:(1) 根据资料使用 Excel 编制组距数列。

(2) 绘制频数分布的曲线图,据此分析学生统计学考试成绩的分布特点。

(3) 计算向上和向下两种累计的人数。

4. 某车间有甲、乙两个生产小组。这两个小组各有 100 名工人。该车间工人某月的生产量分布如表 2-11 所示。

表 2-11 某车间工人某月生产量分布表

月产量/件	工人人数/人	
	甲 组	乙 组
[30,40)	10	7
[40,50)	20	25
[50,60)	35	32
[60,70)	20	18
[70,80)	10	15
[80,90)	5	3

要求:(1) 根据上述数据,画出甲、乙两个生产小组生产量的条形图和环形图。
(2) 比较甲、乙两个生产小组生产量的分布特点。

案例分析

对汽车行业的调研

随着人们生活质量的提高和生活节奏的加快,人们对交通工具有了越来越高的要求,汽车已走进千家万户,自行车逐渐变成人们锻炼身体的工具。2014年,我国汽车销量超2300万辆,同比增长6.86%。

本案例试图通过调查研究准确、客观、全面、真实地掌握某品牌汽车销售情况,了解未来几年汽车的销售趋势,从而为某品牌汽车厂商提供相关有价值的信息,对促进该品牌汽车的销售起到积极的作用。

要求:(1) 根据该案例的目标,请同学们自行组成5人小组,设计一个具体的调查方案和相应的调查问卷。

(2) 设计好调查问卷后进行实际调查(有效问卷在100份以上)。

(3) 将汽车价位按不同的阶段分组处理,并根据所在地区的收入水平分析哪个价位的汽车销量会比较好,并绘制相应的统计图。

第三章

描述统计分析

学习目标

- 了解总量指标的含义、作用、分类与计算方法；
- 了解相对指标的含义和数值表现形式；
- 掌握几种常用相对指标的计算方法；
- 了解平均指标和变异指标的基本知识；
- 掌握算术平均数、调和平均数、几何平均数、众数和中位数的计算方法；
- 掌握全距、四分位差、标准差和变异系数的计算方法；
- 了解数据分布形状的偏度和峰度。

引例

2015年国民经济和社会发展统计公报(节选)

初步核算,全年国内生产总值 676 708 亿元,比上年增长 6.9%。其中,第一产业增加值 60 863 亿元,增长 3.9%;第二产业增加值 274 278 亿元,增长 6.0%;第三产业增加值 341 567 亿元,增长 8.3%。第一产业增加值占国内生产总值的比重为 9.0%,第二产业增加值比重为 40.5%,第三产业增加值比重为 50.5%,首次突破 50%。全年人均国内生产总值 49 351 元,比上年增长 6.3%。全年国民总收入 673 021 亿元。

年末全国大陆总人口 137 462 万人,比上年末增加 680 万人。其中,城镇常住人口 77 116 万人,占总人口比重(常住人口城镇化率)为 56.10%,比上年末提高 1.33 个百分点。

年末全国就业人员 77 451 万人,其中城镇就业人员 40 410 万人。全年城镇新增就业 1 312 万人。年末城镇登记失业率为 4.05%。

全年粮食产量 62 144 万吨,比上年增加 1 441 万吨,增产 2.4%。其中,夏粮产量 14 112 万吨,增产 3.3%;早稻产量 3 369 万吨,减产 0.9%;秋粮产量 44 662 万吨,增产 2.3%。

对于这份报告中提到的数据,你了解它们是怎么得出来的吗?报告中提到的国内生产总值、各产业增加值的绝对数和相对数等反映的实际经济意义是什么呢?实际上,它们就是本章要学习的统计指标。学生在学完本章内容后,就会对这些问题有清晰的认识,并且可以对身边涉及的经济现象做出类似的统计和计算。

第一节 总量指标

一、总量指标的含义和作用

(一) 总量指标的含义

总量指标又称统计绝对数,是用绝对数形式表现的反映客观现象总体在一定时间、地点条件下的总规模、总水平或工作总量的统计指标。其数值大小随总体范围的大小而变化。

(二) 总量指标的作用

1. 总量指标是对社会经济现象总体认识的起点,常用来反映国情国力的基本状况

人们要想了解一个国家或一个地区的国民经济和社会发展状况,必须先准确地掌握客观现象在一定时间、地点条件下的发展规模或水平。例如,为了科学地指导国民经济和社会的协调发展,人们必须通过总量指标正确地反映社会再生产的基本条件和国民经济各部门的工作成果,即反映中国土地面积、人口和劳动资源、自然资源、国民财富、钢产量、工业总产值、粮食产量、农业总产值、国民收入及教育文化等方面的发展状况。

2. 总量指标是编制计划、实行经营管理的主要依据

一个国家或地区为更有效地指导经济建设,保持国民经济协调发展,就必须了解和分析各部门之间的经济关系。它虽然在一定程度上可以用相对数、平均数来反映,但归根到底还是需要用各部门在各个不同时期的总量来反映。例如,国家“十二五”规划、国家“十三五”规划都是依据每年各行业的综合生产情况制定的。

3. 总量指标是计算相对指标和平均指标的基础

总量指标是统计整理汇总后首先得到的能说明具体社会经济总量的综合性数字,是最基本的统计指标。相对指标和平均指标一般都是由两个有联系的总量指标对比而计算出来的。它们是总量指标的派生指标。总量指标计算是否科学、合理、准确,将直接关系到相对指标和平均指标的计算是否准确。

二、总量指标的分类

(一) 总体单位总量指标和总体标志总量指标

按反映总体内容的不同,总量指标可分为总体单位总量指标和总体标志总量指标。

1. 总体单位总量指标

总体单位总量指标是表明总体单位数量的总量指标。例如,全国零售商店的总数、全国高等院校的总数等。

2. 总体标志总量指标

总体标志总量指标是表明总体单位某一数量标志值总和的总量指标。例如,研究某地区工业企业的生产经营情况,那么该地区工业企业的利润总额、工资总额等就是总体标志总量指标。

总体单位是总体标志的直接承担者,标志总量不会独立于单位总量而存在。在一个特定的总体内,只存在一个单位总量,而同时存在多个标志总量。同一总量指标在不同情况下可有不同的性质。例如,对各企业工人总数指标来说,当研究企业平均规模时,以企业为总体单位,企业总数为单位总量,各企业工人总数为标志总量;当研究企业劳动效益时,以工人为总体单位,各企业工人总数为单位总量,这时企业的总产量就成为标志总量。所以说,总体单位总量和总体标志总量并不是固定不变的,两者随研究目的的不同而变化。

(二) 时期指标和时点指标

按反映时间状况的不同,总量指标可分为时期指标和时点指标。

1. 时期指标

时期指标是反映某种社会经济现象在一段时间发展变化结果的总量指标。例如,国内生产总值、工业增加值、人口出生数等。

2. 时点指标

时点指标是反映社会经济现象在某一时间(瞬间)状况上的总量指标。例如,年末人口总数、年末居民储蓄存款余额等。

3. 时期指标与时点指标的区别

时期指标与时点指标的区别如下:

- (1) 时期指标的数值可连续统计;时点指标的数值不能连续统计,只能间断统计。
- (2) 时期指标的各项数值可直接相加,相加后表示现象在更长时间内的的发展变化总量;时点指标的各项数值一般不能直接相加,相加后无意义,会出现同一单位或标志值在不同时点的重复计算。
- (3) 时期指标的数值大小与其所包括的时期长短直接相关;时点指标的数值大小与其所间隔时间的长短无直接关系。

(三) 实物指标、价值指标和劳动量指标

按所采用计量单位的不同,总量指标可分为实物指标、价值指标和劳动量指标。

1. 实物指标

实物指标是以实物单位计量的统计指标。实物单位是根据事物的自然属性和特点采用的自

然、物理计量单位。它包括自然单位、度量衡单位、双重单位、复合单位和标准实物计量单位等。

(1) 自然单位。它是根据事物的自然属性来计量的单位。例如,人口以“人”为单位,汽车以“辆”为单位,鞋以“双”为单位。

(2) 度量衡单位。它是按统一的度量衡制度来计量的单位。例如,钢产量以“吨”为单位,布以“米”为单位,距离以“千米”为单位。

(3) 双重单位。它是采用两种或多种计量单位来表明事物的数量。例如,电动机以“台”“千瓦”计量,船舶以“艘”“马力”“吨位”计量。

(4) 复合单位。它是两种计量单位结合在一起的计量单位。例如,发电量以“千瓦时”计量,货物周转量以“吨千米”计量。

(5) 标准实物计量单位。它是对同类实物产品按统一标准折合的单位。例如,将含热量不同的煤产量统一折算为7 000千卡的标准煤。

实物指标的最大特点是直接反映产品的使用价值或现象的具体内容,因而能够具体地表明事物的规模和水平。它的局限性是指标的综合性比较差,不同的实物的内容、性质、计量单位均不同,无法进行汇总,因而无法反映国民经济的总规模或总的发展速度。

2. 价值指标

价值指标是以货币单位计量的统计指标,如国内生产总值以“元”为单位。

价值指标的最大特点是具有高度的综合性。它可以综合反映不同国家或地区、部门、企业生产不同产品的总成果。其局限性在于脱离了物质内容,比较抽象。价值指标只有和实物指标结合使用,才能充分发挥作用。

3. 劳动量指标

劳动量指标是以劳动单位计量的统计指标。劳动单位是反映劳动力资源及其利用状况所采用的一种复合计量单位,如工时、工日等。这种统计指标虽然不多,但经常遇到。例如,工厂考核职工出勤情况,每天要登记出勤人数,对一个月的出勤人数进行汇总就不能用“人”而应用“工日”来计量。又如,工厂实行计件工资制,要对每个零部件在每道工序上都规定劳动定额。假设工厂规定某零件1小时要生产60件,则1件就是1定额工分。某工人一天生产600件,即生产的产品为600定额工分,即10个定额工时。由于各企业的定额水平不同,因此劳动量指标不适宜在各企业之间进行汇总,往往只限于企业内部的业务核算。

三、总量指标的计算方法

(一) 直接计算法

统计报表或普查中的总量资料基本上都是用直接计算法计算出来的。例如,在统计某企业的职工人数时,只需要把该企业的每个在册职工相加即可;而在统计该企业的本月总产值时,只需要把该企业本月每一天的产量相加即可。这是总量指标的主要计算方法。

(二) 间接推算法

间接推算法是采用社会经济现象之间的平衡关系、因果关系、比例关系,或利用非全面调查资料进行推算总量的方法。目前常用的间接推算法主要有以下几种:

1. 平衡推算法

平衡推算法是依据现象之间的平衡关系式,并根据已知指标来推算未知指标的计算方

法。例如,企业在报告期末推算产品库存量时经常采用如下关系式进行推算:

$$\text{期末产品库存量} = \text{期初产品库存量} + \text{本期产品生产量} - \text{本期产品销售量}$$

2. 因素关系推算法

在社会经济现象中,当一个指标可以分解为两个或两个以上因素的乘积时,就可以根据它们的关系推算出关系式中的未知指标。以下关系式便是如此:

$$\text{工业总产值} = \text{产品产量} \times \text{出厂价格}$$

$$\text{总成本} = \text{产品产量} \times \text{单位成本}$$

3. 比例关系推算法

比例关系推算法是根据已知的比例关系对未来进行推算的计算方法。例如,已知某企业 2015 年完成的销售额为 2.8 亿元,该企业 2014 年的销售利润率为 5%。假设该企业 2015 年的销售利润率与 2014 年基本相同,那么该企业 2015 年的销售利润为 1 400 万元 ($28\,000 \times 5\%$)。

第二节 相对指标

一、相对指标的含义和数值表现形式

在统计中,数字的作用在于进行比较和分析。所谓“比较为统计之母”是非常有道理的,孤立的数字,不进行任何比较分析,是不能说明任何问题的。因此,人们要借助相对指标对事物进行判断、鉴别和比较。

(一) 相对指标的含义

相对指标又称统计相对数,是将两个性质相同或相关的指标数值通过对比求得商数或比例,并借此反映现象总体内部的结构、比例、发展状况等。也可以说,相对指标是两个有联系的现象数值的比率,可以用来反映现象的发展程度、结构、强度、普遍程度或比例关系。例如,产品合格率、人口出生率、人均钢产量等。相对指标的计算公式如下:

$$\text{相对指标} = \frac{\text{对比数}}{\text{基数}}$$

注意:相对指标首先表现为一个抽象化的数值,反映现象之间的相对程度;其次表现为数值的大小与研究总体范围的大小无直接联系。

(二) 相对指标的数值表现形式

相对指标的数值表现形式有两种,即有名数和无名数。

1. 有名数

有名数是在计算相对指标时,保持两个对比指标原来的计量单位。它主要在强度相对指标的计算中采用。例如,2012 年我国人均钢产量为 533 千克,2014 年我国人口密度为 142.5 人/平方千米,等等。

2. 无名数

无名数是一种抽象化了的数值。在计算相对指标时,若分子与分母的数值表现形式相

同,则其数值表现为无名数。无名数包括系数或倍数、成数、百分数和千分数。

(1) 系数或倍数。系数或倍数是指将对比基数抽象为 1 而计算出来的相对数。当分子数值大于分母数值很多时,就形成倍数;当分子数值与分母数值相差不大时,就形成系数。例如,2013 年的产量 \div 2000 年的产量 $=500$ 吨 \div 50 吨 $=10$,则说明 2013 年的产量是 2000 年的产量的 10 倍。

(2) 成数。成数是指将对比基数抽象为 10 而计算出来的相对数。例如,某地区粮食产量 2014 年比 2013 年增长 2 成,即增长 $2/10$ 。

(3) 百分数。百分数是指将对比基数抽象为 100 而计算出来的相对数。例如,2015 年产量 \div 2014 年产量 $\times 100\% = 500$ 吨 \div 400 吨 $\times 100\% = 125\%$ 。

(4) 千分数。千分数是指将对比基数抽象为 1 000 而计算出来的相对数。当分子数值小于分母数值很多时,以千分数表示。例如,2015 年年末,中国大陆总人口为 137 462 万人,比 2014 年年末增加 680 万人,全年出生人口为 1 655 万人。那么,人口出生率 $=$ 人口出生数 \div 年(平均)人口总数 $\times 1\ 000\% = 1\ 655 \div [(137\ 462 + 137\ 462 - 680) \div 2] \times 1\ 000\% = 12.07\%$ 。

此外,在统计中经常使用的还有百分点,即当两个以百分数表示的相对指标,在对比基数相同时,它们相减的结果差距相当于 1%,称为 1 个百分点。例如,2015 年我国第三产业增加值占国内生产总值的比重为 50.5%,2014 年我国第三产业增加值占国内生产总值的比重为 48.2%,那么就此指标而言,2015 年比 2014 年增加了 2.3 个百分点(50.5% $-$ 48.2%)。

二、相对指标的种类及其计算方法

(一) 结构相对指标

1. 结构相对指标的概念和计算方法

结构相对指标就是人们通常所说的“比重”,是在对总体分组的基础上,以总体总量为比较标准,求出各组总量占总体总量的比重来反映总体内部组成情况的综合指标。其计算结果一般是百分数。结构相对指标的计算公式如下:

$$\text{结构相对指标} = \frac{\text{各组总量}}{\text{总体总量}} \times 100\%$$

结构相对指标在经济研究中具有重要作用。人们计算结构相对指标能够反映总体内部结构和现象的类型特征。

【例 3-1】 国家统计局相关资料显示,2015 年我国国内生产总值为 676 708 亿元。其中,第一产业为 60 863 亿元,第二产业为 274 278 亿元,第三产业为 341 567 亿元。试计算各产业占国内生产总值的比重。

$$\text{解: 第一产业占国内生产总值的比重} = \frac{60\ 863}{676\ 708} \times 100\% = 8.99\%$$

$$\text{第二产业占国内生产总值的比重} = \frac{274\ 278}{676\ 708} \times 100\% = 40.53\%$$

$$\text{第三产业占国内生产总值的比重} = \frac{341\ 567}{676\ 708} \times 100\% = 50.48\%$$

2. 计算结构相对指标应注意的问题

(1) 计算条件是统计分组。

- (2) 各部分计算结果小于1,各部分比重之和等于1。
- (3) 分子与分母均为总量指标。
- (4) 分子与分母不能互换。

(二) 比例相对指标

1. 比例相对指标的概念和计算方法

比例相对指标是总体中不同部分数量对比的相对指标,可以用来分析总体范围内各个局部、各个分组之间的比例关系和协调平衡状况。比例相对指标的计算公式如下:

$$\text{比例相对指标} = \frac{\text{总体中某一部分数值}}{\text{总体中另一部分数值}} \times 100\%$$

【例 3-2】 某县有人口 350 000 人,其中男性 180 000 人,女性 170 000 人。试计算该县人口的性别比例。

解: 该县人口的性别比例 = $\frac{\text{男性人口数}}{\text{女性人口数}} \times 100\% = \frac{180\,000}{170\,000} \times 100\% = 105.88\%$

如果以女性人口数为基数 1,则男女性别比例约为 1.06 : 1。

2. 计算比例相对指标应注意的问题

- (1) 计算条件是统计分组。
- (2) 分子与分母一般是总量指标对比,但有时也可以用总体各部分的相对数或平均数对比。例如,男职工平均工资 ÷ 女职工平均工资 × 100% = 500 ÷ 400 × 100% = 125%。
- (3) 各部分之间的比例之和不等于 100%。
- (4) 分子与分母可以互换。例如,女性人口数 ÷ 男性人口数 × 100% = 170 000 ÷ 180 000 × 100% = 94.44%。

(三) 比较相对指标

1. 比较相对指标的概念和计算方法

比较相对指标是以不同单位的同类现象数量对比而确定的相对指标,说明某一同类现象在同一时间内各单位发展的不平衡程度,并表明同类事物在不同条件下的数量对比关系。比较相对指标的计算结果通常用百分数或倍数表示。比较相对指标的计算公式如下:

$$\text{比较相对指标} = \frac{\text{甲总体某指标值}}{\text{乙总体同类指标值}} \times 100\%$$

说明:此处的总体可以是国家、地区、部门或企业等。

【例 3-3】 2015 年,甲市的大米价格为 2 元/千克,乙市的大米价格为 2.5 元/千克。试计算两地大米价格的比较相对数。

解: 乙市的大米价格 ÷ 甲市的大米价格 × 100% = 2.5 ÷ 2 × 100% = 125%

甲市的大米价格 ÷ 乙市的大米价格 × 100% = 2 ÷ 2.5 × 100% = 80%

结果显示,乙市的大米价格是甲市的大米价格的 1.25 倍,甲市的大米价格是乙市的大米价格的 80%。

运用比较相对指标对不同国家、不同地区、不同单位的同类指标进行对比,有助于揭露矛盾、找出差距、挖掘潜力,促进事物进一步发展。

2. 计算比较相对指标应注意的问题

(1) 分子与分母更多采用相对数或平均数。因为总量指标的数值易受总体范围、生产条件等影响,一般不具有可比性。

(2) 分子与分母可以互换。

(四) 强度相对指标

1. 强度相对指标的概念和计算方法

强度相对指标是两个性质不同但有一定联系的总量指标之间的对比,用来表明某一现象在另一现象中发展的强度、密度和普遍程度。它和其他相对指标的根本不同之处在于它不是同类现象指标的对比。强度相对指标以双重计量单位表示,是一种复名数。

$$\text{强度相对指标} = \frac{\text{某个现象总量指标}}{\text{另一个有联系而性质不同的现象总量指标}}$$

【例 3-4】 相关资料显示,截至 2014 年年末,我国人口总数为 136 782 万人,试按 960 万平方千米土地面积计算我国的人口密度。

解: 我国的人口密度 = $\frac{136\,782}{960} = 142.48$ (人/平方千米)

强度相对指标应用十分广泛,既可以反映国民经济和社会发展的基本情况,也可以反映生产条件及公共设施的配备情况,还可以反映经济效益的情况。

2. 计算强度相对指标应注意的问题

(1) 分子与分母为两个性质不同而有联系的现金总量指标。需要说明的是,钢产量和猪的存栏头数这两个性质不同的总量指标在经济上没有联系,因而对比没有意义。

(2) 其数值表现形式多数为有名数,少数为无名数。例如,人口密度以“人/平方千米”为单位、人均钢产量以“千克/人”为单位、商业网密度以“个/千人”为单位等表现为有名数;而人口自然增长率、商品流通过费用率等表现为无名数。

(3) 某些强度相对指标,分子与分母可互换,形成其正、逆指标。举例如下:

每百元固定资产提供的产值 = 年工业产值 ÷ 年固定资产平均余额 (正指标)

每百元产值占用的固定资产 = 年固定资产平均余额 ÷ 年工业产值 (逆指标)

需要说明的是,有些强度相对指标使用人均字眼,如人均钢产量、人均粮食产量、人均教育经费等,但它们与后面要讲的平均指标在含义上是有区别的。因为强度相对指标是两个不同总体但有联系的总量指标的比值,而平均数是同一总体两个不同性质的指标的比值。例如,全班学生的平均成绩是全班学生的总成绩与全班学生的总人数的比值,是平均数而不是强度相对指标。

(五) 计划完成程度相对指标

1. 计划完成程度相对指标的概念和计算方法

计划完成程度相对指标是用来检查、监督计划执行情况的相对指标。它以现象在某一段时间内的实际完成数与计划数的对比来观察计划完成程度。计划完成程度相对指标的计算公式如下:

$$\text{计划完成程度相对指标} = \frac{\text{实际完成数}}{\text{计划任务数}} \times 100\%$$

【例 3-5】 若某企业 2015 年的计划产值为 800 万元,2015 年的实际产值为 1 000 万元。试计算该企业 2015 年产值计划完成程度相对指标。

解:该企业 2015 年产值计划完成程度相对指标 $= \frac{1\ 000}{800} \times 100\% = 125\%$

结论:该企业超额 25% 完成产值计划任务。

【例 3-6】 若某企业 2015 年计划某产品的单位成本为 20 元,实际该产品的单位成本为 18 元。试计算该企业 2015 年该产品单位成本计划完成程度相对指标。

解:2015 年该产品单位成本计划完成程度相对指标 $= \frac{18}{20} \times 100\% = 90\%$

结论:该企业超额 10% 完成单位成本降低的计划任务。

需要说明的是,当计划指标规定为增长率(产值、劳动生产率等)时,计算结果 $\geq 100\%$ 则表明完成或超额完成计划,计算结果 $< 100\%$ 则表明未完成计划;当计划指标规定为降低率(单位成本、单耗等)时,计算结果 $\leq 100\%$ 则表明完成或超额完成计划,计算结果 $> 100\%$ 则表明未完成计划。具体内容如下:

(1) 当计划指标是增长率时,计划完成程度相对指标的计算公式如下:

$$\text{计划完成程度相对指标} = \frac{1 + \text{实际增长率}}{1 + \text{计划增长率}} \times 100\%$$

【例 3-7】 某企业 2015 年计划规定产值要比 2014 年提高 10%,实际比 2014 年提高了 15%。试计算该企业产值计划完成程度相对指标。

解:该企业产值计划完成程度相对指标 $= \frac{1 + 15\%}{1 + 10\%} \times 100\% = 104.55\%$

结论:该企业产值计划完成 104.55%,即超额 4.55% 完成了计划。

(2) 当计划指标是降低率时,计划完成程度相对指标的计算公式如下:

$$\text{计划完成程度相对指标} = \frac{1 - \text{实际降低率}}{1 - \text{计划降低率}} \times 100\%$$

【例 3-8】 某企业 2015 年某产品单位成本计划比 2014 年降低 10%,实际比 2014 年降低 12%。试计算该产品单位成本计划完成程度相对指标。

解:该产品单位成本计划完成程度相对指标 $= \frac{1 - 12\%}{1 - 10\%} \times 100\% = \frac{88\%}{90\%} \times 100\% = 97.78\%$

结论:该产品单位成本实际比计划多降低 2.22% (100% - 97.78%),即超额完成成本降低任务。

2. 计划完成情况的检查

计划完成情况的检查分为短期计划检查和中长期计划检查两种。

(1) 短期计划检查。

① 实际完成数与计划任务数都是同一时期的,可按上述计划完成程度相对指标的公式计算。

② 计划期中某一段实际累计数与全期计划相比,从时间上考核计划执行的均衡性,以便及时发现问题、采取措施,保证计划的完成和超额完成的,可按计划执行进度的公式计算。具体公式如下:

$$\text{计划执行进度} = \frac{\text{累计至本期的实际完成数}}{\text{全期计划任务数}} \times 100\%$$

【例 3-9】 某企业 2015 年全年计划产值为 200 万元,第一季度实际产值为 40 万元,第二季度实际产值为 45 万元,第三季度实际产值为 60 万元。试计算该企业累计至第三季度的产值计划执行进度。

$$\text{解: 产值计划执行进度} = \frac{40+45+60}{200} \times 100\% = 72.5\%$$

结论:时间过去 3/4(75%),计划任务只完成 72.5%。

(2) 中长期计划检查。中长期计划检查的方法有两种,即累计法和水平法。

① 累计法。在检查中长期计划执行情况时,如果计划任务是按计划期各年总和规定的,那么要使用累计法。相关计算公式如下:

$$\text{计划完成程度} = \frac{\text{计划期间实际累计完成量}}{\text{计划规定累计量}} \times 100\%$$

【例 3-10】 某地区“十二五”期间计划基本建设投资总额为 20 亿元,五年实际累计完成 22 亿元。试计算五年基本建设投资总额计划完成程度。

$$\text{解: 五年基本建设投资总额计划完成程度} = \frac{22}{20} \times 100\% = 110\%$$

结论:该地区超额完成计划任务。

按累计法确定提前完成五年计划的时间,是用计划全部时间减去自计划执行日起到实际累计完成规定数量的日期止的时间。

② 水平法。在中长期计划中,如果只规定在整个计划期的末期(最后 1 年)现象应达到的水平,那么用水平法检查计划执行情况。相关计算公式如下:

$$\text{计划完成程度} = \frac{\text{计划期末实际达到的水平}}{\text{计划规定期末应达到的水平}} \times 100\%$$

【例 3-11】 我国“十二五”期间计划某种产品 2015 年的产量应达到 200 万吨,实际完成 260 万吨。试计算该产品产量五年计划完成程度。

$$\text{解: 该产品产量五年计划完成程度} = \frac{260}{200} \times 100\% = 130\%$$

结论:该产品产量超额 30%完成计划。

需要注意的是,按水平法检查中长期计划执行情况时,确定提前完成计划的时间是只要在计划期内有连续 1 年(可以跨日历年度的)数值达到计划规定最后 1 年的水平,即五年计划完成,则其余的时间即提前完成五年计划的时间。

【例 3-12】 某产品五年计划规定,最后 1 年的产量应达到 45 万吨。该产品五年计划执行情况如表 3-1 所示。

表 3-1 该产品五年计划执行情况

单位:万吨

时 间	第 一 年	第 二 年	第 三 年		第 四 年				第 五 年			
			上半年	下半年	一季度	二季度	三季度	四季度	一季度	二季度	三季度	四季度
产 量	30	30	17	19	10	10	11	12	12	13	15	16

解:从表 3-1 可以看出,在自第四年的二季度起到第五年的第一季度止的连续 1 年中,该产品的产量达到了计划所规定的水平,即 45 万吨(10+11+12+12)。这说明该产品提前

三个季度完成五年计划。

3. 计算计划完成程度相对指标应注意的问题

- (1) 分子与分母既可以是绝对数,也可以是相对数或平均数。
- (2) 分子与分母不能互换。

(六) 动态相对指标

动态相对指标又称发展速度,是某一事物报告期数值与基期数值对比的结果,用以说明事物在时间上发展的快慢程度。动态相对指标的计算公式如下:

$$\text{动态相对指标} = \frac{\text{报告期水平}}{\text{基期水平}} \times 100\%$$

通常,作为比较标准的时期称为基期,与基期对比的时期称为报告期。

【例 3-13】 某高校在校生人数 2005 年为 20 000 人,2015 年为 55 000 人。试计算该高校在校生人数的动态相对指标。

解: 该高校在校生人数的动态相对指标 $= \frac{55\,000}{20\,000} \times 100\% = 275\%$

结论:2005—2015 年的 10 年间,该高校在校生人数增加了 175%。

(七) 几种相对指标的区分

在掌握了几种常用的相对指标的概念、作用及计算后,分析人员要注意区分不同的相对指标。

1. 结构相对指标与比例相对指标的区别

结构相对指标是以总体总量为比较标准,通过计算各组总量占总体总量的比重来反映总体内部组成情况的综合指标。例如,各工种的工人占全部工人的比重。

比例相对指标是总体不同部分数量对比的相对数,用以分析总体范围内各个局部之间的比例关系和协调平衡状况。例如,轻重工业比例。

2. 比例相对指标与比较相对指标的区别

(1) 子项与母项的内容不同。比例相对指标是同一总体内,不同组成部分的指标数值的对比;比较相对指标是同一时间同类指标在空间上的对比。

(2) 说明问题不同。比例相对指标说明总体内部的比例关系,比较相对指标说明现象发展的不均衡程度。比较相对指标是不同单位的同类指标对比而确定的相对数,用以说明同类现象在同一时期内各单位发展的不平衡程度。例如,甲地职工平均收入是乙地职工平均收入的 1.3 倍。

3. 强度相对指标与其他相对指标的区别

(1) 其他各种相对指标都属于同一总体内的数量对比,而强度相对指标还可以是两种性质不同但又有联系的属于不同总体的总量指标之间的对比。

(2) 计算结果表现形式不同。其他相对指标用无名数表示,而强度相对指标主要用有名数表示。

(3) 当计算强度相对指标的分子、分母的位置互换后,会产生正指标和逆指标,而其他相对指标不存在正、逆指标之分。

三、相对指标的应用原则

(一) 正确选择对比标准的基数

如果基数的选择不准确,就无法使相对数正确地反映事物之间的数量对比关系。至于选择什么样的基数,必须从现象的性质特点出发,并根据研究目的来确定。例如,人们在计算教育普及程度时,如果用全国识字人口数除以全国人口数,那么显然是不合理的。因为全国人口数包括未达到学龄的人数,不能如实反映教育普及程度,所以必须从全国人口数中扣除6岁以下学龄前儿童人数。

(二) 保持相对指标的可比性

所谓相对指标的可比性,是指两个对比指标在所表明的经济内容、总体范围、计算方法、计量单位、时间长短等方面的一致性。如果违反可比性这一基本原则,计算相对指标就会失去实际意义,会得出不准确的结论。

(三) 相对指标和总量指标必须结合起来运用

在利用相对指标进行分析时,人们要考虑相对数背后所代表的绝对水平,即要将相对指标与总量指标结合起来应用。特别是在动态分析时,要注意到每增长1%增加的绝对值。

【例 3-14】 甲、乙两个钢铁企业的钢产量分别如表 3-2 所示。试计算甲、乙企业钢产量每增长 1% 增加的绝对值。

表 3-2 甲、乙企业的钢产量

单位:万吨

企业名称	2014 年	2015 年	增长量	增长速度/%
甲企业	800	1 000	200	25
乙企业	200	300	100	50

甲企业钢产量每增长 1% 增加的绝对值 = $200 \div 25\% \times 1\% = 8$ (万吨)

乙企业钢产量每增长 1% 增加的绝对值 = $100 \div 50\% \times 1\% = 2$ (万吨)

由此可见,高速度背后可能隐藏低水平,而低速度背后可能隐藏高水平。因此,分析问题时要既要看速度,又要看水平。

(四) 要将多种相对指标结合起来运用

为了从各方面分析和研究问题,分析人员需要把各种相对指标结合起来使用。因为不同的相对指标都是从某一侧面来说明现象之间的数量对比关系的,而现象之间及现象内部的相互联系往往是错综复杂的,所以,在实际统计工作中,只利用某一种相对指标对现象总体进行分析和研究是不能满足需要的,必须灵活运用各种相对指标,以便对所研究的现象总体有一个比较全面、正确的认识。

第三节 平均指标

平均指标又称统计平均数,是用以反映社会经济现象总体各单位某一数量标志在一定

时间、地点条件下所达到的一般水平的综合指标。

平均指标按计算方法不同可分为众数、中位数、分位数、算术平均数、调和平均数和几何平均数。前三个平均指标是根据标志值所处的位置确定的,又称为位置平均数;后三个平均指标是根据总体所有标志值计算的,又称为数值平均数。

一、位置平均数

(一) 众数

1. 众数的概念

众数就是分布数列中最常出现(频数或频率最大)的标志值,用 M_0 表示。某个标志值在数列中最常出现,说明该标志值最具有代表性,可以反映数列的一般水平。

2. 众数的特点和应用条件

(1) 众数的特点。众数是一种位置平均数,不受极端标志值或开口组的影响。所以,当总体出现极端标志值时,众数比算术平均数更能反映总体各单位标志值的一般水平。

(2) 众数的应用条件。在分配数列中,只有在标志值的次数有明显集中趋势的情况下才能确定众数。

3. 众数的计算

(1) 由单项式数列确定众数。

【例 3-15】 某集贸市场的某种商品价格及商户资料如表 3-3 所示,试确定其众数。

表 3-3 某集贸市场的某种商品价格及商户资料

商品价格/(元·千克 ⁻¹)	商户/户
1.5	1
1.6	4
1.8	15
2.2	3
2.4	2
合计	25

解: 商品价格 1.8 元/千克便是众数。

(2) 由组距数列确定众数。在由组距数列确定众数时,应先确定众数组,即次数最多的一组,而后运用下列公式计算众数的近似值:

① 下限公式。

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} d$$

② 上限公式。

$$M_0 = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} d$$

式中, L 为众数所在组的下限, U 为众数所在组的上限, Δ_1 为众数组频数与其前一组频数之差, Δ_2 为众数组频数与其后一组频数之差, d 为众数所在组的组距。

【例 3-16】 某乡农民家庭年人均纯收入资料如表 3-4 所示,试确定其众数。

表 3-4 某乡农民家庭年人均纯收入资料

农民家庭年人均纯收入/元	家庭数/户
[1 000,1 200)	240
[1 200,1 400)	480
[1 400,1 600)	1 050
[1 600,1 800)	600
[1 800,2 000)	270
[2 000,2 200)	210
[2 200,2 400)	120
[2 400,2 600)	30
合计	3 000

解:代入下限公式,该乡农民家庭年人均纯收入的众数计算如下:

$$M_0 = 1\ 400 + \frac{1\ 050 - 480}{(1\ 050 - 480) + (1\ 050 - 600)} \times 200 = 1\ 511.76(\text{元})$$

或代入上限公式,该乡农民家庭年人均纯收入的众数计算如下:

$$M_0 = 1\ 600 - \frac{1\ 050 - 600}{(1\ 050 - 480) + (1\ 050 - 600)} \times 200 = 1\ 511.76(\text{元})$$

(二) 中位数

1. 中位数的概念

将分布数列中各单位标志值依大小顺序排列,位于中间位置的单位标志值称为中位数,用 M_e 表示。中位数表明数列中有一半的单位标志值小于中位数,有一半的单位标志值大于中位数。

2. 中位数的特点和数学性质

(1) 中位数的特点。中位数也是一种位置平均数,不受极端标志值或开口组的影响。因为中位数的确定仅取决于它在数列中的位置,所以它不受少数极端标志值的影响。在这一点上,它优于算术平均数。

(2) 中位数的数学性质。总体各单位标志值与其中位数的绝对离差的总和是所有数值的绝对离差之和的最小值,即 $\sum_{i=1}^n |x_i - M_e| = \text{最小值}$ 。

3. 中位数的计算

(1) 由未分组数列确定中位数。 M_e 等于第 $\frac{n+1}{2}$ 个标志值。分析人员在确定中位数时要注意 n 为奇数和偶数的不同。

① 当 n 为奇数时,中位数取 M_e 位置对应的值。

【例 3-17】 某工厂有 9 名工人,每人日产零件数按从低到高的顺序排列如下:15,17,19,20,22,23,23,24,25。试确定该工厂工人日产零件数的中位数。

解: M_e 的位置 $= \frac{9+1}{2} = 5$

故中位数为第 5 个标志值,等于 22 件。该数字反映了该工厂工人总体日产零件数的一般水平。

② 当 n 为偶数时,中位数取 M_e 位置两侧值的平均数。

【例 3-18】 某工厂有 10 名工人,每人日产零件数按从低到高的顺序排列如下:15,17,19,20,22,23,23,24,25,28。试确定该工厂工人日产零件数的中位数。

解: M_e 的位置 $= \frac{10+1}{2} = 5.5$

故中位数处于第 5 个标志值和第 6 个标志值之间,等于 $22.5[(22+23) \div 2]$ 。该数字也反映了该工厂工人总体日产零件数的一般水平。

(2) 由分组数列确定中位数。

第一步,确定中位数所在组(采用向上或向下累计方法)。

第二步,根据下列公式确定中位数的近似值。

① 下限公式。

$$M_e = L + \frac{\frac{\sum f}{2} - S_{m-1}}{f_m} d$$

② 上限公式。

$$M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} d$$

式中, L 为中位数所在组的下限; U 为中位数所在组的上限; $\sum f$ 为数列的频数总和; $\frac{\sum f}{2}$ 为中位数的位次; f_m 为中位数所在组的频数; S_{m-1} 为在向上累计法中中位数所在组前一组的向上累计频数; S_{m+1} 为在向下累计法中中位数所在组后一组的向下累计频数; d 为中位数所在组的组距。

【例 3-19】 根据【例 3-16】的资料,计算该乡农民家庭年人均纯收入的中位数。

解: 调查该乡农民家庭年人均纯收入的家庭数累计表如表 3-5 所示。

表 3-5 该乡农民家庭年人均纯收入家庭数累计表

农民家庭年人均纯收入/元	家庭数/户	家庭数累计/户	
		向上累计	向下累计
[1 000,1 200)	240	240	3 000
[1 200,1 400)	480	720	2 760
[1 400,1 600)	1 050	1 770	2 280
[1 600,1 800)	600	2 370	1 230
[1 800,2 000)	270	2 640	630
[2 000,2 200)	210	2 850	360
[2 200,2 400)	120	2 970	150
[2 400,2 600)	30	3 000	30
合计	3 000		

由表 3-5 可知,中位数的位置为 1 500(3 000÷2),说明中位数应当在累计次数为 1 500 的组内,即在[1 400,1 600)的组内。

代入下限公式,该乡农民家庭年人均纯收入的中位数计算如下:

$$M_e = L + \frac{\sum f - S_{m-1}}{f_m} d = 1\,400 + \frac{3\,000 - 720}{1\,050} \times 200 = 1\,548.57(\text{元})$$

或代入上限公式,该乡农民家庭年人均纯收入的中位数计算如下:

$$M_e = U - \frac{\sum f - S_{m+1}}{f_m} d = 1\,600 - \frac{3\,000 - 1\,230}{1\,050} \times 200 = 1\,548.57(\text{元})$$

(三) 分位数

1. 分位数的概念

中位数是从中间点将全部数据等分成两部分。与中位数类似的还有四分位数、十分位数和百分位数等。它们分别是用 3 个点、9 个点和 99 个点将数据 4 等分、10 等分和 100 等分后各分位点上的值。这里只介绍四分位数。

四分位数也称四分位点,是一组数据排序后处于 25%、50%和 75%位置上的值。四分位数是通过 3 个点将全部数据等分为 4 部分,每部分均包含 25%的数据。显然,中间的四分位数就是中位数。因此,通常所说的四分位数是指处在 25%位置上的数值(下四分位数)和处在 75%位置上的数值(上四分位数)。

2. 分位数的计算

与中位数的计算方法类似,分析人员在根据未分组数据计算四分位数时,首先要对数据进行排序,然后确定四分位数所在的位置。该位置上的数值就是四分位数。

假设下四分位数为 Q_L ,上四分位数为 Q_U ,根据四分位数的定义, Q_L 与 Q_U 位置的确定公式如下:

$$Q_L \text{ 位置} = \frac{n}{4}$$

$$Q_U \text{ 位置} = \frac{3n}{4}$$

如果位置是整数,四分位数就是该位置对应的值;如果是在 50%的位置上,那么取该位置两侧值的平均数;如果是在 25%或 75%的位置上,那么四分位数等于该位置的下侧值加上按比例分摊位置两侧数值的差值。

【例 3-20】 在某城市中随机抽取 9 个家庭,调查得到每个家庭的人均月收入(单位:元)数据如下:1 500,750,780,1 080,850,960,2 000,1 250,1 630。试计算人均月收入的四分位数。

解: 先将相关数据从低到高排序,结果如下:

750 780 850 960 1 080 1 250 1 500 1 630 2 000

$$Q_L \text{ 位置} = \frac{n}{4} = \frac{9}{4} = 2.25$$

即 Q_L 在第 2 个数值 780 和第 3 个数值 850 之间 25%的位置上。

故 $Q_L = 780 + (850 - 780) \times 25\% = 797.5$ (元)。

$$Q_U \text{ 位置} = \frac{3n}{4} = \frac{3 \times 9}{4} = 6.75$$

即 Q_U 在第 6 个数值 1 250 和第 7 个数值 1 500 之间 75% 的位置上。

故 $Q_U = 1\,250 + (1\,500 - 1\,250) \times 75\% = 1\,437.5$ (元)。

根据四分位数的计算结果,人们可以近似地认为,在排序数据中,至少有 25% 的数据将小于或等于 Q_L ,而至少有 75% 的数据将大于或等于 Q_L ;至少有 75% 的数据将小于或等于 Q_U ,而至少有 25% 的数据将大于或等于 Q_U 。由于 Q_L 和 Q_U 之间包含了 50% 的数据,因此,可以说大约有一半的数据在 Q_L 和 Q_U 之间。

二、数值平均数

(一) 算术平均数

1. 算术平均数的概念

算术平均数是分布数列(总体)中各单位标志值的总和除以全部单位数。算术平均数是计算平均指标的最常用方法,基本公式如下:

$$\text{算术平均数} = \frac{\text{总体标志总量}}{\text{总体单位总量}}$$

【例 3-21】 某车间班组有 5 名工人。他们的月工资分别为 460 元、520 元、600 元、700 元和 850 元。试计算这 5 名工人的月平均工资。

$$\text{解: 工人的月平均工资} = \frac{460 + 520 + 600 + 700 + 850}{5} = \frac{3\,130}{5} = 626 \text{ (元)}$$

算术平均数是最常用的一种平均数,它的计算方法符合众多现象中“总体各单位标志值的算术和等于其总体标志总量”这一客观数量关系。

2. 算术平均数的计算

(1) 算术平均数的计算条件。基本公式的分子(总体标志总量)与分母(总体单位总量)必须是同一总体,并且分子与分母在数量上存在直接的对应关系,即分子的数值要随着分母数值的变动而变动。算术平均数的这一计算条件也是平均指标与强度相对指标的主要区别之一。需要指出的是,虽然平均指标与强度相对指标都是两个总量指标对比,并且有的强度相对指标还带有平均的含义,其计量单位也是双重单位,但两者仍有明显区别。

强度相对指标与平均指标的区别主要表现在以下两个方面:

① 指标的含义不同。强度相对指标说明的是某一现象在另一现象中发展的强度、密度或普遍程度,而平均指标说明的是现象发展的一般水平。

② 计算方法不同。强度相对指标与平均指标虽然都是两个有联系的总量指标之比,但是强度相对指标的分子与分母的联系只表现为一种经济关系,即分子与分母在数量上不存在直接的对应关系;而平均指标是在一个同质总体内标志总量和单位总量的比例关系,其分子与分母的联系是一种内在的联系,即分子与分母在数量上存在直接的对应关系。

(2) 算术平均数的计算形式。在实际工作中,由于掌握资料的不同,算术平均数有两种计算形式,即简单算术平均数和加权算术平均数。

① 简单算术平均数。简单算术平均数适用于计算未分组数列的平均数。其计算公式如下：

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

式中, x 为各单位标志值, n 为总体单位数。

注意:在简单算术平均数中,各单位标志值出现的次数(频数)均相同。例如,在【例 3-21】中,每个标志值出现的次数都是 1。

② 加权算术平均数。加权算术平均数适用于计算分组数列的平均数。其计算公式如下：

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i}$$

或

$$\bar{x} = \sum_{i=1}^n x_i \frac{f_i}{\sum_{i=1}^n f_i}$$

式中, \bar{x} 为加权算术平均数; x_i 为各组标志值; f_i 为各组单位数; $\frac{f_i}{\sum_{i=1}^n f_i}$ 为各组单位数比重。

a. 对单项式数列计算加权算术平均数。下面举例说明对单项式数列计算加权算术平均数。

【例 3-22】 某厂工人各等级月工资和相应工人人数资料如表 3-6 所示。试计算该厂工人的月平均工资。

表 3-6 某厂工人月工资情况

月工资额/元	工人数/人
460	5
520	15
600	18
700	10
850	2
合计	50

解:工人月平均工资计算表如表 3-7 所示。

表 3-7 某厂工人月平均工资计算表

月工资 x_i /元	工人数 f_i /人	月工资总额 $x_i f_i$ /元
460	5	2 300
520	15	7 800

续表

月工资 x_i /元	工人数 f_i /人	月工资总额 $x_i f_i$ /元
600	18	10 800
700	10	7 000
850	2	1 700
合计	50	29 600

该厂工人月平均工资 $\bar{x} = \frac{29\,600}{50} = 592$ (元)

b. 对组距数列计算加权算术平均数。对组距数列计算加权算术平均数,可用组中值代表各组变量值。需要强调的是,用组中值代表各组标志值具有假设性,即假设各组内部的标志值分布是均匀的。因此,计算的平均值只是近似值,而不是准确数值。

【例 3-23】 某公司对所属 6 个企业按生产某产品的平均单位成本的高低分组,各组产量占该公司总产量的比重的资料如表 3-8 所示。试计算该公司所属企业的平均单位成本。

表 3-8 某公司所属企业分组情况

平均单位成本/元	企业数	各组产量占总产量的比重/%
[10,12)	1	22
[12,14)	2	40
[14,18)	3	38
合计	6	100

该公司所属企业的平均单位成本的组中值如表 3-9 所示。

表 3-9 某公司所属企业的平均单位成本计算表

平均单位成本/元	企业数	组中值 x_i	各组产量占总产量的比重 $(f_i / \sum_{i=1}^n f_i) / \%$
[10,12)	1	11	22
[12,14)	2	13	40
[14,18)	3	16	38
合计	6		100

该公司所属企业的平均单位成本计算如下:

$$\bar{x} = 11 \times 0.22 + 13 \times 0.4 + 16 \times 0.38 = 13.7 \text{ (元)}$$

c. 计算加权算术平均数时应注意的问题。由加权算术平均数的计算公式可知,加权算术平均数的大小受两个因素的影响:一是各组标志值 x_i ,二是各组单位数 f_i 或各组单位数比重 $(f_i / \sum_{i=1}^n f_i)$ 。 f 为绝对数权数, $f_i / \sum_{i=1}^n f_i$ 为相对数权数。

当各组标志值已确定时,标志值分配的单位数越多的组,该组标志值对平均数的影响越大;反之,影响越小。也就是说,在一个数列中,当标志值较大的单位数居多时,平均数就会

趋近标志值大的一方；当标志值较小的单位数居多时，平均数就趋近标志值小的一方；当标志值较大的单位数与标志值较小的单位数基本平时，平均数居中。可见，各组标志值的单位数（频数）的多少对平均数的大小有权衡轻重的作用。因此，各组单位数称为权数，用权数乘以各组标志值称为加权，由此计算的平均数称为加权算术平均数。

【例 3-24】 甲、乙两个企业各级别月工资、相应的工人数及所占比重资料如表 3-10 所示。试计算甲、乙两个企业工人的月平均工资，并观察计算结果。

表 3-10 甲、乙两个企业工资情况表

月工资 x_i / 元	工人数 / 人		工人数所占比重 / %	
	甲企业 $f_{i甲}$	乙企业 $f_{i乙}$	甲企业 $f_{i甲} / \sum_{i=1}^n f_{i甲}$	乙企业 $f_{i乙} / \sum_{i=1}^n f_{i乙}$
460	5	20	10	10
520	15	60	30	30
600	18	72	36	36
700	10	40	20	20
850	2	8	4	4
合计	50	200	100	100

解：甲企业工人的月平均工资计算如下：

$$\bar{x}_{甲} = 460 \times 0.1 + 520 \times 0.3 + 600 \times 0.36 + 700 \times 0.2 + 850 \times 0.04 = 592(\text{元})$$

乙企业工人的月平均工资计算如下：

$$\bar{x}_{乙} = 460 \times 0.1 + 520 \times 0.3 + 600 \times 0.36 + 700 \times 0.2 + 850 \times 0.04 = 592(\text{元})$$

当各组单位数（频率）相等时，加权算术平均数等于简单算术平均数。在分组数列中，当各组标志值的单位数或各组单位数比重均相等时，权数就失去了权衡轻重的作用，这时用加权算术平均数计算的结果与用简单算术平均数计算的结果相同。

d. 关于加权算术平均数的权数选择原则如下：

$$\text{各组标志值} \times \text{各组单位数} = \text{各组标志总量}$$

用字母表示如下：

$$x \cdot f = xf$$

此等式必须有实际经济意义（三个量之间存在着客观的数量对等关系），各组单位数 f 才是加权算术平均数的合适权数。

$$\text{被平均的标志值 } x = \frac{\text{各组标志总量 } xf}{\text{各组单位数 } f}$$

例如，在【例 3-24】中，被平均的标志值 x （各组月工资）是绝对数。此时，工人数为合适的权数，其符合权数选择的原则。

【例 3-25】 某工业局所属企业产值计划完成程度、企业数和计划产值如表 3-11 所示。试计算该工业局所属企业的平均产值计划完成程度。

表 3-11 某工业局所属企业产值计划表

产值计划完成程度/%	企业数/个	计划产值/万元
[90,100)	5	100
[100,110)	8	800
[110,120)	2	100
合计	15	1 000

解:根据资料,被平均的标志值(各组产值计划完成程度)是相对数,那么以企业数为权数不符合权数选择原则。这是因为等式“各组产值计划完成程度×企业数=95%×5=475%”无意义。

正确的权数应为各组计划产值,符合权数选择的原则。这是因为等式“各组产值计划完成程度×各组计划产值=各组实际产值=95%×100=95(万元)”有意义。

根据“各组标志值×各组单位数=各组标志总量”,平均产值计划完成程度计算表如表 3-12 所示。

表 3-12 某工业局所属企业平均产值计划完成程度计算表

产值计划完成程度/%	企业数/个	组中值 x_i /%	计划产值 f_i /万元	实际产值 $x_i f_i$ /万元
[90,100)	5	95	100	95
[100,110)	8	105	800	840
[110,120)	2	115	100	115
合计	15		1 000	1 050

平均产值计划完成程度计算如下:

$$\bar{x} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} = \frac{0.95 \times 100 + 1.05 \times 800 + 1.15 \times 100}{100 + 800 + 100} \times 100\% = \frac{1\ 050}{1\ 000} \times 100\% = 105\%$$

由此可得出以下结论:当被平均的标志值是绝对数、相对数或平均数时,要将构成其绝对数、相对数或平均数的分母数值作为各组单位数,即权数;要将构成其绝对数、相对数或平均数的分子数值作为各组标志总量。

3. 算术平均数的数学性质和特点

(1) 算术平均数的数学性质。

① 各变量值与其平均数离差之和等于零。用算式表示如下:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^n (x_i - \bar{x}) f_i = 0$$

这一性质说明算术平均数用来代表总体各单位或分布数列的一般水平。它与总体各单位标志值的正负离差恰好相互抵消,从而使得离差总和等于零。也就是说,所有单位的标志值都用算术平均数代替之后再加起来,仍然等于原来的总体标志总量。

② 各变量值与其平均数离差的平方和是一个最小值,即 $\sum_{i=1}^n (x_i - \bar{x})^2 = \text{最小值}$ 或 $\sum_{i=1}^n (x_i - \bar{x})^2 < \sum_{i=1}^n (x_i - x_0)^2$ 。其中, $\bar{x} \neq x_0$ 。

这一性质说明各标志值与以任意非平均数的数值为中心计算的离差平方和总大于以平均数为中心计算的离差平方和。从公式可以看出,算术平均数是总体最好的代表值。

(2) 算术平均数的特点。算术平均数易受极端标志值(极大值或极小值)和开口组的影响。

(二) 调和平均数

1. 调和平均数的概念

调和平均数是分布数列中各单位标志值倒数的算术平均数的倒数,又称倒数平均数。假设有三个标志值,分别为 x_1 、 x_2 、 x_3 ,则算术平均数为

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

调和平均数为

$$H = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}} = \frac{3}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3}}$$

2. 调和平均数的计算

根据所掌握资料的不同,调和平均数的具体计算可分为简单调和平均数和加权调和平均数。

(1) 简单调和平均数。其计算公式如下:

$$H = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

(2) 加权调和平均数。其计算公式如下:

$$H = \frac{1}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \frac{m_3}{x_3} + \dots + \frac{m_n}{x_n}} = \frac{m_1 + m_2 + m_3 + \dots + m_n}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \frac{m_3}{x_3} + \dots + \frac{m_n}{x_n}} = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n \frac{m_i}{x_i}}$$

当 $m_1 = m_2 = m_3 = \dots = m_n = A$ 时,加权调和平均数等于简单调和平均数。其具体推导过程如下:

$$H = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n \frac{m_i}{x_i}} = \frac{An}{A \sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

3. 调和平均数和算术平均数的关系

令 $m_i = x_i f_i$, 则有:

$$H = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n \frac{m_i}{x_i}} = \frac{\sum_{i=1}^n x_i f_i}{\sum_{i=1}^n f_i} = \bar{x}$$

由此可见,在“ $m_i = x_i f_i$ ”的条件下,根据同一标志值,采用加权调和平均数计算平均指标与采用加权算术平均数计算平均指标的结果完全相同。这是因为两者均符合总体标志总量与总体单位总量的对比关系。所以,加权调和平均数是加权算术平均数的变形。两者的不同之处在于在计算平均指标时应用的权数资料不同,加权算术平均数是以各组单位数为权数,加权调和平均数是以各组标志总量为权数。

【例 3-26】 承【例 3-22】,已知某厂工人各级别的月工资及相对应的月工资总额资料,试计算该厂工人月平均工资。

解:该企业工人月平均工资计算表如表 3-13 所示。

表 3-13 某厂工人平均工资计算表(加权调和平均数)

月工资 x_i /元	月工资总 m_i /元 ($m_i = x_i f_i$)	工人数 f_i /人 ($f_i = m_i / x_i$)
460	2 300	5
520	7 800	15
600	10 800	18
700	7 000	10
850	1 700	2
合计	29 600	50

$$\begin{aligned} \text{该厂工人月平均工资 } H &= \frac{m_1 + m_2 + m_3 + \cdots + m_n}{\frac{m_1}{x_1} + \frac{m_2}{x_2} + \frac{m_3}{x_3} + \cdots + \frac{m_n}{x_n}} \\ &= \frac{2\,300 + 7\,800 + 10\,800 + 7\,000 + 1\,700}{\frac{2\,300}{460} + \frac{7\,800}{520} + \frac{10\,800}{600} + \frac{7\,000}{700} + \frac{1\,700}{850}} \\ &= \frac{2\,300 + 7\,800 + 10\,800 + 7\,000 + 1\,700}{5 + 15 + 18 + 10 + 2} \\ &= \frac{29\,600}{50} \\ &= 592(\text{元}) \end{aligned}$$

这与前面按加权算术平均数计算的结果完全相同。

【例 3-27】 某种蔬菜早晨、中午、晚上的价格及购买金额如表 3-14 所示。试计算该种蔬菜的平均购买价格。

表 3-14 某种蔬菜早晨、中午、晚上的价格及购买金额

时 间	价格/(元·千克 ⁻¹)	购买金额/元
早晨	0.125	2.5
中午	0.10	3
晚上	0.05	3.5
合计		9

解: 该种蔬菜平均购买价格计算过程如表 3-15 所示。

表 3-15 某种蔬菜平均购买价格计算表

时 间	价格 x_i /(元·千克 ⁻¹)	购买金额 m_i /元 ($m_i = x_i f_i$)	购买量 f_i /千克 ($f_i = m_i / x_i$)
早晨	0.125	2.5	20
中午	0.10	3	30
晚上	0.05	3.5	70
合计		9	120

该种蔬菜平均购买价格计算如下:

$$H = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n \frac{m_i}{x_i}} = \frac{2.5 + 3 + 3.5}{\frac{2.5}{0.125} + \frac{3}{0.1} + \frac{3.5}{0.05}} = \frac{9}{120} = 0.075(\text{元} / \text{千克})$$

4. 调和平均数的应用场合

(1) 在采用算术平均数计算平均指标时, 由于资料的限制, 当无法直接得到被平均标志值相对应的各组单位数时, 可通过调和平均数的形式求出所需的各组单位数。

(2) 在由相对数或平均数计算平均指标时, 如果掌握的权数资料是相对数或平均数的母项数值, 就应采用加权算术平均数计算。如果掌握的权数资料是相对数或平均数的子项数值, 就应采用加权调和平均数计算。

5. 调和平均数的特点

调和平均数易受极端标志值和开口组的影响。当数列中某项标志值为零时, 调和平均数无法计算得出。

(三) 几何平均数

1. 几何平均数的概念及计算

几何平均数是分布数列中 n 个单位标志值连乘积的 n 次方根。其计算公式如下:

$$G = \sqrt[n]{x_1 x_2 x_3 \cdots x_n} = \sqrt[n]{\prod x_i}$$

式中, G 为几何平均数; x_i 为用相对数表示的各单位标志值; n 为标志值的个数; \prod 为连乘符号。

几何平均数适合于计算现象的平均比率或平均速度。当变量值的连乘积等于总比率或

总速度时,分析人员宜用几何平均法进行计算。

【例 3-28】 某省 2010—2015 年钢产量如表 3-16 所示。试计算该省 2010—2015 年钢产量年平均发展速度。

表 3-16 某省 2010—2015 年钢产量

年份	2010 年	2011 年	2012 年	2013 年	2014 年	2015 年
钢产量/万吨	9 400	10 110	10 757	11 559	12 426	12 850

解:该省 2010—2015 年的钢产量年平均发展速度计算表如表 3-17 所示。

表 3-17 该省 2010—2015 年钢产量年平均发展速度计算表

年份	2010 年	2011 年	2012 年	2013 年	2014 年	2015 年
钢产量/万吨	a_0 9 400	a_1 10 110	a_2 10 757	a_3 11 559	a_4 12 426	a_5 12 850
环比发展速度/%		$x_1 = \frac{a_1}{a_0}$ 107.55	$x_2 = \frac{a_2}{a_1}$ 106.40	$x_3 = \frac{a_3}{a_2}$ 107.46	$x_4 = \frac{a_4}{a_3}$ 107.50	$x_5 = \frac{a_5}{a_4}$ 103.41

该省 2010—2015 年的钢产量平均发展速度计算如下:

$$G = \sqrt[n]{x_1 x_2 x_3 \cdots x_n} = \sqrt[5]{1.0755 \times 1.0640 \times 1.0746 \times 1.0750 \times 1.0341} = 106.45\%$$

2. 几何平均数的特点

几何平均数易受极端标志值的影响。当数列(总体)中某一标志值为零或负数时,几何平均数无法计算得出。

3. 几何平均数、算术平均数和调和平均数的关系

几何平均数、算术平均数和调和平均数三种平均数有其各自的应用条件和特点,但从数量关系上看,存在着一定的规律。对同一资料分别用三种方法计算,其结果是算术平均数最大,几何平均数次之,调和平均数最小。只有当所有变量值都相同时,三者方法的计算结果才相等。三者之间的关系可用不等式表示如下:

$$\bar{x} \geq G \geq H$$

第四节 变异指标

平均指标是说明总体各单位某一数量标志一般水平的综合指标,可以反映总体变量(各单位标志值)分布的集中趋势。但是,平均指标将总体各单位标志值的差异抽象化了,是总体各单位标志值的代表水平,不能反映总体各单位标志值的差异情况。

变异指标即标志变异指标,又称标志变动度,是反映分布数列(总体)中各单位标志值差异程度的综合指标。它从另一个角度反映总体的特征。

变异指标是衡量平均数代表性的尺度。它与平均数的代表性成反比,表明总体各单位标志值的分散程度。变异指标的数值越大,平均数的代表性越小;变异指标的数值越小,平

均数的代表性越大。

变异指标的作用是在与平均指标结合中产生的。离开了平均指标,它就失去了意义。而它与平均指标相结合,可全面反映总体的特征,并对平均指标的代表性做出评价。

变异指标包括全距、四分位差、标准差和变异系数。

一、全距

全距又称极差,是标志值数列(总体)中最大值与最小值之差。它说明标志值的变动范围,一般用 R 表示。其用公式表示如下:

$$\text{全距} = \text{最大标志值} - \text{最小标志值}$$

概念清楚、意义明确、计算简单是全距的优点。但是,因它仅由最大值与最小值求得,所以易受两极端数值的影响。如果两极端有偶然性或异常值时,全距就会不稳定、不可靠。它不考虑中间数值的差异,即其他数据不起作用,反应不灵敏。它明显地受取样变动的影

基于上述原因,全距只是一种低效的差异量,只能作为差异量的粗略指标。它一般只在研究的预备阶段使用,用以检查数据的大概散布范围,以便确定统计分组。

二、四分位差

四分位差也称内距,是上四分位数与下四分位数之差,记作 Q_d 。其计算公式如下:

$$Q_d = Q_U - Q_L$$

四分位差反映了中间 50% 数据的离散程度。其数值越小,说明中间的数据越集中;反之,说明中间的数据越分散。四分位差不受极值的影响。

此外,由于中位数处于数据的中间位置,因此四分位差在一定程度上也说明了中位数对一组数据的代表程度。

四分位差主要用于测度顺序数据的离散程度,也可用于测度数值型数据,但它不适合测度分类型数据。

【例 3-29】 承【例 3-20】,计算家庭人均月收入的四分位差。

解:由【例 3-19】的计算结果可知, Q_L 为 797.5 元, Q_U 为 1 437.5 元。

$$Q_d = 1\,437.5 - 797.5 = 640(\text{元})$$

三、标准差

标准差是分布数列(总体)中各单位标志值与其算术平均数离差平方的平均数的平方根。换句话说,标准差是各变量值离差平方平均数的平方根。标准差又叫均方差,用 σ 表示, σ^2 称为方差。

由于掌握的资料不同,标准差的计算可分为简单标准差和加权标准差两种形式。

(1) 简单标准差。简单标准差的计算公式如下:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (\text{未分组数列})$$

(2) 加权标准差。加权标准差的计算公式如下:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2 f_i}{\sum_{i=1}^n f_i}} \quad (\text{分组数列})$$

【例 3-30】 A、B 两组学生的成绩如下：

A 组：65 70 75 80 85

B 组：68 70 76 80 81

请问哪组学生的平均成绩更具代表性？

解：相关计算分析过程如表 3-18 所示。

表 3-18 A、B 两组学生成绩的平均数离差与离差平方

学生 序号			A 组		B 组	
	x_{iA}	x_{iB}	平均数离差 $x_{iA} - \bar{x}_A$	离差平方 $(x_{iA} - \bar{x}_A)^2$	平均数离差 $x_{iB} - \bar{x}_B$	离差平方 $(x_{iB} - \bar{x}_B)^2$
	甲	65	68	-10	100	-7
乙	70	70	-5	25	-5	25
丙	75	76	0	0	1	1
丁	80	80	5	25	5	25
戊	85	81	10	100	6	36
合计	375	375		250		136

$$\bar{x}_A = \bar{x}_B = 75(\text{分})$$

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^n (x_{iA} - \bar{x}_A)^2}{n}} = \sqrt{\frac{250}{5}} = 7.07(\text{分})$$

$$\sigma_B = \sqrt{\frac{\sum_{i=1}^n (x_{iB} - \bar{x}_B)^2}{n}} = \sqrt{\frac{136}{5}} = 5.22(\text{分})$$

由于 $\sigma_A > \sigma_B$ ，故 B 组学生的平均成绩比 A 组学生的平均成绩更具代表性。

四、变异系数

标准差的数值不但取决于数列各单位标志值的差异程度，而且受数列平均水平的影响。标准差在反映标志值的差异程度时还带有计量单位。因此，如果两个数列平均水平不同或标志值的计量单位不同，而比较其数列的变动度，即比较其数列平均数的代表性大小，就需要消除平均水平不同或计量单位不同的影响，计算出变异系数。

变异系数是分布数列(总体)中变异指标(标准差)与其算术平均数之比，反映标志值差异的相对水平。也就是说，变异系数反映单位平均水平下标志值的离散程度。分析人员常用标准差系数。

标准差系数的计算公式如下：

$$v_{\text{甲}} = \frac{\sigma}{x}$$

【例 3-31】 A、C 两组学生的成绩如表 3-19 所示。请问哪组学生的平均成绩更具代表性？

表 3-19 A、C 两组学生的成绩

学生序号	成绩/分	
	A 组	C 组
甲	65	79
乙	70	85
丙	75	90
丁	80	95
戊	85	100

解：相关计算分析过程如表 3-20 所示。

表 3-20 A、C 两组学生成绩的平均数离差与离差平方

学生序号	成绩/分		A 组		C 组	
	x_{iA}	x_{iC}	平均数离差	离差平方	平均数离差	离差平方
			$x_{iA} - \bar{x}_A$	$(x_{iA} - \bar{x}_A)^2$	$x_{iC} - \bar{x}_C$	$(x_{iC} - \bar{x}_C)^2$
甲	65	79	-10	100	-11	121
乙	70	85	-5	25	-5	25
丙	75	90	0	0	0	0
丁	80	95	5	25	5	25
戊	85	100	10	100	10	100
合计	375	449		250		271

$$\bar{x}_A = 75(\text{分}) \quad \bar{x}_C = 90(\text{分})$$

$$\sigma_A = \sqrt{\frac{\sum_{i=1}^n (x_{iA} - \bar{x}_A)^2}{n}} = \sqrt{\frac{250}{5}} = 7.07(\text{分})$$

$$v_A = \frac{\sigma_A}{x_A} = \frac{7.07}{75} = 0.0943$$

$$\sigma_C = \sqrt{\frac{\sum_{i=1}^n (x_{iC} - \bar{x}_C)^2}{n}} = \sqrt{\frac{271}{5}} = 7.36(\text{分})$$

$$v_C = \frac{\sigma_C}{x_C} = \frac{7.36}{90} = 0.0818$$

因为 $v_A > v_C$ ，所以 C 组学生的平均成绩比 A 组学生的平均成绩更具代表性。

需要注意的是，标准差与标准差系数的应用条件不同：在比较两个不同数列（总体）的标

志变异程度或说明平均数的代表性时,当两者的平均水平相同时,分析人员可通过直接计算标准差进行比较;当两者的平均水平不同或计量单位不同时,分析人员需要先消除平均水平不同或计量单位不同的影响,再通过计算标准差系数进行比较。

第五节 数据分布的偏度和峰度

一、偏度

(一) 偏度的概念

偏度是用来帮助判断数据序列的分布规律性的指标。测度偏度的统计量就是偏度系数,记作 SK。

(二) 偏度的类型

在数据序列呈对称分布(正态分布)的状态下,其均值、中位数和众数重合,而且在这三个数的两侧,其他所有的数据完全以对称的方式左右分布。如果数据序列的分布不对称,那么均值、中位数和众数必定分处不同的位置。这时,以中位数为参照点,位于均值左侧的数据较多,称为右偏;位于中位数右侧的数据较多,称为左偏。

(三) 偏度的测定

分析人员在根据未分组的原始数据计算偏度系数时,通常采用以下公式:

$$SK = \frac{n \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

式中, s^3 为样本标准差的三次方。

偏度系数测度了数据分布的非对称程度。若一组数据的分布是对称的,则偏度系数等于零;若偏度系数不等于零,则表明数据分布是非对称的。

分析人员在根据分组数据计算偏度系数时可采用以下公式:

$$SK = \frac{\sum_{i=1}^n (x_i - \bar{x})^3 f_i}{ns^3}$$

偏度系数是离差三次方的平均数再除以标准差的三次方。当数据分布对称时,离差三次方后正、负离差可以相互抵消,因而 SK 的分子等于 0,则 $SK=0$;当数据分布不对称时,正、负离差不能抵消,形成了正或负的偏度系数 SK。当 SK 为正时,表示正偏离差值较大,可判断为正偏或右偏;当 SK 为负时,表示负离差值较大,可判断为负偏或左偏。在计算 SK 时,将离差三次方的平均数除以 s^3 是将偏度系数转化为相对数。SK 的绝对值越大,表示偏斜的程度越大。

二、峰度

(一) 峰度的概念

峰度是对数据分布平峰或尖峰程度的测度。峰度一般用 K 表示。

(二) 峰度的类型

峰度通常是与标准正态分布相比较而言的。若一组数据服从标准正态分布,则峰度系数的值等于零;若峰度系数的值明显不等于零,则表明数据分布比正态分布更平或更尖,通常称为平峰分布或尖峰分布。

(三) 峰度的度量

分析人员在根据未分组数列计算峰度系数 K 时,通常采用以下公式:

$$K = \frac{n(n+1) \sum_{i=1}^n (x_i - \bar{x})^4 - 3 \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2 (n-1)}{(n-1)(n-2)(n-3)s^4}$$

根据分组数据计算峰度系数的方法是用离差四次方的平均数除以标准差的四次方。其计算公式如下:

$$K = \frac{\sum_{i=1}^n (x_i - \bar{x})^4 f_i}{ns^4} - 3$$

式中,将离差的四次方除以 s^4 是为了将峰度系数转化为相对数。

用峰度系数说明分布的偏尖和偏平程度是通过与标准正态分布的峰度系数进行比较得出的。正态分布的峰度系数为零。当 $K > 0$ 时,数据分布为尖峰分布;当 $K < 0$ 时,数据分布为平峰分布。

需要注意的是,分组数据的峰度系数公式也可以不减 3,则此时的比较标准是 3。也就是说,当 $K > 3$ 时,数据分布为尖峰分布;当 $K < 3$ 时,数据分布为平峰分布。

【例 3-32】 一家服装公司的 20 名销售人员 12 月的销售情况(单位:万元)如下:7,10,8,12,9,6,7,5,11,13,15,17,12,12,12,14,8,9,16,18。试用 Excel 中的描述统计工具计算这 20 个数据的集中趋势、离散程度、偏度和峰度等。

解:用 Excel 中的描述统计工具操作的步骤如下:

第一步,打开 Excel 表格,输入相关数据。

第二步,单击“文件”按钮,在打开的下拉菜单中选择“Excel 选项”按钮,在弹出的对话框中选择“加载项”项目,并将“管理”栏设为“Excel 加载项”。

第三步,单击“转到”按钮,在弹出的对话框中勾选全部选项,然后单击“确定”按钮。

第四步,单击“数据”按钮,在打开的下拉菜单中选择“数据分析”按钮,然后选择“描述统计”选项,单击“确定”按钮,弹出“描述统计”对话框,如图 3-1 所示。

第五步,在“描述统计”对话框的“输入区域”中输入需要计算的数据区域,在“输出选项”中选择“输出区域”“汇总统计”等,再单击“确定”按钮,结果显示如图 3-2 所示。

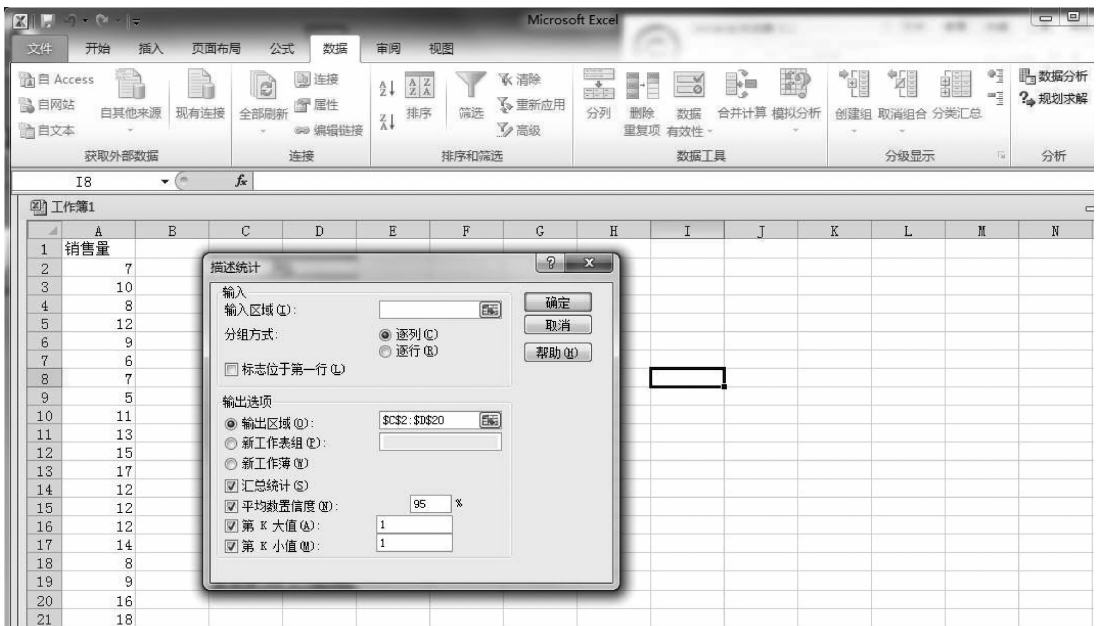


图 3-1 “描述统计”对话框

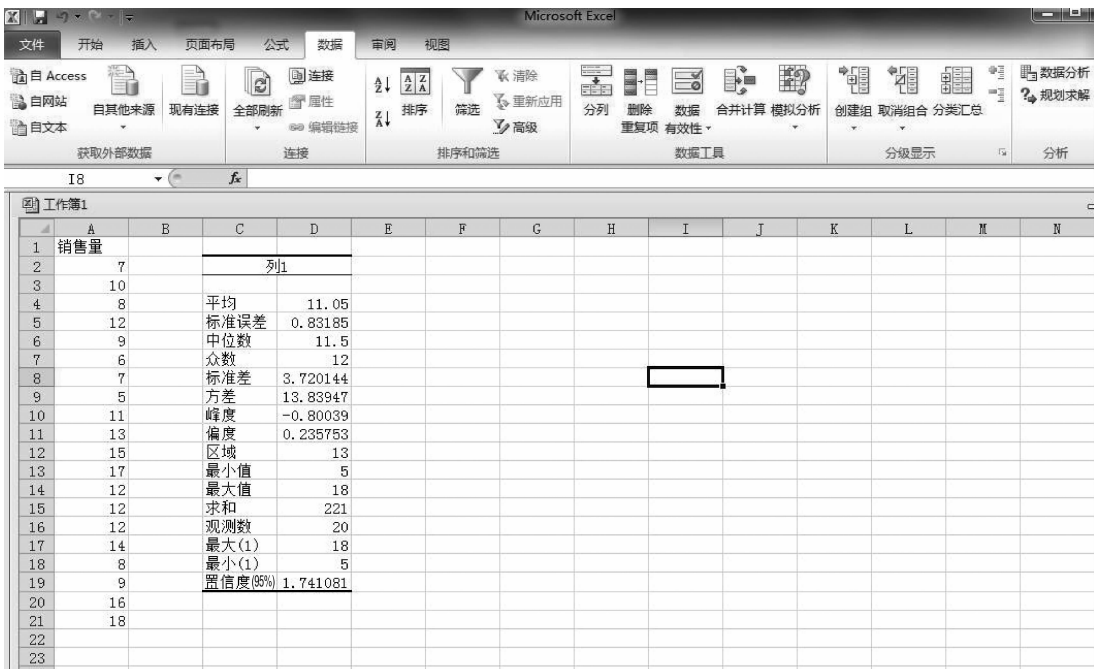


图 3-2 描述统计结果

描述统计工具操作的结果的含义如下：

- (1) 平均,即算术平均数。

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{7+10+8+\cdots+16+18}{20} = 11.05$$

(2) 标准误差,即标准差(样本标准差)除以样本单位数的平方根。

$$\frac{s}{\sqrt{n}} = \frac{3.720144}{\sqrt{20}} = 0.83185$$

(3) 中位数的计算。

$$M_e \text{ 位置} = \frac{20+1}{2} = 10.5$$

对数列由低到高进行排序,得到

$$M_e = \frac{11+12}{2} = 11.5$$

(4) 众数。本例中, $M_0 = 12$ 。

(5) 标准差,即样本标准差。

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = 3.720144$$

(6) 方差,即样本方差。它等于样本标准差的平方,即 $s^2 = 13.83947$ 。

(7) 峰度,即峰度系数。本例中,峰度为 -0.80039 。

(8) 偏度,即偏度系数。本例中,偏度为 0.235753 。

(9) 区域,即极差。它等于数据中的最大值与最小值之差,即 $18-5=13$ 。

(10) 求和,即所有数据的总和。本例中,求和为 221。

(11) 观测数,即数值的个数。本例中,观测数为 20。

(12) 置信度(95%),即当置信度为 95%时 t 分布的临界值。本例中,置信度(95%)约为 1.741081。



思考练习 >>>>

一、单项选择题

1. 社会经济统计中最基本的统计指标是()。

- A. 总量指标
- B. 相对指标
- C. 平均指标
- D. 变异指标

2. 总量指标按其说明总体内容的不同可分为()。

- A. 时期指标和时点指标
- B. 总体单位总量指标和总体标志总量指标
- C. 总体数量指标和总体质量指标
- D. 实物指标和价值指标

3. 下列用有名数表现的相对指标是()。
- 计划完成情况相对指标
 - 结构相对指标
 - 比例相对指标
 - 强度相对指标
4. 按人口平均的地区粮食产量属于()。
- 比例相对指标
 - 比较相对指标
 - 强度相对指标
 - 平均指标
5. 加权算术平均数中的权数为()。
- 变量值
 - 次数的总和
 - 变量值的总和
 - 次数比重
6. 某车间三个班组生产同种产品。这三个班组 6 月份的劳动生产率分别为 2 件/工日、3 件/工日和 4 件/工日,产量分别为 400 件、500 件和 600 件,则该车间平均劳动生产率的计算式应为()。
- $\frac{2+3+4}{3}=3$
 - $\frac{2 \times 400 + 3 \times 500 + 4 \times 600}{1\ 500}=3.13$
 - $\sqrt[3]{2 \times 3 \times 4}=2.88$
 - $\frac{1\ 500}{\frac{400}{2} + \frac{500}{3} + \frac{600}{4}}=2.9$
7. 某企业 2014 年完成产值 200 万元。该企业 2015 年计划增长 10%,实际完成 231 万元,则产值超额完成计划的()。
- 115.5%
 - 5.5%
 - 15.5%
 - 5%
8. 计算平均比率和平均速度最适宜的方法是()。
- 算术平均法
 - 调和平均法
 - 几何平均法
 - 综合指数法
9. 简单算术平均数和加权算术平均数的计算结果相同,是因为()。
- 各组次数相等
 - 各组次数不相等
 - 各组变量值相等
 - 各组变量值不相等
10. 下列变异指标中,易受极端值影响的是()。
- 全距
 - 四分位差
 - 标准差
 - 标准差系数

- D. 总体标志总量指标
E. 质量指标
3. 在相对指标中,分子和分母可以互换的有()。
- A. 计划完成情况相对指标
B. 动态相对指标
C. 比较相对指标
D. 比例相对指标
E. 结构相对指标
4. 反映国民经济产业结构的相对数有()。
- A. 国民生产总值
B. 各产业产值之比
C. 各产业增长速度
D. 各产业比上年的增长量
E. 各产业所占比重
5. 时点指标的特点包括()。
- A. 不同时间数值可以相加
B. 不同时间数值不可以相加
C. 调查资料需要连续登记
D. 数值与时期长短有关
E. 数值只能间断登记
6. 比较相对数适用于()。
- A. 计划水平与实际水平之比
B. 先进与落后之比
C. 不同国家之间的对比
D. 不同时间状态之比
E. 实际水平与标准水平之比
7. 某企业计划 2015 年产品成本降低 8%,但产品成本实际降低了 10%。下列说法中正确的有()。
- A. 该企业产品成本的计划完成程度为 125%
B. 该企业的成本实际比计划多降低了 2 个百分点
C. 该企业产品成本的计划完成程度为 97.83%
D. 该企业未完成成本降低任务
E. 该企业超额完成成本降低任务
8. 下列说法中正确的有()。
- A. 各组变量值作用的大小由各组频数的多少反映
B. 各组变量值作用的大小由各组变量值的大小反映
C. 频数越大的变量值对总体一般水平的影响越大
D. 频数越大的变量值对总体一般水平的影响越小
E. 频数越大,变量值也越大

9. 下列说法中正确的有()。
- 加权算术平均数受各组变量值的影响
 - 加权算术平均数受各组频数的影响
 - 加权算术平均数与各组变量值无关
 - 加权算术平均数与各组频数无关
 - 加权算术平均数受各组变量值和频数的共同影响
10. 下列现象中,应采用算术平均数公式计算平均数的有()。
- 已知粮食总产量与播种面积,求亩产量
 - 已知计划完成百分比与实际销售额,求销售额的平均计划完成百分比
 - 已知某厂 2011—2015 年的工业产值,求产值的平均发展速度
 - 已知甲、乙两个市场某商品的销售价格与销售量,求该商品的平均价格
 - 已知甲、乙、丙三种产品的合格率,求这三种产品的平均合格率
11. 下列各项中,应用加权算术平均法计算平均数的有()。
- 已知各组职工工资水平和各组职工人数,求平均工资
 - 已知各组职工工资水平和各组工资总额,求平均工资
 - 已知各组计划完成百分数和各组计划产值,求平均计划完成百分数
 - 已知各组计划完成百分数和各组实际产值,求平均计划完成百分数
 - 已知各组职工的劳动生产率和各组职工人数,求平均劳动生产率
12. 下列各项中,应采用几何平均法计算的有()。
- 生产同种产品的三个车间的平均合格率
 - 平均发展速度
 - 前后工序的三个车间的平均合格率
 - 平均劳动生产率
 - 以复利支付利息的年平均利率
13. 下列对平均指标的理解,正确的有()。
- 把某一数量标志在总体各单位之间的数量差异抽象化
 - 反映总体各单位某一数量标志值的一般水平,是一个代表性数值
 - 现象总体的同质性是计算和应用平均数的前提条件
 - 反映总体各单位标志值之间的差异
 - 反映总体各单位变量值分布的集中趋势
14. 比较两组工作成绩:算术平均数,甲组小于乙组;标准差,甲组大于乙组。那么,下列说法中正确的有()。
- 乙组算术平均数的代表性高于甲组
 - 甲组算术平均数的代表性高于乙组
 - 乙组工作的均衡性高于甲组
 - 甲组工作的均衡性高于乙组
 - 甲组的离散程度大于乙组
15. 将所有变量值都减去 10,结果可能有()。
- 算术平均数不变

- B. 算术平均数减去 10
- C. 方差不变
- D. 标准差不变
- E. 标准差系数不变

三、简答题

1. 什么是总量指标？它有什么作用？
2. 如何理解权数的含义？
3. 如何区分结构相对指标、比例相对指标、比较相对指标和强度相对指标？

四、计算题

1. 某企业下属三个分公司 2015 年上半年的销售情况如表 3-21 所示。

表 3-21 某企业下属三个分公司 2015 年上半年的销售情况

公司名称	第一季度 销售额 /万元	第二季度					第二季度销售 额与第一季度 销售额之比/%
		销售额/万元		销售额比例/%		销售计划 完成程度/%	
		计划	实际	计划	实际		
甲公司	420	550					130
乙公司	820	830				120	
丙公司	260		270	15			
合计							

要求：根据资料填写所缺的数字，并说明各指标属于何种指标。

2. 某工厂生产某种产品。根据五年计划的要求，最后 1 年的产量应达到 100 万吨。生产计划执行情况如表 3-22 所示。

表 3-22 某工厂生产计划执行情况

时间	第一年	第二年	第三年		第四年				第五年			
			上半 年	下半 年	一季 度	二季 度	三季 度	四季 度	一季 度	二季 度	三季 度	四季 度
产量 /万吨	78	82	44	45	23.5	24	24.5	25	25	26	26.5	27.5

要求：计算该种产品计划完成情况相对指标，并说明该工厂提前多长时间完成五年计划。

3. 某地区 100 个工业企业的利润完成程度资料如表 3-23 所示。

表 3-23 某地区工业企业的利润完成程度

计划完成程度/%	企 业 数	实际利润额/万元
[80,90)	12	425
[90,100)	18	950
[100,110)	40	8 400
[110,120)	20	3 450
[120,+∞)	10	2 500
合计	100	15 725

要求:计算该地区 100 个工业企业的平均利润计划完成程度。

4. 2015 年 6 月,甲、乙两个农贸市场西红柿的成交资料如表 3-24 所示。

表 3-24 甲、乙农贸市场西红柿成交资料

产品品种	价格/(元·千克 ⁻¹)	甲市场成交额/元	乙市场成交量/千克
A	2.4	12 000	10 000
B	2.8	28 000	5 000
C	3	15 000	5 000
合计		55 000	20 000

要求:试比较两个农贸市场西红柿的平均价格。

案例分析

根据我国有关部门进行的薪金调查得知,某地区高等学校毕业生的平均薪金为每年 34 500 元。将 30 名 2015 年毕业的高等学校毕业生作为一个样本,其年薪数据如表 3-25 所示。

表 3-25 某地区高等学校毕业生年薪情况

学生编号	年薪/元	学生编号	年薪/元
1	36 800	16	35 000
2	35 800	17	36 500
3	37 300	18	36 400
4	38 300	19	37 200
5	38 300	20	36 700
6	38 800	21	36 400
7	34 900	22	36 700
8	36 800	23	38 400
9	38 200	24	37 000
10	36 000	25	36 200

续表

学生编号	年薪/元	学生编号	年薪/元
11	36 400	26	36 600
12	35 400	27	39 000
13	35 200	28	37 900
14	36 100	29	39 400
15	36 300	30	36 400

要求:计算该地区高等学校毕业生年薪的平均数、中位数、众数、四分位差、标准差,并对它们进行分析。