大数据丛书系列之六

总主编◎曾　羽　龙奋杰

大数据分析
——R语言方法

DASHUJU FENXI
R YUYAN FANGFA

主　编◎李云冀

电子科技大学出版社

大数据丛书系列之六

总主编◎曾　羽　龙奋杰

大数据分析
——R语言方法

DASHUJU FENXI
R YUYAN FANGFA

主　编◎李云冀

电子科技大学出版社

图书在版编目(CIP)数据

大数据分析：R语言方法 / 李云冀主编. -- 成都：电子科技大学出版社，2017.7
ISBN 978-7-5647-4827-2

Ⅰ. ①大… Ⅱ. ①李… Ⅲ. ①数据处理软件－程序语言－程序设计 Ⅳ. ①TP274②TP312

中国版本图书馆CIP数据核字（2017）第175664号

大数据分析——R语言方法

李云冀　主编

策划编辑　杨仪玮　李燕芩
责任编辑　李燕芩

出版发行　电子科技大学出版社
　　成都市一环路东一段159号电子信息产业大厦　邮编 610051
主　　页　www.uestcp.com.cn
服务电话　028-83203399
邮购电话　028-83201495

印　　刷　成都市火炬印务有限公司
成品尺寸　165mm × 240mm
印　　张　20.5
字　　数　336千字
版　　次　2017年7月第一版
印　　次　2017年7月第一次印刷
书　　号　ISBN 978-7-5647-4827-2
定　　价　69.00元

目　录

第 1 篇　统计学基础

第2篇 R语言及其应用

第1篇

统计学基础

第1章　统计数据的搜集与整理

1.1　数据的计量与类型

大数据研究与应用离不开数据，统计数据是很重要的一个数据来源。统计数据是对客观现象进行计量的结果。对统计数据的属性、特征进行分类、标示和计算，称为统计测定统计研究客观事物的数量方面，离不开统计数据，或统计度量，统计度量有定性和定量测定之别，并且可分不同的层次。根据计量学的一般分类方法，按照对事物计量的精确程度，可将所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。采用不同计量尺度可以得到不同类型的统计数据，而不同类型的逃难数据又适用于不同的统计分析方法。

1.1.1　数据的计量尺度

(1)定类尺度

定类尺度是最粗略、计量层次最低的计量尺度。它按照事物的属性进行平行的分类或分组。定类尺度只能区分事物是同类或不同类，因此它具有等于或不等于的数学特性。

(2)定序尺度

定序尺度是对事物之间等级差或顺序差别的一种测度。它不仅可以将事物分成不同的类别，还可以确定这些类别的优劣或顺序。因此该尺度有大于和或者小于的数学性质，其计量结果不仅能对事物分门别类，还可以比较大小。

(3)定距尺度

定距尺度不仅能区分事物的类别和进行排序，还可以准确地指出类别之间的差距是多少。它是对事物类别或次序之间间距的测度，它不仅有定类和定序尺度的特性，其结果还可以进行加、减运算。

(4)定比尺度

定比尺度计量的结果也表示为数值。它除了要满足上述三种计量尺度的

全部特性外，还可以计算两个测度值之间的比值。这就要求定比尺度中必须有一个绝对固定的“零点”，这也是它与定距尺度的唯一差别。因此，采用定比尺度计量的结果通常不会出现“0”值。现实生活中，大多数情况下使用的都是定比尺度。

定距尺度和定比尺度的区别可以理解为：定距尺度是从桌面上开始测量高度，定比尺度则是从地面上开始测量高度。定比尺度中由于“0”表示不存在，因而其数值不仅可以比较大小、计算差值，还可以计算数值之间的比值。它可以进行加、减、乘、除运算，如表 1-1 所示。

表 1-1　不同类型数据进行的可进行的运算

	定类尺度	定许尺度	定矩尺度	定比尺度
$=$ $\neq$	√	√	√	√
$>$ $<$		√	√	√
$+$ $-$			√	√
$\times$ $\div$				√

1.1.2　数据的类型与分析方法

(1)数据类型与分析方法

统计数据是采用某种计量尺度对事物进行计量的结果，采用不同的计量尺度会得到不同类型的统计数据，有以下四种类型。

定类数据：表现为类别，不区分顺序，是有定类尺度计量形成的。

定序数据：表现为类别，但有顺序，是有定序尺度计量形成的。

定距数据：表现为数值，可进行加、减运算，是由定距尺度计量形成的。

定比数据：表现为数值，可进行加、减、乘、除运算，是由定比尺度计量形成的。

前两类数据说明的是事物的品质特征，称为定性数据或品质数据；后两类数据说明的是现象的数量特征，能够用数值来表现，称为定量数据或数量数据。

(2) 变量及其类型

在统计中，把说明现象某种特征的概念称为变量，变量的具体表现称为变量值。统计数据就是统计变量的具体表现，变量可分为以下几种类型。

①定类变量:变量是由定类数据来记录则称为定类变量。

②定序变量:变量是由定序数据来记录则称为定序变量。

③数字变量:变量是由数量数据来记录则称为数字变量。

1.1.3 统计调查的种类

(1)按调查的组织方式,分为统计报表和专门调查。

(2)按调查对象的范围,分为全面调查和非全面调查。全面调查即对调查对象的全部单位无一例外的进行调查;非全面调查即对调查对象中的一部分单位进行调查,包括抽样调查、重点调查、典型调查和非全面统计报表。

(3)按调查登记时间连续与否,分为经常性调查和一次性调查。经常性调查是指随着被研究现象的变化,连续不断地进行登记,以取得这些现象在一段时期内发展的总量。一次性调查是指对被研究现象每间隔一般相当长的时间所进行的登记,以取得这些现象在一定时间状况上的总量。

(4)按搜集数据的方法,分为直接观察法、凭证(报告)法、询问(采访)法。直接观察法由调查人员亲临现场对被调查单位进行观察、点数、计量;凭证法是以各种原始记录和核算凭证为基础,依据统一的表格形式和要求,按照隶属关系逐级向有关部门提供统计数据的方法;询问法只指派调查员对被调查者询问、采访,提出所要了解的问题,根据被调查者的答复来搜集统计数据的方法。

1.2 统计数据的搜集

统计数据主要来源于两种渠道:一是直接的调查和科学实验;二是别人调查和科学实验。

1.2.1 统计数据的直接来源

统计数据的直接来源主要有两个渠道:专门组织的调查和科学试验。

(1)统计调查方式

统计调查是获得直接统计数据的重要手段。常用的统计调查方式有以下几种。

1)普查。普查是为某一特定目的而专门组织的一次性全面调查,它是使用于特定目的、特定对象的一种调查方式。

2)统计报表。统计报表是统计数据的一种重要形式。统计报表是按国家有关法规的规定,自上而下地统一布置、自下而上地逐级提供基本统计数据的一种调查方式。

3)抽样调查。它是实际中应用最广泛的一种调查方式和方法,它是从调查对象的总体中随机抽取一部分单位作为样本进行调查,并根据样本调查的结果来推断总体数量特征的一种非全面调查。

4)重点调查。它是专门组织的一种非全面调查,它是在调查对象中只选择一部分重点单位所进行的调查,借以了解总体的基本情况。

5)典型调查。它是根据调查研究的目的和要求,在对总体进行全面分析的基础上,有意识地选择其中有代表性的典型单位进行深入细致的调查,借以认识事物的本质特征、因果关系和发展变化的趋势。

(2)数据的搜集方法

数据的搜集方法即统计调查方法可分为两大类:询问调查和观察实验。

1)询问调查。调查者与被调查者直接或间接接触以获得数据的一种方法。具体包括以下几种。

a. 访问调查:访问调查是调查者与被调查者通过面对面地交谈从而得到所需统计数据的调查方法。

b. 邮寄调查:是通过邮寄或宣传媒体等方式将调查表或调查问卷送至被调查者手中,由被调查者填写,然后将调查表寄回或投放到指定收集点的一种调查方法。

c. 电话调查:是调查人员利用电话同受访者进行语言交流,从而获得信息的一种方式。它具有时效快、费用低的特点。

d. 电脑辅助调查:电整个调查过程,包括问卷的设计和显示、样本设计、数据处理等也多可以由电脑来控制和完成。

e. 座谈会:也称为集体访谈法,是将被调查者集中在调查现场,让他们对调查的主题发表意见,从而获取调查数据的方法。

f. 个别深度访问:深度访问是一种一次只有一名受访者参加的特殊的定性研究。它要求不断深入的受访者的思想当中,努力发掘其行为的真实动机。是一种无结构的个人访问。

2)观察与实验。观察与实验是调查者通过直接的观察或实验获得数据的一种方法。

a. 观察法:是指就调查对象的行动和意识,调查人员边观察边记录以收集信息的方法。

b. 实验法:是一种特殊的观察调查方法,它是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数据的一种调查方法。

1.2.2 统计数据的间接来源

统计数据的间接来源指通过其他渠道获取别人调查或科学实验的第二手数据。第二手数据主要是公开出版的或公开报道的数据,当然也可以是尚未公开的数据。还可以是从网络上获取的数据。

使用第二手数据既经济又方便,但应注意统计数据的含义、计算口径和计算方法,以避免误用或滥用。同时,在引用第二手数据时,一定要注明数据的来源,以尊重他人的劳动成果。

1.3 统计调查方案

为了使调查得以顺利的实施和完成,在进行统计调查之前,需要制定一个周密、完整的调查方案。一个完整的调查方案,至少应考虑以下几个方面的问题(五个"W",即 why, who, what, where, when, 一个"H",即 how),即为什么进行调查、向谁调查、调查什么、何时调查、调查何时、怎样调查。

1.3.1 调查目的

调查目的是调查所要达到的具体目标,它所回答的是为什么调查,要解决什么样的问题,调查具有什么样的意义等。

1.3.2 调查对象和调查单位

调查对象是根据调查目的确定的调查研究的总体或调查范围。调查单位是构成调查对象的每一个单位,调查对象和单位所解决的是向谁调查,由谁提供所需数据的问题。

1.3.3 设计调查项目和调查表

调查项目是回答调查什么的问题,调查项目是调查的具体内容,可以是调

查单位的数量特征，也可以是调查单位的某种属性或品质特征。

调查项目通常以表格的形式来表现，称为调查表。它是用于登记调查数据的一种表格，一般由表头、表体和表外附加三部分组成。表头是调查表的名称，用来说明调查表的内容、被调查单位的名称、性质、隶属关系等；表体是调查表的主要部分，包括调查的具体项目；表外附加通常有填表人签名、填表日期、填表说明等内容组成。

1.4 统计数据的质量

1.4.1 统计数据的误差

统计数据的误差通常是指统计数据与客观真实值之间的差距，误差主要有登记性误差和代表性误差两类。

登记性误差是调查过程中由于调查者或被调查者的人为因素所造成的误差。如：调查方案中有关的规定或解释不明确导致的错误、抄录错误、汇总错误等；因人为因素干扰形成的有意虚报或瞒报调查数据这种误差在统计调查中应予以特别重视。从理论上讲，登记性误差是可以消除的。

代表性误差主要是指用样本数据进行推断时所产生的随机误差。

1.4.2 统计数据的质量要求

就一般的统计数据而言，可将其质量评价标准概括为六个方面：

a. 精度，即最低的抽样误差或随机误差；

b. 准确性，即最小的非抽样误差或偏差；

c. 关联性，即满足用户决策、管理和研究的需要；

d. 及时性，即在最短的时间里取得公布数据；

e. 一致性，即保持时间序列的可比性；

f. 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。

1.5 统计数据的整理

统计数据的整理与可视化是统计工作的一个重要环节，是统计分析的前提。

1.5.1 统计整理的意义

统计整理：指根据统计研究的目的要求，对统计调查所取得的各项数据进行科学的分组和汇总的工作过程；对已整理过的数据（包括历史数据）进行再加工也属于统计整理。

1）通过统计调查可以取得第一手数据，但这种数据是分散、零碎、表面的。要说明总体情况，揭示出总体的内在特征，还需要对这些数据进行加工整理，使之系统化，以便通过综合指标对总体做出概括性的说明。

2）统计整理是统计调查的继续，也是统计分析的基础。统计调查所搜集到的数据，只有通过科学的审核、分类、汇总等整理工作，才能使统计由特殊到一般、由现象到本质、由感性到理性的转化，才能从整体上反映出事物的数量特征。

3）统计整理还是积累历史数据的必要手段。统计研究中经常要用动态分析，对已有的统计数据进行筛选，以及按历史的口径对现有的统计数据重新调整、分类和汇总等，都必须通过统计整理工作来完成。

1.5.2 统计整理的程序

统计整理的包括对统计数据的审核、分组、汇总和编制统计图表四个环节，需要按照一定的步骤进行。

（1）对搜集到的数据进行全面审核，以确保统计数据符合统计研究目的的要求，数据准确无误。

（2）根据研究目和统计分析的需要，选择整理的指标，并进行划类分组。统计分组是统计整理的重要内容和统计分析的基础，只有正确的分组才能整理出有科学价值的综合指标，并借助这些指标来揭示现象的本质与规律。

（3）在分组的基础上，将各项数据进行汇总，得出反映各组和总体数量特征的各种指标。

（4）统计数据的显示。即通过编制统计表和绘制统计图或者利用相关的可视化软件，将整理出的数据简捷明了、系统有序地显示出来。

1.6 统计数据的审核

对调查数据进行审核是统计整理的第一步，包括以下内容。

1.6.1 审核数据的完整性和及时性

审核数据的完整性，就是看调查单位或填报单位是否齐全；规定的项目是否都有答案，应报数据的份数是否符合规定。

审核数据的及时性，是看填报单位是否按时报送了有关数据。对不报、漏报或迟报的现象都要及时查清。

1.6.2 审核数据的正确性

审核数据的正确性，是检查所填报的数据是否准确可靠。常用的审核方法有两种。

(1)逻辑检查

首先，从理论上或常识上检查数据是否有悖常理、有无不切实际或不符合逻辑的地方。其次，是检查各项目之间有无相互矛盾的地方。

(1)计算检查

即检查各项指标的计算口径、计量单位是否符合规定，并通过各种计算方法来检查各指标间的数字是否相互衔接。

1.6.3 历史数据的审核

在利用历史数据(或其他间接数据)时，应审核数据的可靠程度、指标含义、所属时间与空间范围、计算方法和分组条件与规定的要求是否一致。

1.6.4 数据审核后的订正

如在上述审核中发现有缺报、缺份和缺项等情况，应及时催报、补报；如有不正确之处，则应分别不同情况作如下处理。

(1)对于可以肯定的一般错误，应及时代为更正，并通知原报单位。

(2)对于可疑之数或无法代为更正的错误，应要求原单位复查更正。

(3)如果所发现的差错在其他单位也可能发生时，应将错误情况通报所有单位，以免发生类似错误。

(4)对于严重的错误，应发还重新填报，并查明发生错误的原因，若属于违法行为，则应依法严肃处理。

对于用大数据工具和技术进行处理时，对错误数据和缺失数据有相应的处理机制，可参见后面相关的章节。

第 2 章　统计数据的基本概念

2.1　统计分组

2.1.1　统计分组的概念

根据统计研究的目的要求，按照某个指标(或几个指标)把总体划分为若干不同性质的组，称为统计分组。

2.1.2　统计分组的原则

统计分组必须遵循两个原则：穷尽原则和互斥原则。

穷尽原则，是使总体中的每一个单位都应有组可归，或者说各分组的空间足以容纳总体所有的单位。

互斥原则，就是在特定的分组指标下，总体中的任何一个单位只能归属于某一组。

2.1.3　统计分组的作用

统计分组在统计研究中的重要作用可概括为三个方面。

(1)划分社会现象的不同类型

社会经济现象千差万别，必须根据某种指标把它们划分为性质不同的类型，以便揭示不同社会经济现象的质的差异。

(2)揭示社会现象的内部结构

从数量上反映总体内部的结构是统计研究的重要任务。总体的内部结构可体现部分与整体的关系以及各部分之间存在的差别和相互联系，反映事物从量变到质变的过程，帮助人们掌握事物的特征，认识事物的性质。

表 2-1　我国出口产品构成表(%)

年份	农副产品	农副产品加工品	工矿产品	合计
1950	57.5	33.2	9.3	100
1960	31.0	42.3	26.7	100
1970	36.7	37.7	25.6	100
1980	18.7	29.5	51.8	100
1985	17.5	26.9	55.6	100
1990	13.0	29.2	57.8	100
1995	7.3	26.2	66.5	100
2000	5.2	25.1	69.7	100

如表 2-1 所示，就从我国出口商品构成的变化，反映出我国经济发展水平和经济结构的变化。

(3)分析社会现象之间的依存关系

社会经济现象之间广泛地存在着相互依存的关系，所有这些依存关系，都可通过统计分组分析出影响因素与结果因素之间的变动规律。

2.1.4 统计分组的种类

(1)按分组的作用或目的不同，分为类型分组、结构分组和分析分组。

1)类型分组：将复杂的现象总体，划分为若干个不同性质的部分。

2)结构分组：在对总体分组的基础上计算出各组对总体的比重，以此来研究总体各部分的结构。类型分组和结构分组往往紧密联系在一起。

3)分析分组：为研究现象之间依存关系而进行的统计分组。分析分组的分组指标称为原因指标，与原因指标相对应的指标称为结果指标。原因指标不同，结果指标也会不同；同一原因指标由于分组的不同，结果指标也会不同。

(2)按分组指标的多少，可分为简单分组、复合分组和并列分组。

1)简单分组：对总体只按一个指标进行分组。例如国民生产总值按产业分为第一、第二、第三产业三组；货运量按运输方式分为铁路运输、公路运输、水陆运输、航空运输与管道运输等五组。

2)复合分组：对总体按两个或两个以上的指标进行的重叠式分组，即在按某一指标分组的基础上再按另一指标进一步分组。

3)并列分组：同时用两个或两个以上的指标，分别从不同的角度，进行不重叠的多种分组。也就是说，很多简单分组从不同角度说明同一个总体，就构成一个并列的分组体系。

(3)按分组指标的性质，分为品质分组和数量分组。

品质分组：按品质指标进行的分组，即按事物的某种属性分组。这种分组可以反映总体的构成和不同属性事物在总体中的地位和作用。

数量分组：按数量指标进行的分组。

按品质指标分组和按数量指标分组是一对重要的统计分组，统计分组方法主要是围绕这两种分组来阐述的。