

第 1 章 数据的收集与整理

1.1 数据的类型

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数
据，通常按数据的收集途径将数据进行如下分类。

1.1.1 按度量尺度分

按度量尺度，数据可分为定性数据和定量数据。

(1) 定性数据（也称计数数据，quantitative data）

度量事物进行分类的结果。数据表现为类别，用文字来表述，如性别、区域、产品分类等。假如某班学生按性别分为男、女两类，那么性别就构成了一个定性数据。

性别:女,男,男,女,男,男,女,男,女,男,⋯,女,男,女,女,男,男,女,男,女

具体见 1.2.1 节例 1.1 调查数据。

(2) 定量数据（也称计量数据，quantitive data）

度量事物的精确测度。结果表现为具体的数值，如身高、体重、家庭收入、成绩等，假如测量某班每个学生的体重，那么体重就构成了一个定量数据。

体重:67,66,83,68,70,90,70,58,63,72,⋯,65,76,71,66,65,68,65,77,70

具体见 1.2.1 节例 1.1 调查数据。

这类数据的详细分析参见王斌会编著的《数据统计分析及 R 语言编程》一书。

1.1.2 按时间状况分

这类数据为动态数列（也称时间序列数据，time series data）：是按照一定的时间间隔对某一变量在不同时间的取值进行观测得到的一组数据，反映在不同时间收集到的数据描述现象随时间变化的情况。比如，收集 2001—2015 年各季度我国各地区国内生产总值（GDP，单位：万亿元）的数据，这些数据便形成时间序列数据。

下面是 GDP 的季度数据：

季度	2001Q1	2001Q2	2001Q3	2001Q4	⋯	2015Q1	2015Q2	2015Q3	2015Q4
GDP	2.330	2.565	2.687	3.384	⋯	14.067	17.351	17.316	18.937

下面是 GDP 的年度数据：

年份	2001	2002	2003	2004	...	2012	2013	2014	2015
GDP	10.966	12.033	13.582	15.988	...	51.947	58.802	63.646	67.671

具体见 1.2.2 节例 1.3 日期数据。

这类数据的详细分析参见王斌会编著的《计量经济学模型及 R 语言应用》一书。

1.2 数据的收集

数据收集有一定的格式，当对一个观察指标测量了每一观察单位的数据时，通常以向量的形式展现，如 $\mathbf{x}$: $x_1, x_2, \dots, x_n$ 。

当对每一观察单位测量了多个指标时，通常以双向表的矩阵形式展现，即

$$\mathbf{X}: \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m$$

这里 $\mathbf{X}_j$ ($j=1, 2, \dots, m$) 为 $n \times 1$ 向量， $\mathbf{X} = (x_{ij})_{n \times m}$ ，如表 1-1 所示。

表 1-1 关系型数据库的结构化数据

id	$\mathbf{X}_1$	$\mathbf{X}_2$	...	$\mathbf{X}_m$
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	...	x_{nm}

不同领域对该数据的观察单位和指标的叫法不同：数学上称它们为行（row）和列（column）的数组或矩阵；统计学上称它们为观测（observation）和变量（variable）的数据集；数据库中称它们为记录（record）和字段（field）的数据表，人工智能中称它们为示例（example）和属性（attribute）的数据集。

为便于大家将注意力集中在如何进行数据分析，而不是将精力花在对数据的收集和输入上，本书采用一种新的数据分析策略，即通篇使用几组数据讲解如何进行数据分析。

1.2.1 横向数据的收集

这类数据通常是一个个单独的数据变量，都可单独拿来进行分析。

【例 1.1 调查数据】

某高校想给其研究生开设一门有关数据分析方面的通识课程，该校 2015 年共有研究生 2600 名，现按 2% 的比例随机抽取 52 名学生进行问卷调查，为了收集这些学生的信息，我们设计了一个简单调查表，其中前 4 项指标是为进一步的数据分析辅助用的。共调查了这些学生的 8 项指标：学生编号（按年份、学院、专业、序号排列，简记为【学号】）；学生

性别（定性变量，简记为【性别】）；学生身高（定量变量，单位 cm，简记为【身高】）；学生体重（定量变量，单位 kg，简记为【体重】）和学生个人年消费支出额（定量变量，单位千元，简记为【支出】）；开设课程的必要性（简记为【开设】）；是否学过相关课程（定性变量，简记为【课程】）；是否学过或用过何种数据分析相关软件（定性变量，简记为【软件】）。

研究生【数据分析】开课信息调查表

【学号】1510111001，.....

【性别】男，女

【身高】165

【体重】67

【支出】7.5

【开设】有必要，不必要，不清楚

【课程】编程技术，概率统计，统计方法，都学习过，都未学过

【软件】SAS，SPSS，Matlab，R，Excel，Python，No

数据由一些变量及其观测值所组成。本例共有 8 个变量：编号（定性或定量）、开设、课程、软件和性别（定性变量），以及身高、体重和支出（定量变量）等。

为了充分体现现代问卷调查的能力，我们使用问卷星设计网络化调查问卷，只要进入问卷星（<https://www.wjx.cn/>）网站便可快速设计。

下面是我们设计的网络调查问卷，可以输入网址

<https://www.wjx.top/jq/21099021.aspx>

或扫二维码进行答卷。



【数据分析】开课调查表



1. 学号 *

2. 性别 *

☐男

☐女

3.身高 (cm) * (从100到200)

4.体重 (kg) * (从40到100)

5.支出 (千元) * (最大值500)

6. 数据分析开设情况 *

有必要

不必要

不清楚

7. 课程学习情况 * [多选题]

编程技术

统计方法

都未学过

概率统计

都学习过

8. 数据分析软件 * [多选题]

Excel

SAS

Python

SPSS

R

No

表 1-2 是 52 名研究生的个人信息调查数据，按照这个数据的格式，每列为一个指标的不同观测值（变量）；而每行则称为一个观测单位（样品），它是由定量值和定性值组成的向量，每个值相应于一个变量。于是就构成了表 1-2 的数据集，该数据保存在 DaPy_data.xlsx 文档的基本数据【BSdata】表单中。有时为了方便编程运算，也可将变量名改成英文或拼音格式。

表 1-2 52 名研究生的个人信息调查数据

学 号	性 别	身 高	体 重	支 出	开 设	课 程	软 件
1510248008	女	167	71	46.0	不清楚	都未学过	No
1510229019	男	171	68	10.4	有必要	概率统计	Matlab
1512108019	女	175	73	21.0	有必要	统计方法	SPSS
1512332010	男	169	74	4.9	有必要	编程技术	Excel
1512331015	男	154	55	25.9	有必要	都学习过	Python
1516248014	男	183	76	85.6	不必要	编程技术	Excel
1516352030	女	169	71	9.1	有必要	编程技术	Excel

续表

学 号	性 别	身 高	体 重	支 出	开 设	课 程	软 件
1516171019	女	166	66	2.5	不必要	都未学过	Excel
1516391008	女	165	69	35.6	不必要	都未学过	Excel
1520395019	男	173	63	22.8	有必要	统计方法	R
1520100029	男	184	82	10.3	有必要	都学习过	SAS
1520324035	男	163	66	13.0	有必要	概率统计	Matlab
1522186005	男	162	63	9.8	有必要	都学习过	SPSS
1522160006	女	168	72	35.3	不必要	统计方法	SPSS
1522274026	女	164	66	50.5	有必要	统计方法	SPSS
1523376027	男	180	81	64.1	有必要	统计方法	Excel
1523368030	女	158	63	20.6	不清楚	都学习过	Excel
1524225006	男	179	75	5.8	有必要	编程技术	Python
1524105026	女	163	65	69.4	有必要	编程技术	Python
1524286013	男	160	62	4.8	有必要	都未学过	R
1525235027	女	168	70	8.2	有必要	都学习过	R
1525352033	男	185	83	5.1	有必要	都学习过	SPSS
1526177005	男	174	76	15.8	有必要	概率统计	Excel
1526196010	男	167	72	9.8	不清楚	统计方法	SPSS
1527173011	女	160	62	11.5	不必要	都学习过	Matlab
1527237032	女	163	65	19.4	有必要	统计方法	R
1527289024	男	155	50	10.8	有必要	概率统计	SPSS
1529107020	男	178	78	8.9	不清楚	概率统计	Matlab
1529314037	男	170	70	15.1	有必要	概率统计	SAS
1529245023	男	164	58	21.9	有必要	统计方法	Excel
1529365032	男	172	71	10.4	有必要	都学习过	SPSS
1530273031	男	178	77	35.6	不必要	统计方法	R
1530243029	男	186	87	9.5	不必要	都未学过	No
1531364037	女	171	69	7.3	有必要	都学习过	Excel
1531316038	女	156	56	52.8	有必要	统计方法	Excel
1532304031	女	166	68	47.9	不清楚	统计方法	SAS
1532208040	男	176	78	75.5	不必要	概率统计	Excel
1532292012	男	178	78	28.4	不必要	概率统计	No
1532185004	女	155	54	13.4	不清楚	编程技术	Excel
1533219013	女	163	62	11.1	不清楚	概率统计	Matlab
1533384028	男	158	60	6.1	有必要	编程技术	R
1533172017	女	167	68	27.2	不必要	都未学过	Excel
1537288004	女	173	70	19.1	不清楚	编程技术	Python
1537359035	女	174	71	17.6	不清楚	概率统计	No
1438391022	女	164	62	10.3	有必要	编程技术	Python
1538399025	男	169	65	9.5	有必要	统计方法	SAS
1438120022	男	166	70	35.6	有必要	统计方法	R

							续表
学 号	性 别	身 高	体 重	支 出	开 设	课 程	软 件
1538319004	男	175	68	44.4	不清楚	统计方法	SAS
1538254010	女	166	65	5.3	不清楚	编程技术	Python
1540294017	女	159	58	71.4	不清楚	都学习过	SPSS
1540365026	女	169	73	5.5	有必要	统计方法	Excel
1540388036	女	165	67	56.8	不必要	概率统计	SAS

1.2.2 纵向数据的收集

纵向数据是一类比较特殊的数据，这类数据也称为序列数据，它对数据的格式有一定要求，特别是时间序列数据，须注意时间序列数据的输入格式。

【例 1.2 季节数据：经济数据】

年度数据有时太过宏观，须研究季节（季度或月度）数据，以了解不同季度或月度 GDP 的变化。现从国家统计局网站（<http://data.stats.gov.cn/>）收集到 2001—2015 年每个季度我国 GDP 的数据，就形成了一个时间序列数据集，共 15 年 60 个数据，该数据存放在 DaPy_data.xlsx 文档的季度数据【QTdata】表中。2001—2015 年我国国内生产总值的季度数据如表 1-3 所示。

表 1-3 2001—2015 年我国国内生产总值的季度数据

年 份	一 季 度	二 季 度	三 季 度	四 季 度
2001	2.330	2.565	2.687	3.384
2002	2.536	2.797	2.972	3.728
2003	2.886	3.101	3.346	4.249
2004	3.342	3.699	3.956	4.991
2005	3.912	4.280	4.474	5.828
2006	4.532	5.011	5.191	6.897
2007	5.476	6.124	6.410	8.571
2008	6.628	7.419	7.655	9.702
2009	6.982	7.839	8.310	10.96
2010	8.250	9.238	9.729	12.934
2011	9.748	10.901	11.586	15.076
2012	10.837	11.963	12.574	16.573
2013	11.886	12.916	13.908	20.092
2014	12.821	14.083	15.086	21.656
2015	14.067	17.351	17.316	18.937

【例 1.3 日期数据：股票数据】

今从某证券网站（此类网站很多）收集到 2005—2017 年苏宁易购（股票代码为 002024）每个交易日的股票基本数据（包括开盘价 Open、最高价 High、最低价 Low、收盘价 Close、

成交量 Volume 及调整收盘价 Adjusted), 这是一种典型的日期时间序列数据集, 共 13 年 3180 组数据, 该数据存放在 DaPy_data.xlsx 文档的股票数据【Stock】表中。苏宁电器日交易数据如表 1-4 所示。

表 1-4 苏宁电器日交易数据

date	Open	High	Low	Close	Volume	Adjusted
2005-1-3	0.702	0.717	0.702	0.712	0	0.618
2005-1-4	0.709	0.721	0.694	0.695	10958717	0.603
2005-1-5	0.695	0.708	0.695	0.705	6165072	0.611
2005-1-6	0.702	0.706	0.696	0.696	9845971	0.604
2005-1-7	0.695	0.709	0.694	0.702	13667162	0.608
...	...	...	...	...	...	...
2017-12-25	12.73	12.74	12.25	12.38	65681626	12.38
2017-12-26	12.46	12.54	12.37	12.52	30913299	12.52
2017-12-27	12.54	12.57	12.1	12.18	53813380	12.18
2017-12-28	12.2	12.28	12.06	12.18	33692919	12.18
2017-12-29	12.18	12.33	12.14	12.29	25372331	12.29

进一步, 还可以收集股票指数的时数据、分数据、秒数据、毫秒数据和微秒数据, 这类数据就形成了高频数据, 是一种大数据, 限于篇幅, 本书将不涉及。

上述数据都是结构化数据, 随着大数据时代的来临, 出现了大量的非结构化数据, 这些数据不只是由数字构成的数据库, 还包括大量的文字、图像、影像和视频数据, 关于这类数据的分析, 限于篇幅, 将不重点介绍。

1.3 数据的管理

数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的 过程。对于一般的数据分析而言, 电子表格软件已经足以胜任分析所需要的数据管理 工作。最常用的电子表格软件有微软 Office 的 Excel 表格软件 (收费) 和金山 Office 的 WPS 表格软件 (免费)。

1.3.1 表格管理数据

如果仅做一般数据管理, 数据量不太大, 而且要求系统免费、跨平台, 那么首选的 数据管理软件应该是 WPS 表格软件 (WPS 表格是与 Excel 兼容度最高的电子表格软 件, 但 WPS 是免费的, 建议使用)。下页是采用 WPS 表格对前面的数据进行管理的 界面。

数据保存在 DaPy_data.xlsx 文档中，可在网址 blog.leanote.com/DaPy 下载该数据。

	A	B	C	D	E	F	G	H	I	J	K	L
1	学号	性别	身高	体重	支出	开设	课程	软件				
2	1510248008	女	167	71	46.0	不清楚	都未学过	No				
3	1510229019	男	171	68	10.4	有必要	概率统计	Matlab				
4	1512108019	女	175	73	21.0	有必要	统计方法	SPSS				
5	1512332010	男	169	74	4.9	有必要	编程技术	Excel				
6	1512331015	男	154	55	25.9	有必要	都学习过	Python				
7	1516248014	男	183	76	85.6	不必要	编程技术	Excel				
8	1516352030	女	169	71	9.1	有必要	编程技术	Excel				
9	1516171019	女	166	66	2.5	不必要	都未学过	Excel				
10	1516391008	女	165	69	35.6	不必要	都未学过	Excel				
11	1520395019	男	173	63	22.8	有必要	统计方法	R				
12	1520100029	男	184	82	10.3	有必要	都学习过	SAS				
13	1520324035	男	163	66	13.0	有必要	概率统计	Matlab				
14	1522186005	男	162	63	9.8	有必要	都学习过	SPSS				
15	1522160006	女	168	72	35.3	不必要	统计方法	SPSS				
16	1522274026	女	164	66	50.5	有必要	统计方法	SPSS				
17	1523376027	男	180	81	64.1	有必要	统计方法	Excel				
18	1523368030	女	158	63	20.6	不清楚	都学习过	Excel				
19	1524225006	男	179	75	5.8	有必要	编程技术	Python				
20	1524105026	女	163	65	69.4	有必要	编程技术	Python				
21	1524286013	男	160	62	4.8	有必要	都未学过	R				
22	1525235027	女	168	70	8.2	有必要	都学习过	R				
23	1525352033	男	185	83	5.1	有必要	都学习过	SPSS				
24	1526177005	男	174	76	15.8	有必要	概率统计	Excel				
25	1526196010	男	167	72	9.8	不清楚	统计方法	SPSS				

1.3.2 数据库管理数据

当分析的数据量很大时，采用电子表格类软件有很大问题，须采用数据库来管理数据表格，详见相关文献。

数据及练习

将下面的数据统一放入一个 Excel 或 WPS 电子表格中，每个 sheet 放一组，并给文档起名为 mydata1.xlsx，以备后用。

1. 某厂对 50 个计件工人某月的工资进行登记，获得以下原始资料（单位：元）。

1465, 1405, 1355, 1225, 1000, 1760, 1755, 1710, 1605, 1535,
1985, 1965, 1910, 1845, 1810, 2270, 2240, 2190, 2040, 2010,
2980, 2820, 2600, 2430, 2290, 1375, 1295, 1265, 1175, 1125,
1735, 1645, 1625, 1595, 1575, 1940, 1880, 1865, 1835, 1815,
2220, 2110, 2095, 2030, 2030, 2670, 2550, 2520, 2370, 2320,

试将这组数据输入电子表格 mydata1.xlsx 的表单中。

2. 一份调查学生抽烟与每天学习时间关系的问卷，具体数据见表 1-5。

表 1-5 部分学生抽烟与每天学习时间调查表

编 号	是否抽烟	每天学习时间
1	是	少于 5 小时
2	否	5~10 小时
3	否	5~10 小时
4	是	超过 10 小时
5	否	超过 10 小时
6	是	少于 5 小时
7	是	5~10 小时
8	是	少于 5 小时
9	否	超过 10 小时
10	是	5~10 小时

试将这组数据输入电子表格 mydata1.xlsx 的表单中。

3. 从某大学统计系的学生中随机抽取 24 人，对数学和统计学的考试成绩进行调查，数据如表 1-6 所列。

表 1-6 某大学统计系部分学生数学和统计学考试成绩

编 号	性 别	数 学	统 计 学	编 号	性 别	数 学	统 计 学
1	M	81	72	13	F	83	78
2	F	90	90	14	F	81	94
3	F	91	96	15	M	77	73
4	M	74	68	16	M	60	66
5	F	70	82	17	F	66	58
6	F	73	78	18	M	84	87
7	M	88	89	19	F	80	86
8	M	78	82	20	F	85	84
9	M	95	96	21	M	70	82
10	F	63	75	22	M	54	56
11	F	85	86	23	F	93	98
12	M	60	71	24	M	68	76

试将这组数据输入电子表格 mydata1.xlsx 的表单中。