

# 第 1 章

## 统计和统计数据收集

本章将介绍统计学的基本知识，统计学的基本术语，统计数据的收集和调查，以及变量的类型。

### 1.1 统计学概述

#### 1.1.1 统计的广泛应用

人类活动的各个方面都离不开统计工作和统计数据。

在个人生活中，人们的任何决策都依赖于有关的统计信息。无论是报考学校、选择工作单位、购买房屋、进行股票交易、外出旅游，还是购买日常生活用品、生活开支预算等，都离不开有关的统计信息。尽管个人生活中的许多统计信息是以非常简单粗糙的形式出现的，但正是在对过去发生的事情或经验所获得的信息进行综合的基础上，人们才能作出正确的判断和决策。

在政府的层次上，统计更是渗透于每一个部门的管理工作之中。政府部门通过不断收集经济、社会、人口等各个领域的统计数据，在综合分析的基础上对国民经济的发展进行预测、规划、指导和调控，并对全社会提供各种统计信息。

在科学和学术研究的各个领域，包括自然科学、经济学、社会学、体育、医疗卫生、环境保护等领域，都广泛使用统计学方法进行分析和推断。

在商务活动中，统计学有更广泛的应用，经营管理人员在作决策时需要应用统计方法来归纳分析各种可以获得的统计资料(参看如图 1-1 所示的由数据向信息和决策的转换过程)。商务活动中经常使用统计方法的有以下几个方面。

##### 1. 财务分析

企业的财务报告和各项财务分析指标都是对企业的财务状况和经营成果的统计分析结果；管理会计各种方法的应用需要以成本和收益为基础的经营业绩的统计分析；企业的投资和融资决策更离不开各种内外部统计信息的支持。

## 2. 产品开发

企业产品开发计划的制定，需要对经济发展趋势、商业竞争、顾客需求、财务收支估算等方面的数据进行统计分析，并在此基础上进行产品开发的可行性分析。

## 3. 计划

企业各种计划的制定，都需要对销售、资金、人力资源需求、成本和利润等因素进行预测，而预测就是以过去和当前的统计数据为依据，运用统计推断方法对未来情况进行估计。

## 4. 市场研究

在市场研究中，需要对消费者的需求偏好及其变化趋势，竞争对手的情况，本企业产品和服务的顾客满意度等方面的数据进行统计分析。

## 5. 工序管理和质量控制

运用统计分析方法，可以帮助确定影响产品质量的主要因素，制定有效的质量控制标准和工艺规范，达到降低成本、提高生产效率的目标。质量控制和质量管理是企业经营管理中运用统计分析方法最多的领域。

## 6. 人力资源管理

在企业的人力资源管理中，需要经常使用统计方法来分析人事变动、出勤状况、工作业绩等情况，并在此基础上进行业绩评定、奖励和惩罚、制定有效的激励约束机制等。

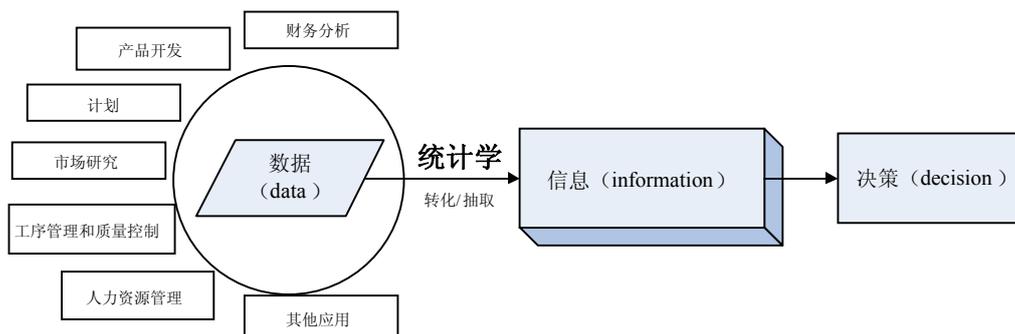


图 1-1 基于统计学的数据转换应用

### 1.1.2 统计与统计学

关于“统计”一词，可以有三种理解：统计工作、统计资料、统计学。

- 统计工作是指应用科学的方法对调查研究的对象进行数据搜集、整理和分析的全部工作，其成果是形成各种统计数据资料。
- 统计资料是统计工作所取得的成果，包括两大类型：原始资料和再生信息。
- 统计学是关于如何搜集、整理和分析客观现象数量规律的一门方法论的科学，是对统计工作实践加以总结升华而产生的理论，并用以指导统计实践。

### 1.1.3 统计研究对象的特点

统计研究的对象有以下4个方面的特点。

#### 1. 数量性

统计学研究的对象是客观现象的数量特征和规律性。统计学是通过大量的观察方法来研究事物总体数量方面的特征及其规律性，反映客观事物在一定时间、地点等条件下的数量表现及其发展趋势和变化规律，为人类活动的各种决策提供依据。凡是产生数据的领域，都需要统计工作和统计学；同样，凡是能以数量来描述的事物，都可以作为统计学的研究对象。

#### 2. 总体性

统计学研究的是客观现象总体的数量特征与规律性，而不是个体的量。总体由具有某种相同性质的一定范围内的全体事物组成。例如：对全国农村家庭收入情况进行统计分析，此时全国所有农村家庭的收入就构成研究的总体，但统计学对总体数量规律性的认识是通过对大量个体的观察和分析后获得的。

#### 3. 具体性

统计学研究的对象是一定时间、地点等条件下具体事物的量，而不是抽象对象的量，这是统计学和数学的一个重要区别。因此，统计工作需要具体事物进行调查研究，收集和整理特定时间、地点等条件下所研究对象中大量个体的有关数据是统计分析的基础。

#### 4. 差异(变异)性

组成统计研究对象总体的个体应是有差异的，否则就不需要进行统计分析。客观事物是错综复杂的，受到多种因素的影响，因而不同个体在数量方面必定存在差异，这就需要通过总体中大量的个体进行观察并进行综合分析，才能获得总体的数量特征，如平均值、方差、偏斜度等方面的分布特征。

### 1.1.4 统计学的分类

根据研究的重点和观察问题的角度不同，统计学主要有以下两种分类方法。

#### 1. 描述统计学和推断统计学

这一分类方法既反映了统计学发展的两个主要阶段，同时也反映了各自不同的侧重。

**描述统计学**是指研究如何对客观现象进行数量的计量、加工、概括和表示的方法。在20世纪之前，统计学基本上处于描述阶段，描述统计学是统计学的基础。

**推断统计学**是指研究如何根据样本数据来推断总体的分布情况，概率论是推断统计学的主要理论基础。推断统计学是近代统计学的核心，也是统计学中的主要内容。

## 2. 理论统计学和应用统计学

**理论统计学**主要研究统计学的数学原理，它基于概率论的原理，还包括不属于传统概率论的一些内容，如随机化原理的理论、各种估计的原理、假设检验的原理以及一般决策的原理。在统计实践中经常会遇到一些原有的统计方法不能解决的新问题，需要创造新的统计模型和统计分析方法，这就需要统计理论的研究与指导。

将统计学的基本原理应用于各个领域就形成各种**应用统计学**的分支。它包括适用于各个领域的一般性的统计分析方法，如参数估计、假设检验、方差分析、回归分析等，还包括在某一领域中特定的统计分析方法，如经济领域中的指数分析法等。应用统计学需要既熟悉统计知识又熟悉某一领域业务知识的专门人才，它侧重于阐明统计的基本原理，并将理论统计学的结论作为工具应用于各个领域。

## 1.2 统计基本术语

**变量**是事物的特征，是运用统计方法所分析的对象。例如，在商业模型中，销售额、每年的开支和每年的净利润都是企业想要分析的变量。

**数据**是与变量相关的值。变量可能随时间变动，如某一公司的期望销售额、开支和净利润每年都有所不同。这些不同的值就是与变量相关的数据，或者简单地说，就是统计所要分析的“数据”。

除了时间，变量的数据也可能因其他原因而不同。例如，如果你要分析参加一个大型讲座的人员组成，包括的变量可能有年级、性别和专业。因为班级里每位同学都各不相同，所以这些变量的值也会有所不同。一名学生也许是经济学专业的大一男生，而另一名则有可能是金融学专业的二年级女生。

但需要记住的是，除非赋予实际操作定义，否则所有变量的值(或数据)都是没有任何意义的。而且，进行分析时我们必须非常清楚这些实际操作的定义并一致接受，否则就会产生歧义。例如，对销售额的操作性定义可能会发生这样错误的理解：一个人认为年销售额是指全部连锁店的年销售额，而另一个人则认为是每家店的年销售额。即使是变量中的单个值有时也需要赋予操作性定义。例如年级变量，到底什么是二年级和三年级？

对变量和数据之间区别的理解有助于其他基本术语的学习。

**总体**是指所研究对象的全体，或者具体指研究对象的某项数量指标的值的全体。

**个体**是指总体中的每个元素或单元。总体依其所包含的个体总数分为有限总体和无限总体。

**样本**是从总体中挑选出来用于分析的一部分。这种挑选通常是相对独立的。样本是进行推断统计的依据。

**参数**是描述总体特征的数值。

**统计量**是描述样本特征的数值。进行推断统计时，往往不是直接利用样本本身，而是通过基于样本构造样本的适当函数(例如，样本平均值、样本方差等)获得统计量进行分析。

以上术语可以通过下面的示例——沃尔玛某一年所有的销售交易进行解释。

总体：沃尔玛某一年所有的销售交易。

个体：沃尔玛某一年所有销售交易的每一条交易记录。

样本：从沃尔玛某一年所有销售交易记录中随机选取 200 条，样本中的交易数代表了组成总体的事物的一部分。

参数：沃尔玛某一年所有销售交易记录的消费值。

统计量：随机选取的 200 次交易记录的平均消费值。

## 1.3 数据的收集

统计数据的收集是统计整理和分析，以及统计推断和预测的基础。统计数据的收集就是根据统计研究的目的和要求，有组织、有计划地向调查对象搜集原始资料的过程。确定数据最合适的来源和收集数据的合理方法是非常重要的任务，因为如果收集的数据有偏差，模糊不清或有其他类型的错误时，即使最复杂的统计方法也无法得到有用的信息，即需要避免“Garbage In Garbage Out(垃圾数据产生垃圾统计结果)”的统计应用情况。

在实际应用中，我们要依据特定的研究目的或工作任务，结合研究对象所具有的性质和特点，相应地选择适合的调查方法，必要时也可以几种调查方法结合使用。这里，我们介绍几种通常采用的调查方法，包括：普遍调查、重点调查、典型调查、抽样调查和网上调查。

### 1.3.1 普遍调查

普遍调查简称普查，是专门组织的一次性的全面调查。如全国的人口普查、能源普查、工业普查等。普查的组织方式一般有两种：一种是建立专门的普查机构，配备大量的普查人员；另一种是利用调查单位的原始记录和核算资料，发放调查表，由登记单位填报，如物资库存普查等。普查时必须注意以下几个原则。

(1) 规定统一的标准时点。标准时点是指对被调查对象登记时所依据的统一时点。

(2) 规定统一的普查期限。在普查范围内各调查单位或调查点尽可能同时进行登记。

(3) 规定普查的项目和指标。普查时必须按照统一规定的项目和指标进行登记，不准任意改变或增减，以免影响汇总和综合，降低资料质量。

**【例 1.1】** 2002 年在国务院统一部署下，开展全国“基本单位普查”。关于这次普查的部分重要要求与规定如下。

**调查目的：**统计分析全国和各省市地区的单位总数、性质、隶属关系、分布情况等。

**标准时间：**2001 年 12 月 31 日。

**调查对象：**各类法人单位，各类法人单位所属的产业活动单位。

**调查项目：**单位总数；

性质分类 (企业法人、事业法人、社团法人、机关法人、其他法人)；

基本情况——性质划分、隶属关系 (中央、省市、区县、街道等)；

主要构成 (产业结构、行业分布、地区分布、经济成分、规模结构)；

分布情况。

### 1.3.2 重点调查

重点调查是专门组织的一种非全面调查,它是在总体中选择个别的或部分重点单位进行调查,藉以了解总体的基本情况。这些单位虽然少,但它们调查的标志值在总体标志总量中占有绝大部分的比重,通过对这些单位的调查,就能掌握总体的基本情况。例如,要了解全国钢铁企业的生产情况,只要调查宝钢、鞍钢、马钢、包钢、首钢等大型钢铁公司,就能达到调查的目的。因为这些钢铁企业虽在全国只占少数,但它们的产量在全国钢产量中占有绝大部分的比重。因此当调查的目的只是掌握调查对象的基本情况,而在总体中却有部分单位能较集中地反映所研究的问题时,采用重点调查是比较合适的。

重点调查的特点如下:

(1) 重点调查适用于调查对象的标志值比较集中于某些单位的场合,这些单位的管理比较健全,统计力量比较充实,能够及时取得准确资料。

(2) 重点调查的目的在于了解总体现象某些方面的基本情况,而不要求全面准确地推算总体数字。

(3) 重点调查比实际调查的单位数目少,在满足调查目的所要求的前提下,可以比全面调查节省人力、物力和时间。

### 1.3.3 典型调查

典型调查是根据调查的目的,选择在同类对象中最具典型性的部分和个体进行调查。典型调查也是一种非全面调查。例如,选择第一汽车制造厂作为国有企业改革情况调查分析的样板。典型调查的作用如下:

- 典型调查可用来研究新生事物。
- 典型调查可用来研究事物变化的规律。
- 典型调查可用来分析事物的不同类型,研究它们之间的差别和相互关系。
- 典型调查的资料可用来补充和验证全面统计的数字,推论和测算有关现象的总体。

### 1.3.4 抽样调查

抽样调查是一种专门组织的非全面调查,它按照随机原则,从总体中抽取部分单位进行观察,用观察的结果来推算全部总体的某些数值,即以部分推断全体。抽样调查是现代推断统计的核心,因为无论是对总体的参数估计或假设检验,都是以测定样本得到样本指标——统计量为依据的。

在进行抽样调查时,应尽可能避免下列四种误差,以使抽样调查结果更好地反映总体特性。

### 1. 涵盖误差

涵盖误差是指当某一组有代表性的样本被排除在抽样调查之外时所引起的选择偏差。例如,抽样统计上海市全市超市运营性能时,如果没有包括易初莲花超市所得到的统计结果,那么该统计在一定程度上就存在涵盖误差。

### 2. 无回应误差

抽样时,对样本个体数据收集失败会导致无回应偏差。例如,某项调查共发出 1 000 份调查表,但是最终进行统计分析时只得到 240 份有效调查,其余 760 份未能得到回复。由于存在 760 份未知的个体数据,这样的统计结果难以确保正确性和全面性。

### 3. 抽样误差

选择抽样调查是因为这种方法简单、低成本和有效。但同时也意味着有的个体被抽中,有的个体没有被抽中。抽样误差就是反映样本间的这种波动,其大小是基于某些特殊样本情况被抽中的概率而确定的。

当你读到调查报告、报纸或杂志中的民意调查时,其中往往会给出波动的误差。例如,“本次民意调查与实际情况的误差在 $\pm 4$ 个百分点之内”,这样的误差就是抽样误差。如何正确描述这种抽样误差是推断统计学的重要内容之一。

### 4. 测量误差

测量误差是指由于样本数据测量程序的设计和应用不当所引起的误差。例如,抽样调查时由于问题设计的不明确性而引起的变量数据的模糊性。

## 1.3.5 网上调查<sup>1</sup>

互联网的发展把我们带入了网络经济时代,传统的调查理论与国际互联网技术结合的要求,使得网上调查应运而生。1999年10月16日,北京零点专业市场调查公司与爱特信搜狐网络公司正式携手,创立了搜狐—零点网上调查公司,共同拓展网上调查业务,这标志着中国调查业正式步入了“网络时代”。

1997年初,我国网民是20万人,1999年7月就增加到了400万人。截至2008年底,我国互联网普及率以22.6%的比例首次超过21.9%的全球平均水平。这是继2008年6月中国网民规模超过美国,一举成为全球第一之后,中国的互联网普及再次实现飞跃,赶上并超过了全球平均水平。截至2009年6月底,中国网民规模达到3.38亿人,较2008年底增长13.4%,上网普及率达到25.5%。网民规模持续扩大,互联网普及率平稳上升。其中宽带网民数达到3.2亿,占比高达94.3%,较

---

1. 本节内容引自高志敏:《网上调查——一种新兴的调查方式》, <http://dms.zje.net.cn/jhtj/jhtj-blgxl.html>, 经过编者加工整理。

2008 年底上升了 3.7 个百分点。另外,截至 2009 年 6 月底,作为国家域名, CN 域名总计达到 1 296 万个。作为我国互联网的重要基础资源, CN 域名已经广泛地应用在金融、汽车等各个行业中。截至 2009 年 6 月我国互联网用户的发展情况如图 1-2 所示。



图 1-2 截至 2009 年 6 月我国互联网用户的发展情况  
(资料来源:第 24 次中国互联网络发展状况统计报告)

网上调查的主要方式有如下几种。

#### (1) E-mail 法

E-mail 法即电子邮件法,是以较为完整的 E-mail 地址清单作为样本框,使用随机的方法发送问卷进行调查。这种调查方法主要用于对特定群体——网民——多方面的行为模式、消费规模、网上消费者心理特征等的研究,在调查实施过程中,还可通过多媒体技术,向被调查者展示包括问卷、图像、样品在内的多种测试工具,以获得更加客观、全面的资料。在样本框较为全面的情况下,调查结果可用于推论研究的总体。由于几乎每个网民上网时都会处理 E-mail,因此 E-mail 调查法是当前最主要的网上调查法。

#### (2) Web 站点法

Web 站点法又称主动浏览访问法,即将调查问卷放置在访问率较高的 Web 站点的页面上,由对该问题感兴趣的访问者完成并提交。

#### (3) Net-meeting 法

Net-meeting 法即网络会议法、视频会议法、焦点团体座谈会。通过直接在网上人士中征集与会者,并在约定时间举行网上座谈会,在主持人的引导下,对某一问题进行深入的或探索性的讨论、研究的一种网上调查方法。

#### (4) Internet Phone 法

Internet Phone 法即网络电话法,是以 IP 地址为抽样框,采用 IP 自动拨叫技术,邀请用户参与调查。比如:可将 IP 地址排序,每隔 100 个进行一次抽样,被抽中的用户会自动弹出一个小窗口,询问其是否愿意接受调查,回答“是”,则弹出调查问卷;回答“否”,则呼叫下一个 IP 地址。这种调查方法类似于传统调查方式中的电话调查。

## 1.4 问卷设计

统计调查采用问卷的形式比较普遍,如网上调查、抽样调查、重点调查和典型调查等。问卷是一种特殊形式的调查表。其特点是表中用一系列按照严密逻辑结构组成的问题,向被调查者调查具体事实和个人对某问题的反映、看法,它不要求被调查者填写姓名。问卷设计一般要遵循以下原则。

### 1.4.1 合理性

合理性指的是问卷必须与调查主题紧密相关。违背了这一点,再漂亮或精美的问卷都是无益的。而所谓问卷应体现调查主题其实质是在问卷设计之初要找出与“调查主题相关的要素”。

例如:“调查某化妆品的用户消费感受”——这里并没有一个现成的选择要素的法则。但从问题出发,特别是结合一定的行业经验与商业知识,要素是能够被寻找出来的:一是使用者(可认定为购买者),包括她(他)的基本情况(自然状况,如性别、年龄、皮肤性质等),以及使用化妆品的情况(是否使用过该化妆品、周期、使用化妆品的日常习惯等);二是购买力和购买欲,包括她(他)的社会状况、收入水平、受教育程度、职业等,化妆品的消费特点(品牌、包装、价位、产品外观等)和使用该化妆品的效果(评价:问题应具有一定的多样性,但又限制在某个范围内,如价格、使用效果、心理满足等);三是产品本身,包括对包装与商标的评价、广告等促销手段的影响力、与市场上同类产品的横向比较等。应该说,具有了这样几个要素对于调查主题的结果是有直接帮助的。被访问者也相对容易了解调查员的意图,从而予以配合。

### 1.4.2 一般性

一般性,即问题的设置是否具有普遍意义。

应该说,这是问卷设计的一个基本要求,但我们仍然能够在问卷中发现这类带有一定常识性的错误。这一错误不仅不利于调查成果的整理分析,而且会使调查委托方轻视调查者的水平。如搞一个“居民广告接受度”的调查:

问题:你通常选择哪一种广告媒体?

答案:a.报纸 b.电视 c.杂志 d.广播 e.其他

而如果答案是另一种形式:

a.报纸 b.车票 c.电视 d.幕墙广告 e.气球 f.大巴士 g.广告衫 h.……

那情况将会怎样?

如果我们的统计指标没有那么细(或根本没必要),那我们就犯了一个“特殊性”的错误,从而导致某些问题的回答实际上是对调查无助的!

在一般性的问卷技巧中,需要注意的是,不能犯问题内容上的错误。如:

问题: 你拥有哪一种信用卡?

答案: a. 长城卡 b. 牡丹卡 c. 龙卡 d. 维萨卡 e. 金穗卡

其中“d”的设置是错误的, 应该避免。

### 1.4.3 逻辑性

问卷的设计要有整体感, 即问题与问题之间要具有逻辑性, 独立的问题本身也不能出现逻辑上的谬误, 从而使问卷成为一个相对完善的小系统。例如, 下列问题:

(1) 你通常每日读几份报纸?

a. 不读报 b. 1份 c. 2份 d. 3份以上

(2) 你通常用多长时间读报?

a. 10分钟以内 b. 30分钟左右 c. 1小时 d. 1小时以上

(3) 你经常读的是下面哪类(或几类) 报纸?

a. ×市晚报 b. ×省日报 c. 人民日报 d. 参考消息 e. 中央广播电视报 f. 足球……

在以上的几个问题中, 由于问题设置紧密相关, 因而能够获得比较完整的信息。调查对象也会感到问题集中, 提问有章法, 相反, 假如问题是发散的、带有意识流痕迹的, 问卷就会给人以随意性而不是严谨性的感觉。那么, 将市场调查作为经营决策的一个科学过程的企业就会对调查失去信心。

因此, 逻辑性的要求是与问卷的条理性、程序性分不开的。在一个综合性的问卷中, 调查者应将差异较大的问卷分块设置, 从而保证了每个“分块”的问题都密切相关。

### 1.4.4 明确性

所谓明确性, 事实上是问题设置的规范性。这一原则具体是指: 命题是否准确; 提问是否清晰明确, 便于回答; 被访问者是否能够对问题作出明确的回答等等。

如上文问题中“10分钟”、“30分钟”、“1小时”等设计即是十分明确的。统计后会告诉我们: 用时极短(浏览)的概率为多少; 用时一般(粗阅)的概率为多少; 用时较长(详阅)的概率为多少; 反之, 答案若设置为“10~60分钟”, 或“1小时以内”等, 则不仅不明确, 难以说明问题, 而且令被访问者也很难作答。

再则, 问卷中常有“是”或“否”一类的是非式命题。例如:

问题: 您的婚姻状况如何?

答案: a. 已婚 b. 未婚

显而易见, 此题还有第三种答案(离婚/丧偶/分居)。如按照以上方式设置, 则不可避免地会发生选择上的困难和有效信息的流失! 其症结即在于问卷违背了“明确性”的原则。

### 1.4.5 非诱导性

不成功的记者经常会在采访中使用诱导性的问题。这种提问方式如果不是刻意地要得出某种

结论而甘愿放弃客观性的原则，就是彻头彻尾的职业素质的缺乏。在问卷调查中，因为有充分的时间作提前准备，这种错误大大地减少了。但这一原则之所以成为必要，是在于高度竞争的市场对调查业的发展提出了更高的要求。

非诱导性指的是问题要设置在中性位置、不参与提示或主观臆断，完全将被访问者的独立性与客观性摆在问卷操作的限制条件的位置上。例如：

问题：你认为这种化妆品对你的吸引力在哪里？

答案：a. 色泽 b. 气味 c. 使用效果 d. 包装 e. 价格 f. ……

这种设置是客观的。若换一种答案设置：

a. 迷人的色泽 b. 芳香的气味 c. 满意的效果 d. 精美的包装 e. ……

这样一种设置则具有了诱导和提示性，从而在不知不觉中掩盖了事物的真实性。

#### 1.4.6 便于整理、分析

成功的问卷设计除了要紧密结合调查主题与方便信息收集外，还要考虑调查结果是否容易得出和调查结果的说服力。这就涉及问卷在调查后的整理与分析工作：首先，要求调查指标是能够累加和便于累加的；其次，指标的累计与相对数的计算是有意义的；再次，能够通过数据清楚地说明了所要调查的问题。

只有这样，调查工作才能收到预期的效果。

## 1.5 变量

### 1.5.1 变量类型

统计学中将变量分为分类变量和数值变量。图 1-3 显示了变量之间的关系，并举例说明了每种变量。

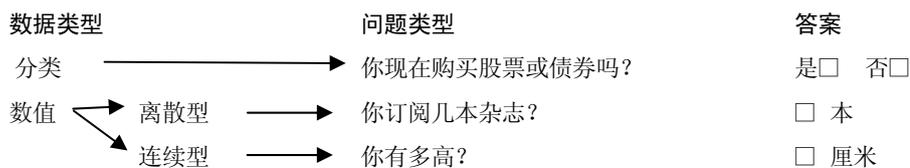


图 1-3 变量类型

**分类变量**(也称为定性变量)的值只能按类别分开，例如图 1-3 中的“是”和“否”。分类变量也可以有多个答案。例如，顾客写出他们在周几购买了商品，可能是周一到周日中的一天或几天。

**数值变量**(也称为定量变量)的值表示数量。数值变量可进一步分为离散变量和连续变量。

离散变量的值随计数过程逐渐增加。“杂志的订阅数量”就是一个离散变量的例子，因为反应变量值是整数中某个确定的数值，如订阅 0、1、2 等数量的杂志。从订购杂志那天起到收到杂志那天之间的天数是一个离散变量，因为是以天来计数的。

连续变量的数值随度量过程而逐渐增加。例如，在银行等待出纳员服务的时间是连续变量，因为反应变量值可以是一个闭区间或开区间内的任何一个值，取决于测量设备的精度。例如，等候时间可以是 1 分钟、1.1 分钟、1.11 分钟或 1.113 分钟，取决于你测量时间所使用的工具。

理论上讲，如果测量设备的精度足够高，任何两个连续变量值都不会相同。但是实际上，大部分测量设备都不够精确到可以发现微小的差异，因此，在试验或调查数据中经常会见到两个或两个以上相同的连续变量值。

### 1.5.2 度量水平和度量等级

应用度量水平是将数据分类的另一种方法。有四个广为认可的度量水平：定类尺度、定序尺度、定距尺度和定比尺度。

**定类尺度(nominal measurement)**，又称为**名义尺度**。分类变量数据的度量是在定类尺度上进行的。定类尺度(见图 1-4)是将数据分为不同的类别，这些类别中没有排序。例如，1.4.2 节中“a. 长城卡 b. 牡丹卡 c. 龙卡 d. 金穗卡”就是定类尺度变量的一个例子。你最喜欢的软饮料、你所属的政治党派和你的性别等也是这样的例子。定类尺度是度量的最弱形式，因为无法对不同的类别进行排序。

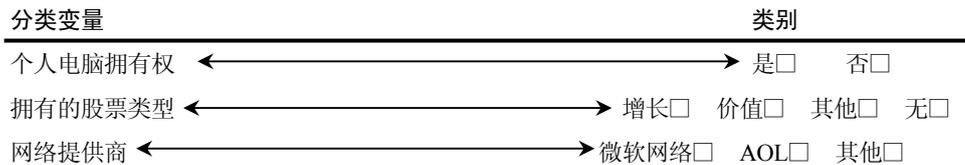


图 1-4 定类尺度举例

**定序尺度(ordinal measurement)**是将数据分为不同的类别，但可以进行排序。如 1.4.3 节中“a. 10 分钟以内 b. 30 分钟左右 c. 1 小时 d. 1 小时以上”就表示一个定序尺度变量，因为这是按时间长度排序的。另外，常见的用户反应值“非常好、很好、一般和差”是按照满意度排序的。图 1-5 列举了几个定序尺度变量的例子。

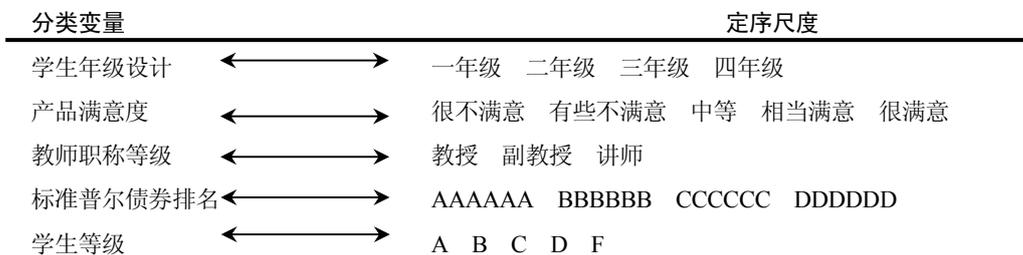


图 1-5 定序尺度举例

定序尺度是比定类尺度具有更强形式的度量, 因为其数值被赋予的性质多于仅仅被归为某一个类别。但是, 定序尺度仍然是一个相对较弱形式的度量, 因为该尺度没有度量出类别之间的数量差距。

定距尺度(interval measurement)和定比尺度(ratio measurement)是将数值变量数据在区间或比例的尺度上进行度量。定距尺度(见图 1-6)是一个顺序尺度, 度量值之间的差异是一个有意义的数量, 但是缺乏真正意义上的参考值(“0”值)。例如, 中午温度读数摄氏 28 度比摄氏 26 度温暖 2 度。另外, 中午温度读数摄氏 2 度的差值与摄氏 32 度和 30 度的差值意义是相同的, 因为在这个尺度上所有的差值意义都是相同的。定比尺度(见图 1-6)是一个定序尺度, 其中度量值之间的差异存在真正的参考值, 如高度、重量、年龄或薪水。如一个体重为 240 磅的人的重量是一个体重为 120 磅的人的体重的两倍。温度是一个特殊的例子: 华氏度和摄氏度都是定距尺度而非定比尺度; 你不能说中午温度读数华氏 4 度是华氏 2 度的两倍热。但是在开氏温度读数中, 0 度意味着没有分子运动, 是定比尺度。相反的, 华氏和摄氏是随意选取的 0 度起始点。

数值变量	度量水平
温度(摄氏度或华氏度)	定距
标准化考试分数(例如全国英语等级考试或注册会计师考试)	定距
高度(英寸或厘米)	定比
重量(磅或千克)	定比
年龄(岁或天)	定比
薪水(美元或人民币)	定比

图 1-6 定距和定比尺度举例

在定距尺度或定比尺度上度量的数据是最高水平的度量, 是比定序尺度更强形式的度量, 因为不仅可以确定哪个观察值最大, 而且可以确定大多少。

上述四种度量水平, 具有各自的特点(如表 1-1)。这些不同层次的度量本身形成了一个累积尺度, 即高一层次的尺度除自己的特性外, 必包含下一层次尺度的所有特性。高层次度量具有向下的兼容性, 而低层次度量不具有向上的兼容性。

表 1-1 四种测量尺度的数学特性

分 类	定 类 尺 度	定 序 尺 度	定 距 尺 度	定 比 尺 度
类别区分	有	有	有	有
次序区分		有	有	有
距离区分			有	有
比例区分				有

## 习题一

1. 样本和总体的区别是什么?
2. 统计量和参数的区别是什么?
3. 什么是普查? 其特点和作用如何?
4. 什么是重点调查、典型调查和抽样调查? 它们各自有哪些特点和作用?
5. 一家大型百货连锁商店的市场主管想要在某大城市进行一项调查, 以确定某个月内上班女性在购买衣服上所用的时间。

(1) 描述总体和样本, 并指出主管可能想要收集的变量类型。

(2) 草拟(1)中所需的问卷, 写出你认为适合这项调查的三个分类变量和三个数值变量问题。

6. 盖洛普公司会在其网站 [www.gallupoll.com](http://www.gallupoll.com) 上发布最近的一些投票结果。登录该网站并阅读今天的头条分析。

(1) 给出一个投票中的分类变量例子。

(2) 给出一个投票中的数值变量例子。

(3) 分析在(2)中选择的变量是离散变量还是连续变量?

7. 使用下面的问卷调查, 调查对象包含 50 名本科学生, 请他们回答下列问题:

(1) 你的性别?     女     男

(2) 你的年龄(以最后一个生日为准)?

(3) 你有多高(厘米)?

(4) 你现在注册的年级是几年级?

一年级     二年级     三年级     四年级

(5) 你学习的主要专业是什么?

会计     经济/金融     信息系统     国际贸易  
管理     市场营销/零售     其他     未定

(6) 此时你计划升学进入研究生院吗?

是     否     不确定

(7) 你现在的总平均绩点是多少?

(8) 如果你在获得学士学位后立刻找工作, 期望年薪起薪(千元)是多少?

(9) 你期望自己在拥有五年的工作经验之后的薪水(千元)是多少?

(10) 你现在的雇佣状况是什么?

全职     兼职     无业

(11) 你在校园内现在参加了多少俱乐部、小组、组织或团队?

(12) 你对校园学生广告服务的满意度如何?

非常满意     满意     中等     不满意

(13) 你本学期在教科书和日用品方面花费了多少钱?

请将调查的结果保存在 Excel 文件中, 并明确:

(1) 调查中哪些是分类变量?

(2) 调查中哪些是数值变量?

(3) 哪些变量是离散数值变量?

8. 对于智商、肤色、社会地位、体重、温度、职业、民族、职业声望、文化程度、态度、收入水平, 分别说出它们所能达到的最高度量水平是什么。