第 3 章　数据的图表展示

内容提要

3.1 数据的预处理

3.2 分类数据的整理与展示

3.3 数值数据的整理与展示

3.4 合理使用图表

3.1 数据的预处理

数据审核

- 检查数据中的错误

数据筛选

- 找出符合条件的数据

数据排序

- 寻找数据的基本特征

数据审核

原始数据（raw data）

- 完整性审核
  - 应调查的单位或个体是否有遗漏
  - 所有调查项目是否填写齐全
- 准确性审核
  - 是否有错误
  - 是否存在异常值
    - 记录错误的异常值，纠正或剔除
    - 非记录错误的异常值，保留

说明

- 问卷缺失数据：问卷填写参考
- 缺失数据的标记
  - 分类数据：NA
  - 数值数据：0

二手数据（second hand data）

- 适用性审核
  - 弄清楚数据：来源、口径、背景材料
  - 是否符合分析研究的需要
- 时效性审核
  - 是否为最新的数据

数据筛选、数据排序

数据筛选

- 场景
  - 数据中存在大量错误
  - 存在大量不符合要求的数据
- 方法
  - 借助计算机程序
    - Excel中的数据筛选
    - 数据库中的条件查询

数据排序

- 目的
  - 发现特征或趋势
  - 检查纠错
  - 重新归类分组
  - 排序分析
- 方法
  - 分类数据：按字母，升序或降序
  - 数值数据：按大小，递增或递减

数据整理和展示：原则、步骤

原则

- 数据类型不同，处理方法不同
- **低层次**数据的整理和图示方法，也适合于**高层次**的数据
- **高层次**数据的整理和图示方法，不适合于**低层次**的数据

步骤

- 分组
- 计算各组频数
- 制作频数分布表
- 图形展示

3.2 分类数据的整理与展示

分类数据的整理

分类数据的展示

分类数据的整理：步骤与概念

基本步骤

- 列出各个类别
- 计算各个类别的频数
- 制作频数分布表

说明:

- 在Excel中，分类数据的频数计算与频数分布表可通过**数据透视表**得到

相关概念

- 频数(frequency)
  落在各类别中的数据个数
- 比例(proportion)、
  某一类别数据个数占全部数据个数的比值
- 百分比(percentage)
  - 将对比的基数作为100而计算的比值
- 比率(ratio)
  不同类别数值个数的比值
- 累积频数(cumulative frequencies)
  各类别频数的逐级累加
- 累积频率(cumulative percentages)
  各类别频率(百分比)的逐级累加

分类数据的整理：数据透视表

原始数据

表 3—3　　顾客性别及购买的饮料类型

	A	B	C	D	E	F
1	顾客性别	饮料类型	顾客性别	饮料类型	顾客性别	饮料类型
2	女	碳酸饮料	女	碳酸饮料	女	其他
3	男	绿茶	男	绿茶	女	碳酸饮料
4	男	矿泉水	男	其他	女	其他
5	女	矿泉水	女	碳酸饮料	女	果汁
6	男	碳酸饮料	男	绿茶	男	绿茶
7	男	矿泉水	男	绿茶	女	果汁
8	女	碳酸饮料	女	碳酸饮料	女	碳酸饮料
9	女	绿茶	男	碳酸饮料	女	果汁
10	男	果汁	女	绿茶	男	矿泉水
11	男	碳酸饮料	男	矿泉水	女	碳酸饮料
12	女	矿泉水	女	绿茶	女	绿茶
13	女	其他	女	碳酸饮料	女	其他
14	男	碳酸饮料	女	矿泉水	女	果汁
15	男	绿茶	男	其他	男	绿茶
16	男	碳酸饮料	男	碳酸饮料	女	其他
17	女	其他	女	果汁	女	矿泉水
18	男	矿泉水	男	矿泉水		

数据透视表

表 3—4　　不同类型饮料和顾客性别的频数分布表

	A	B	C	D
1	计数项:饮料类型	顾客性别		
2	饮料类型	男	女	总计
3	果汁	1	5	6
4	矿泉水	6	4	10
5	绿茶	7	4	11
6	其他	2	6	8
7	碳酸饮料	6	9	15
8	总计	22	28	50

分类数据的展示：条形图

定义

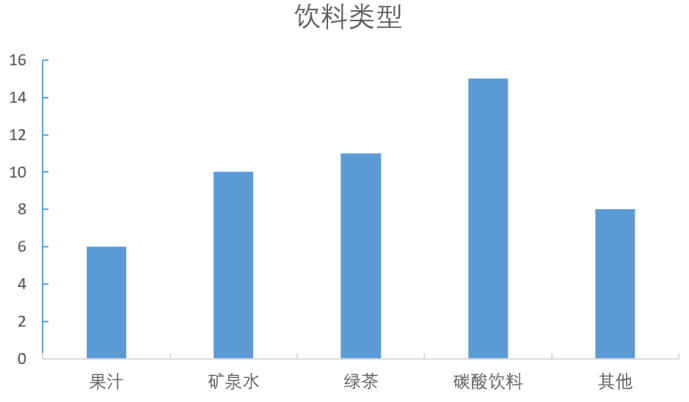
用条形的高度或长短来表示数据多少的图形。

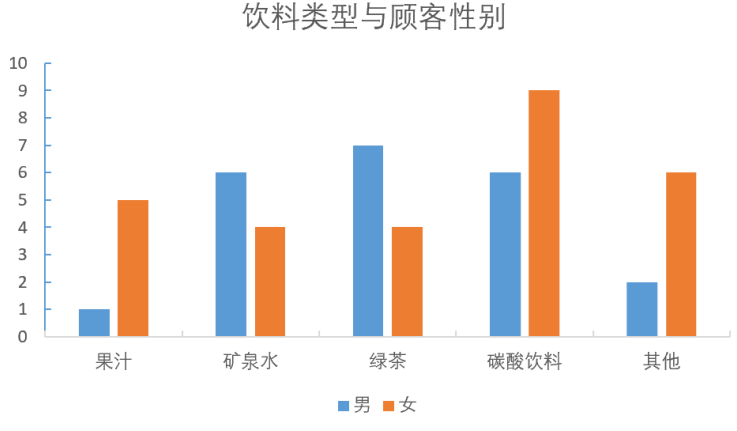
分类

- 单式条形图、复式条形图
- 条形图(横置)、柱形图(纵置)

用途

- 分类数据的频数分布
- 数值数据(分组)的频数分布





分类数据的展示：帕累托图

定义

按各类别数据出现的频数多少排序后绘制的柱形图 。

用途

- 分类数据的频数(排序)分布

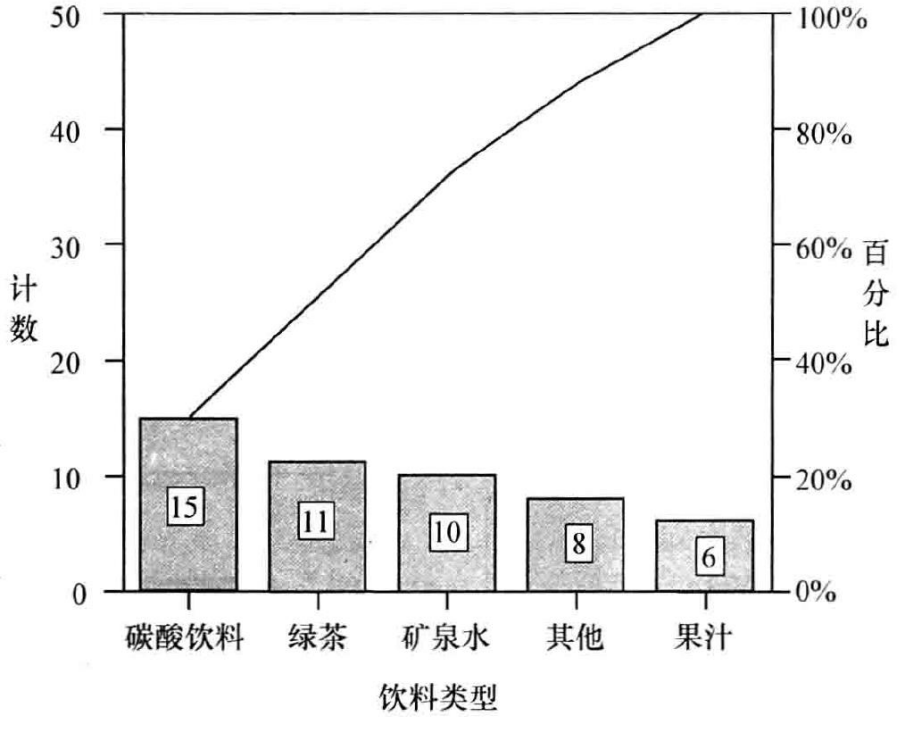


图 3—15　不同类型饮料的帕累托图

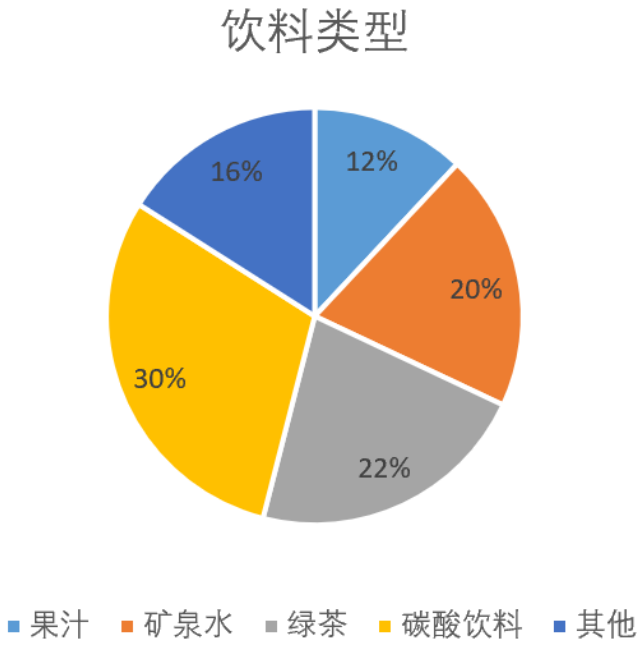
分类数据的展示：饼图

定义

用圆形及圆内扇形的角度来表示数值大小的图形。

用途

分类数据的样本(或总体)中各部分所占的**比例**，用于研究结构性问题



分类数据的展示：环形图

定义

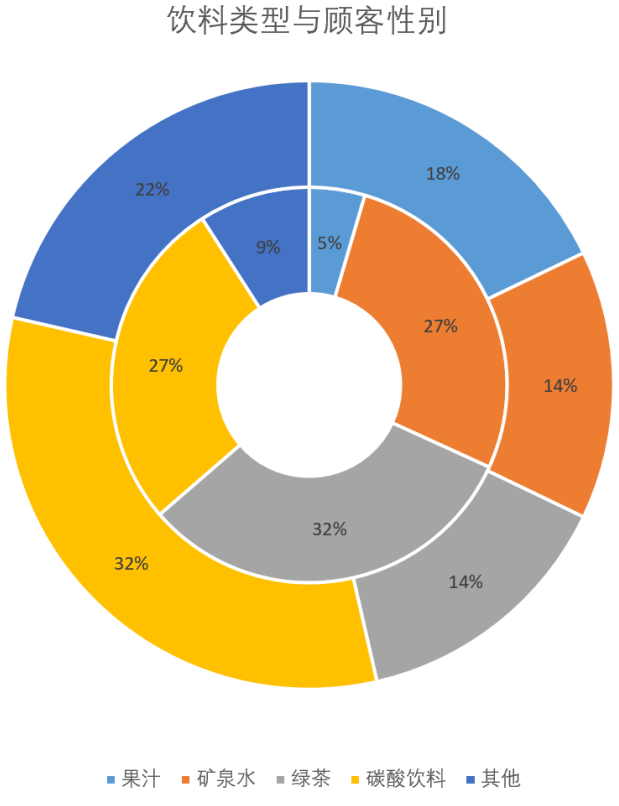
环形图中间有一个“空洞”，样本或总体中的每一部分数据用环中的一段表示。

用途

- 分类数据的多个样本(或总体)结构比较

比较

- 饼图：一个样本(或总体)各部分所占的比例
- 环形图：多个样本(或总体)各部分的比例



3.3 数值数据的整理与展示

数值数据的分组

数值数据的展示

数值数据的分组：定义、概念与步骤

定义

统计分组，是指根据统计研究的需要，将原始数据按照某种标准分成不同的组别。

概念

- 下限值(lower limit)： 一个组的最小值
- 上限值(upper limit)： 一个组的最大值
- 组距(class width)： 上限与下限之差
- 组中值(class midpoint) ： 下限与上限之间的中点值

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

分组步骤

- **确定组数**

  以显示数据的**分布特征**为目的

  在实际分组时，组数一般为5≤K ≤15

- **确定组距**

  组距 = (最大值 - 最小值)÷ 组数

- **制作频数分布表**

  在Excel中通常使用Frequency函数统计频数

数值数据的分组：要点

- **组距分组：**以变量值的某一个区间作为一组
- **连续变量 vs 离散变量**
  - 离散变量的处理
  - 例子：年龄的表示
- **等距分组 vs 不等距分组**
  - 默认为等距分组，某些情形也可采用不等距分组
  - 例子：年龄分组 1-12,13-17,18-29,30-39,40-49……

    考试成绩分组：0-59,60-69,70-79,80-89,90-100
- **“不重不漏”的原则**
  - 不重：某一数据只能属于某一组，不能重复，即“**上组限不在内**”
  - 不漏：所有数据均进入分组，不能遗漏，即**首尾的开口分组涵盖最小/大值**

数值数据的分组：例子

原始数据

表 3—12　　某电脑公司连续 4 个月的销售量

	A	B	C	D	E	F	G	H	I	J
1	234	159	187	155	172	183	182	177	163	158
2	143	198	141	167	194	225	177	189	196	203
3	187	160	214	168	173	178	184	209	176	188
4	161	152	149	211	196	234	185	189	196	206
5	150	161	178	168	174	153	186	190	160	171
6	228	162	223	170	165	179	186	175	197	208
7	153	163	218	180	175	144	178	191	197	192
8	166	196	179	171	233	179	187	173	174	210
9	154	164	215	233	175	188	237	194	198	168
10	174	226	180	172	190	172	187	189	200	211
11	156	165	175	210	207	181	205	195	201	172
12	203	165	196	172	176	182	188	195	202	213

开口分组

表 3—15　　某电脑公司销售量的频数分布表（三）

	A	B	C
1	按销售量分组(台)	频数（天）	频率（%）
2	150以下	4	3.33
3	150~160	9	7.50
4	160~170	16	13.33
5	170~180	27	22.50
6	180~190	20	16.67
7	190~200	17	14.17
8	200~210	10	8.33
9	210~220	8	6.67
10	220~230	4	3.33
11	230以上	5	4.17
12	**合计**	**120**	**100**

上下组限重叠

表 3—13　　某电脑公司销售量的频数分布表（一）

	A	B	C
1	按销售量分组(台)	频数（天）	频率（%）
2	140~150	4	3.33
3	150~160	9	7.50
4	160~170	16	13.33
5	170~180	27	22.50
6	180~190	20	16.67
7	190~200	17	14.17
8	200~210	10	8.33
9	210~220	8	6.67
10	220~230	4	3.33
11	230~240	5	4.17
12	**合计**	**120**	**100**

上下组限间断

表 3—14　　某电脑公司销售量的频数分布表（二）

	A	B	C
1	按销售量分组(台)	频数（天）	频率（%）
2	140～149	4	3.33
3	150~159	9	7.50
4	160~169	16	13.33
5	170~179	27	22.50
6	180~189	20	16.67
7	190~199	17	14.17
8	200~209	10	8.33
9	210~219	8	6.67
10	220~229	4	3.33
11	230～239	5	4.17
12	**合计**	**120**	**100**

数值数据的展示：直方图

定义

直方图（Histogram）是一种用于展示数值数据分布的一种图形，它是用矩形的宽度和高度来表示频数分布。

用途

已分组的数值型数据的分布展示

直方图 vs 柱形图

	直方图	柱形图
宽度含义	组距	类别
排列方式	连续排列	分开排列
适用对象	数值型数据	分类数据

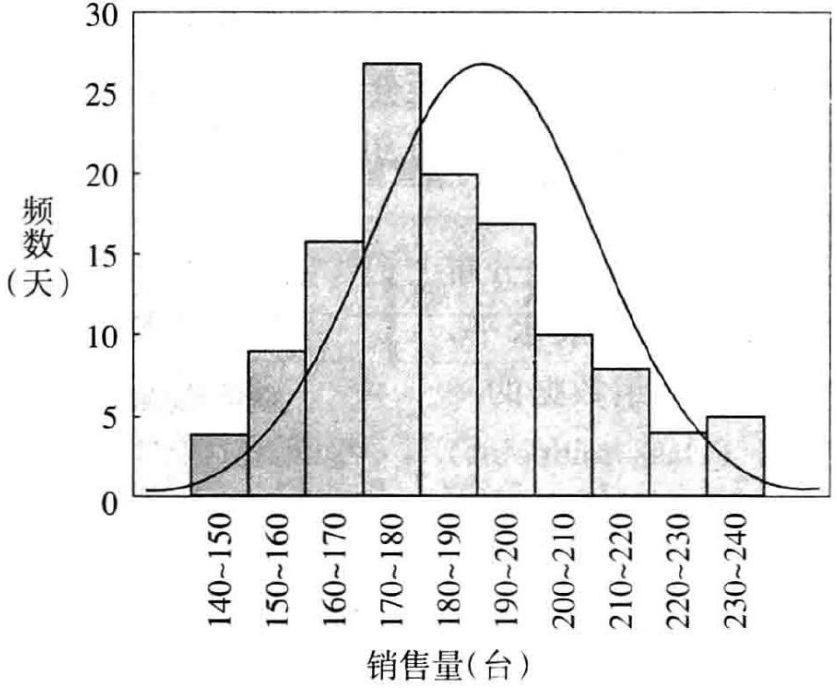


图 3—20　某电脑公司销售量分布的直方图

数值数据的展示：箱线图

用途

未分组的数值型数据的分布展示

绘制方法

- 找出未分组数据的**5个特征值**：最大值、最小值、中位数Me、下四分位数和上四分位数
- 连接两个四分位数画出箱子
- 再将两个极值点与箱子相连接

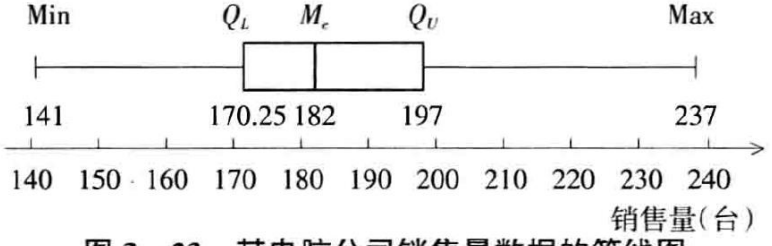
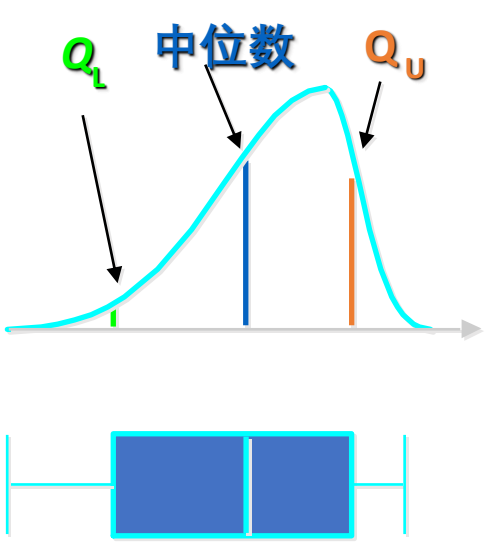


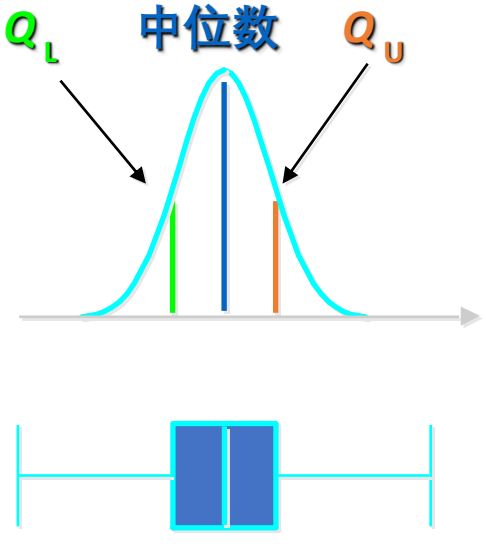
图 3—23　某电脑公司销售量数据的箱线图

数值数据的展示：箱线图与分布



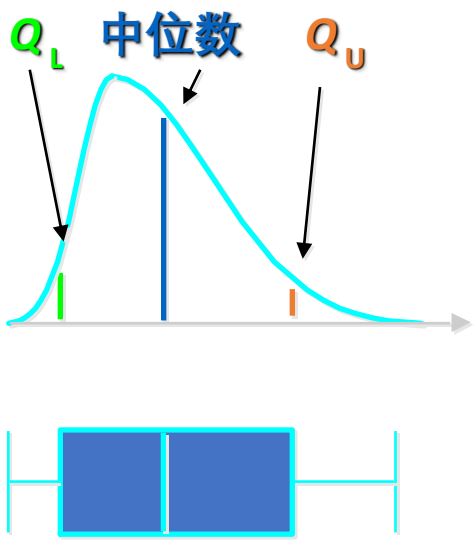
左偏分布

Left-skewed distribution



对称分布

Bell-shaped distribution



右偏分布

Right-skewed distribution

数值数据的展示：茎叶图

用途

未分组的数值型数据的分布展示

绘制方法

- 以未分组数据的高位数值作**树茎**，低位数字作**树叶**
- 树叶上只保留最后一位数字

树茎	树 叶	数据个数
14	1349	4
15	023345689	9
16	0011233455567888	16
17	01122222334445556677888999	27
18	0012234566777888999	20
19	0012445566667788	17
20	0123356789	10
21	00113458	8
22	3568	4
23	33447	5

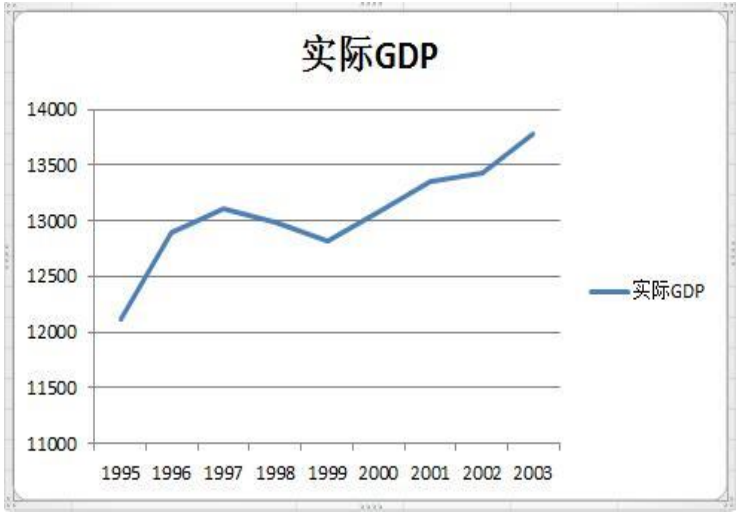
数值数据的展示：折线图

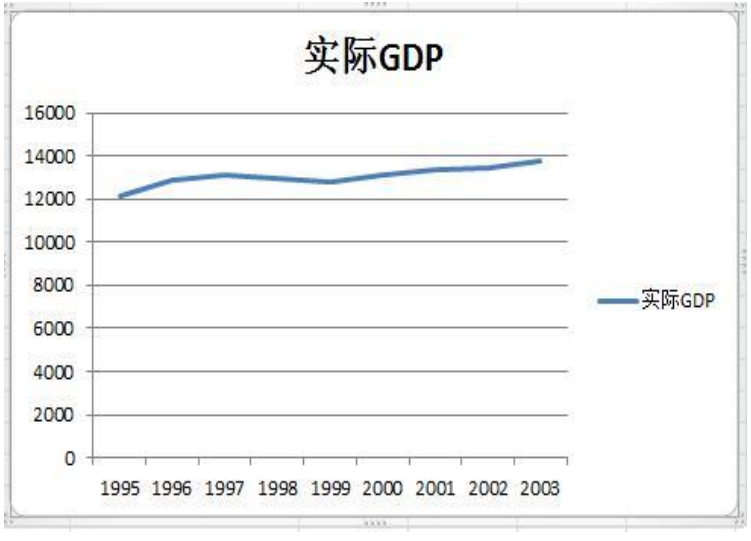
用途

展示**时间序列数据**趋势的图形

绘制方法

- 以时间为横轴，数据为纵轴
- 纵轴刻度从“0”开始
- 参考比例为 10：7





数值数据的展示：散点图

用途

展示**两个数值变量**之间的关系

绘制方法

- 横轴代表变量x，纵轴代表变量y
- 每组数据(x，y)坐标表示一个点
- n组数据的坐标形成n个散点
- 所有散点形成的二维数据图

表 3—18　　小麦产量与降雨量和温度的数据

	A	B	C
1	温度(°C)	降雨量(mm)	产量(kg/ hm^2)
2	6	25	2250
3	8	40	3450
4	10	58	4500
5	13	68	5750
6	14	110	5800
7	16	98	7500
8	21	120	8250

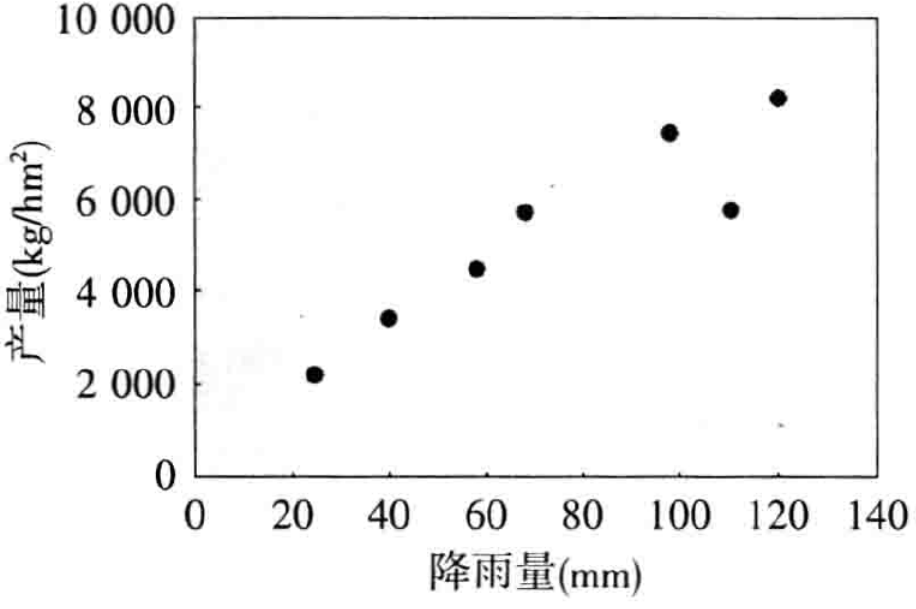


图 3—28　小麦产量与降雨量的散点图

数值数据的展示：气泡图

用途

展示**三个数值变量**之间的关系

绘制方法

- 横轴代表变量x，纵轴代表变量y
- 第三个变量取值决定气泡大小

表 3—18　　小麦产量与降雨量和温度的数据

	A	B	C
1	温度(°C)	降雨量(mm)	产量(kg/ hm^2)
2	6	25	2250
3	8	40	3450
4	10	58	4500
5	13	68	5750
6	14	110	5800
7	16	98	7500
8	21	120	8250

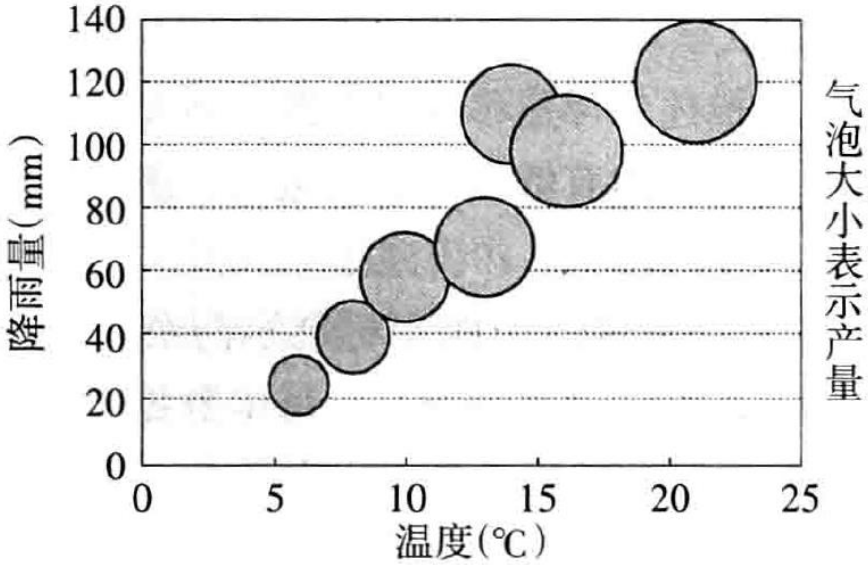


图 3—29　小麦产量与降雨量和温度的气泡图

数值数据的展示：雷达图

用途

展示**多个样本之间的**的相似程度

绘制条件

- 变量取值必须为正

表 3—19　　2003 年城乡居民家庭人均消费支出构成（%）

	A	B	C
1	项 目	城镇居民	农村居民
2	食品	37.12	45.59
3	衣着	9.79	5.67
4	家庭设备用品及服务	6.30	4.20
5	医疗保健	7.31	5.96
6	交通通讯	11.08	8.36
7	教育文化娱乐服务	14.35	12.13
8	居住	10.74	15.87
9	杂项商品与服务	3.30	2.21

资料来源：《中国统计年鉴 2004》，359 页，北京，中国统计出版社，2004。

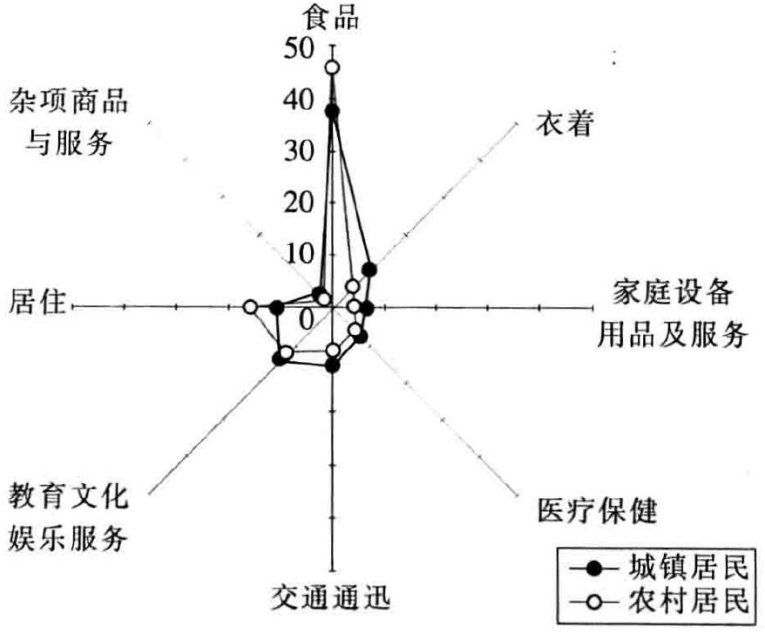
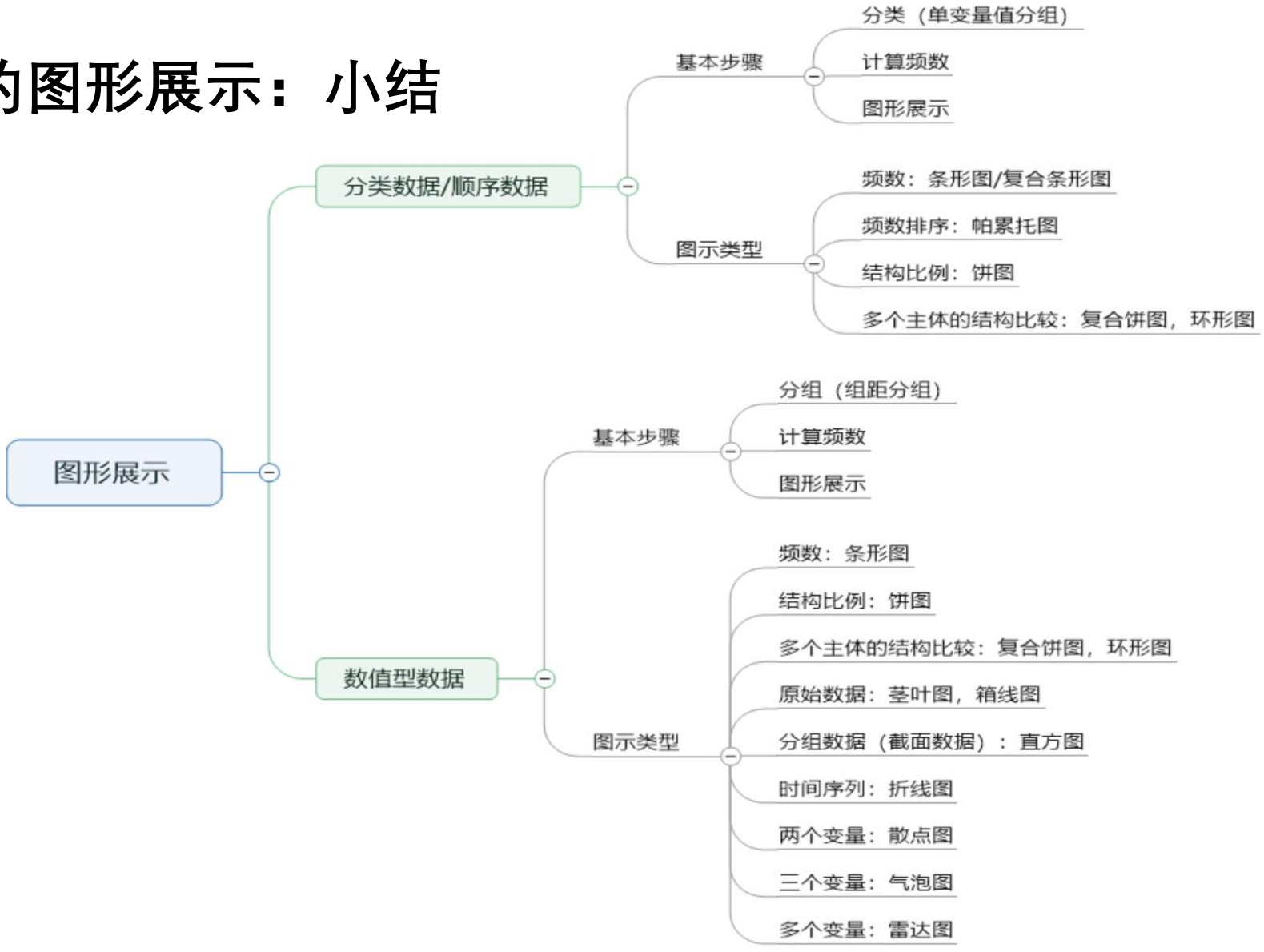


图3—30　2003 年城乡居民家庭人均消费支出构成的雷达图

数据的图形展示：小结


图形展示
分类数据/顺序数据
基本步骤
分类（单变量值分组）
计算频数
图形展示
图示类型
频数：条形图/复合条形图
频数排序：帕累托图
结构比例：饼图
多个主体的结构比较：复合饼图，环形图
数值型数据
基本步骤
分组（组距分组）
计算频数
图形展示
图示类型
频数：条形图
结构比例：饼图
多个主体的结构比较：复合饼图，环形图
原始数据：茎叶图，箱线图
分组数据（截面数据）：直方图
时间序列：折线图
两个变量：散点图
三个变量：气泡图
多个变量：雷达图

3.2 合理使用图表

表格展示

图表展示的准则

表格展示：数据透视表

作用

- 从复杂的数据中提取有用的信息
- 按使用者的习惯或分析要求，对重要信息进行分类汇总

实质

根据需要分类汇总的交叉表(列联表)

表 3—4　　不同类型饮料和顾客性别的频数分布表

	A	B	C	D
1	计数项:饮料类型	顾客性别		
2	饮料类型	男	女	总计
3	果汁	1	5	6
4	矿泉水	6	4	10
5	绿茶	7	4	11
6	其他	2	6	8
7	碳酸饮料	6	9	15
8	总计	22	28	50

表格展示：统计表

统计表(三线表)设计的原则

- **表头：**内容应满足3W 要求
- **单位：**相同时，放在表的右上角
  不同时，单列标明
- **线条**：上下两条横线一般用粗线
  其他线用细线
- **版式**：左右两边不封口
- **对齐**：通常右对齐，有小数点时应以小数点对齐，小数点的位数应统一
- **缺失数据标注**：用“—”表示
- **注释**：表的下方

表 3—20　　2002—2003 年城镇居民家庭抽样调查资料　　←表头

项　目	单位	2002 年	2003 年
调查户数	户	45 317	48 028
平均每户家庭人口	人	3.04	3.01
平均每户就业人口	人	1.58	1.58
平均每户就业面	%	51.97	52.49
平均每一就业者负担人数	人	1.92	1.91
平均每人全部年收入	元	8 177.40	9 061.22
其中：可支配收入	元	7 702.80	8 472.20
平均每人消费性支出	元	6 029.88	6 510.94

←列标题　　行标题　　数据资料

注：本表为城镇居民家庭收支抽样调查资料。
资料来源：《中国统计年鉴 2004》，359 页，北京，中国统计出版社，2004。　　附加

图表展示的准则

好图表的**6个基本特征**

- 展示数据
- 让读者把注意力集中在图表的内容上，而不是制作图表的程序上
- 避免歪曲
- 强调数据之间的比较
- 服务于一个明确的目的
- 有对图表的统计描述和文字说明

鉴别图表优劣的**5条准则**

- 精心设计、有助于洞察问题的实质
- 使复杂的观点得到简明、确切、高效的阐述
- 能在最短的时间内以最少的笔墨给读者提供最大量的信息;
- 包含多个维度;
- 表述数据的真实情况。

本章小结


数据的图表展示
数据的预处理
数据审核
一手数据：完整性，准确性
二手数据：适用性，时效性
数据筛选
数据排序
字母型：升序/降序
数值型：升序/降序
图形展示
分类数据/顺序数据
基本步骤
分类（单变量值分组）
计算频数
图形展示
图示类型
频数：条形图/复合条形图
频数排序：帕累托图
结构比例：饼图
多个主体的结构比较：复合饼图，环形图
数值型数据
基本步骤
分组（组距分组）
计算频数
图形展示
图示类型
频数：条形图
结构比例：饼图
多个主体的结构比较：复合饼图，环形图
原始数据：茎叶图，箱线图
分组数据（截面数据）：直方图
时间序列：折线图
两个变量：散点图
三个变量：气泡图
多个变量：雷达图
表格展示
数据透视表：分类汇总
统计表（三线表）：期刊文献
图表展示的准则
好图的六个特征
鉴别图形优劣的五条准则