第03章 数据的图表展示

# 数据的预处理

## 数据预处理的主要内容

数据预处理包括以下主要方面：

1. 数据审核：检查数据中的错误
2. 数据筛选：找出符合条件的数据
3. 数据排序：升序和降序；寻找数据的基本特征
4. 数据透视：按需要汇总

## 数据审核

### 原始数据审核

1. **完整性审核**：
	* 检查应调查的单位或个体是否有遗漏
	* 确认所有调查项目或变量是否填写齐全
2. **准确性审核**：
	* 验证数据是否真实反映实际情况
	* 检查数据是否有错误，计算是否正确

### 二手数据审核

1. **适用性审核**：
	* 明确数据的来源、口径以及背景材料
	* 确定数据是否符合分析研究需要
2. **时效性审核**：
	* 尽可能使用最新的数据
	* 确认是否需要进一步加工整理

## 数据筛选

数据筛选的主要内容：

1. 剔除不符合要求或有明显错误的数据
2. 筛选出符合特定条件的数据

## 数据排序

数据排序的作用：

1. 按特定顺序排列数据，发现特征或趋势
2. 有助于数据检查纠错
3. 为重新归类或分组提供依据
4. 某些场合排序本身就是分析目的

## 数据透视表

数据透视表（pivot table）的主要功能：

1. 从复杂数据中提取有用信息
2. 按使用者习惯或分析要求进行汇总
3. 形成符合需要的交叉表（列联表）
4. 数据源表中首行必须有列标题

# 数据清洗

## 常用数据清洗操作

### 文本处理

1. **文本提取**：
	* 从左侧：LEFT(text, [num\_chars])
	* 从右侧：RIGHT(text, [num\_chars])
	* 从指定位置：MID(text,start\_num,num\_chars)
2. **大小写处理**：
	* 全部小写：LOWER(text)
	* 全部大写：UPPER(text)
	* 首字母大写：PROPER(text)
	* 自定义大小写：UPPER(LEFT(A2,1)&LOWER(MID(A2,2,60))

### 数据清理

1. **重复值处理**：
	* 识别重复值
	* 删除重复记录
2. **文本合并与清理**：
	* 合并文本：[Cell 1]&[Cell 2]
	* 清除空格：TRIM(text)
	* 清除非打印字符：CLEAN(text)

## 数据审核要点

1. **真实性审核**：
	* 确认数据来源
	* 验证数据真实性
2. **完整性审核**：
	* 检查样本完整性
	* 检查选项应答情况
3. **可用性审核**：
	* 检查编码完成情况
	* 确认权数可行性
	* 处理缺失值标记
4. **错误性审核**：
	* 检查样本错误
	* 检查应答人错误
	* 检查应答方式错误

## 数据变换

### 变换的目的

* 使数据更适合建模分析
* 改善数据分布特征
* 标准化数据尺度

### 主要变换方法

* 中位数变换
	+ 公式：$income\\_normalized=\frac{income\_{i}}{median\\_income}$
	+ 适用于消除地区差异
* 均值变换
	+ 公式：$income\\_byMean=\frac{income\_{i}}{\overline{income}}$
	+ 适用于中心化处理
* 标准化变换
	+ 公式：$income\\_bySd=\frac{income\_{i}−\overline{income}}{sd\\_income}$
	+ 适用于消除量纲影响
* 对数变换
	+ 公式：$ln\left(X\_{i}\right)$
	+ 适用于处理偏态分布
* 离散化
	+ 连续变量转换为离散变量
	+ 便于分类分析

# 数据子集

## 数据子集的概念

数据子集（data subset）是对数据集进行抽样（sampling）的过程，形成的数据子集一般用于后续的建模分析。主要包括：

1. **训练集**（training subset）
	* 主要用于参数估计
	* 用于得到最终估计模型
2. **测试集**（test subset）
	* 主要用于评估模型
	* 用于评估预测准确性

## 数据子集的抽取方法

1. **随机抽样**
	* 使用均匀随机分布
	* 设定训练集和测试集的比例
2. **等比例抽样**
	* 按照某个分类变量进行分层
	* 在各层中保持相同的抽样比例

# 品质数据的整理与展示

## 统计制表

### 单变量制表

* 类别分组表
* 频次表
* 频率表

### 双变量制表

* 交叉分组表
* 频次表
* 列频率表
* 行频率表
* 复合表

## 统计制图

### 条形图/柱状图

* 条形图：类别在纵轴
* 柱状图：类别在横轴
* 可用于单变量和多变量展示

### 饼图

* 用途：表示各部分所占比例
* 绘制要点：
	+ 扇形角度表示频次/百分比
	+ 进行极坐标变换
	+ 注意排列顺序和标签显示

### 环形图

* 用途：用于结构比较研究和展示分类、顺序数据
* 特点：
	+ 中间有”空洞”
	+ 可同时展示多个数据系列
	+ 每个数据系列为一个环

# 顺序数据的统计量和图表

## 主要统计指标

### 累积频数

* 定义：各类别频数的逐级累加
* 类型：
	+ 较小制累积（向上累积）
	+ 较大制累积（向下累积）

### 累积频率

* 定义：各类别频率的逐级累加
* 类型：
	+ 较小制累积频率
	+ 较大制累积频率

## 统计图表

### 累积频数/频率表

* 较小制累积表
* 较大制累积表
* 综合累积表

### 累积图

* 累积频数图
* 累积频率图
* 可按较小制或较大制绘制

# 数据分组

## 数据分组的概念和作用

### 概念

数据分组是把同质总体中具有不同特点的单位分开，从而正确地认识事物的本质及其规律性。

### 作用

1. **类型分组**：揭露社会经济现象的类型，反映各类型的特点
2. **结构分组**：说明社会经济现象的内部结构
3. **分析分组**：研究经济现象之间的依存关系

## 分组标志的选择

### 选择原则

1. 科学性
2. 完备性
3. 互斥性

### 选择方法

1. 根据研究问题的目的来选择
2. 选择最能反映被研究现象本质特征的标志
3. 结合现象所处的具体历史条件或经济条件来选择

## 数据分组的类型

### 按分组标志的特征分类

1. **品质标志分组**：反映事物属性差异
	* 简单分组：如人口按性别分组
	* 复杂分组：如人口按职业分组
2. **数量标志分组**：反映事物数量差异
	* 单项式数量分组：适用于变量变动幅度小、项目少的分组
	* 组距式分组：适用于变量变动幅度大、项目多的分组

### 按总体所选择标志的个数分类

1. **单一分组**：按一个标志对总体进行分组
2. **复合分组**：按两个或两个以上标志对同一总体进行分组

## 组距式分组的方法

### 组距式分组的要素

1. **组距**：每组上限与下限之差
2. **组界**：每组的上限和下限
3. **组限**：每组的实际界限
4. **组中值**：每组上、下限的算术平均数

### 组距的确定方法

1. **等距分组法**
	* 计算总距离：最大值与最小值之差
	* 确定组数：一般为5-15组
	* 计算组距：总距离除以组数
	* 确定组界：从最小值开始，依次加上组距
2. **不等距分组法**
	* 适用于数据分布不均匀的情况
	* 可根据实际需要设定不同的组距
	* 常用于特殊区间需要更详细分析时

### 分组注意事项

1. **组距大小的选择**
	* 不宜过大：会掩盖数据特征
	* 不宜过小：会使规律性不明显
	* 应保持整数：便于计算和分析
2. **组界的设置**
	* 避免重叠：确保数据只能归入一组
	* 保持连续：不应有遗漏区间
	* 注意实际意义：符合数据的实际含义
3. **特殊情况处理**
	* 极端值处理：可设置开放式组距
	* 零散数据：可适当合并组别
	* 缺失数据：需要单独说明处理方法

# 数值型数据的整理与展示

## 数值型数据的分组

### 分组步骤

1. **确定组数**：
	* 根据数据特点和研究目的确定
	* 一般5-15组为宜
	* 组数过多或过少都会影响分析效果
2. **确定组距**：
	* 等距分组：组距 = (最大值 - 最小值) / 组数
	* 不等距分组：根据数据分布特点确定
3. **确定组限**：
	* 第一组下限应小于或等于最小值
	* 最后一组上限应大于或等于最大值
	* 组限应避免重叠

### 分组方法

1. **单项式分组**：
	* 适用于离散型变量
	* 变量值较少的情况
	* 每个变量值作为一组
2. **组距式分组**：
	* 适用于连续型变量
	* 变量值较多的情况
	* 将变量值分成若干区间

## 数值型数据的统计图表

### 直方图

* 用途：展示连续型变量的分布特征
* 特点：
	+ 横轴表示变量值
	+ 纵轴表示频数或频率
	+ 矩形高度表示频数或频率
	+ 矩形宽度表示组距

### 折线图

* 用途：展示数据随时间变化的趋势
* 特点：
	+ 横轴表示时间
	+ 纵轴表示变量值
	+ 用线段连接各点
	+ 可展示多个数据系列

### 箱线图

* 用途：展示数据的分布特征和异常值
* 特点：
	+ 展示最小值、Q1、中位数、Q3、最大值
	+ 可识别异常值
	+ 可比较不同组别的分布特征

### 散点图

* 用途：展示两个变量之间的关系
* 特点：
	+ 横轴表示自变量
	+ 纵轴表示因变量
	+ 每个点代表一个观测值
	+ 可观察变量间的相关关系

# 高级统计图表

## 气泡图

* 用途：展示三个变量之间的关系
* 特点：
	+ 横轴和纵轴表示两个变量
	+ 气泡大小表示第三个变量
	+ 可添加颜色区分不同类别
	+ 适合展示多维度数据关系

## 雷达图

* 用途：展示多个变量的综合特征
* 特点：
	+ 也称为蜘蛛图
	+ 各变量取值具有相同正负号
	+ 图形面积与变量总和成正比
	+ 可用于比较多个样本的相似程度

# 数据分布形态

## 分布类型

1. **典型分布**：
	* 正态分布（正态曲线）
	* 偏态分布（偏态曲线）
2. **其他分布**：
	* U型分布
	* 双峰分布
	* J型分布
	* 反J型分布

## 偏度系数

偏度系数（SK）的计算公式：

$$SK=\frac{n}{\left(n−1\right)\left(n−2\right)}\sum\_{1}^{n}\left(\frac{X\_{i}−‾}{S\_{X}}\right)^{3}$$

其中： - $n$为样本数 - $S\_{X}$为样本标准差 - $S\_{X}=\sqrt{\frac{\sum\_{1}^{n}\left(X\_{i}−‾\right)^{2}}{n−1}}$

### 偏度系数的解释

1. **对称分布**：
	* $SK=0$
	* 均值等于中位数：$‾=M\_{e}$
2. **左偏分布**：
	* $SK<0$
	* 均值小于中位数：$‾<M\_{e}$
3. **右偏分布**：
	* $SK>0$
	* 均值大于中位数：$‾>M\_{e}$

## 峰度系数

峰度系数（KT）的计算公式：

$$KT=\frac{n\left(n+1\right)}{\left(n−1\right)\left(n−2\right)\left(n−3\right)}\frac{\sum\_{i=1}^{n}\left(X\_{i}−‾\right)^{4}}{S\_{X}^{4}}−\frac{3\left(n−1\right)^{2}}{\left(n−2\right)\left(n−3\right)}$$

其中： - $n$为样本数 - $S\_{X}$为样本标准差 - $S\_{X}=\sqrt{\frac{\sum\_{1}^{n}\left(X\_{i}−‾\right)^{2}}{n−1}}$

### 峰度系数的解释

1. **常峰态**：
	* $KT=0$
	* 分布形态与正态分布相似
2. **低峰态**：
	* $KT<0$
	* 分布比正态分布更平坦
3. **尖峰态**：
	* $KT>0$
	* 分布比正态分布更陡峭

# 统计制表

## 统计表的结构

1. **形式结构**：
	* 总标题
	* 横行标题
	* 纵栏标题
	* 指标数值
2. **内容结构**：
	* 主词：说明总体或总体的分组
	* 宾词：用指标数值说明总体或分组

## 统计表的特点

1. **基本特征**：
	* 开口式设计
	* 上下有基线
	* 编号系统（主词A、B、C…，宾词1、2、3…）
	* 计量单位标注
	* 空格处理（“-”表示不需要，“……”表示暂缺）
2. **设计准则**：
	* 合理安排结构
	* 总标题满足3W要求（When、Where、What）
	* 计量单位标注位置
	* 线条粗细区分
	* 数据对齐方式
	* 注释添加位置

# 合理使用图表

## 图表体系和要素

### 制图体系

1. **画布区（Background）**：
	* 整个图表的背景区域
	* 包含所有图表元素
	* 可设置背景颜色和边框
2. **绘图区（Panel）**：
	* 实际绘制数据的区域
	* 包含坐标轴和网格线
	* 可设置边框和背景
3. **网格系统**：
	* 主网格（Major Grid）：主要刻度线
	* 次网格（Minor Grid）：次要刻度线
	* 帮助定位和读取数据

### 坐标轴系统

1. **横轴（X轴）**：
	* 下横轴（Bottom）：主要横轴
	* 上横轴（Top）：次要横轴
	* 用于显示自变量或分类变量
2. **纵轴（Y轴）**：
	* 左纵轴（Left）：主要纵轴
	* 右纵轴（Right）：次要纵轴
	* 用于显示因变量或数值变量

### 图表要素

1. **标题系统**：
	* 主标题：说明图表主题
	* 副标题：补充说明
	* 轴标题：说明坐标轴含义
2. **图例系统**：
	* 位置：通常位于图表右侧或下方
	* 内容：说明不同颜色或形状代表的含义
	* 样式：与图表风格保持一致
3. **注释系统**：
	* 数据标签：显示具体数值
	* 说明文字：解释特殊点或趋势
	* 参考线：突出重要数值或范围

## 图表评价标准

### 良好图表的基本特征

1. **目的明确**：
	* 服务于特定分析目的
	* 突出关键信息
	* 避免无关内容
2. **数据展示**：
	* 准确显示数据
	* 便于数据比较
	* 突出重要趋势
3. **信息完整**：
	* 包含必要的统计描述
	* 提供适当的文字说明
	* 标注数据来源
4. **视觉设计**：
	* 重点突出数据内容
	* 避免过度装饰
	* 保持简洁清晰

### 图表优劣的鉴别准则

1. **真实性**：
	* 准确反映数据特征
	* 避免数据歪曲
	* 保持客观公正
2. **有效性**：
	* 简明阐述复杂观点
	* 高效传递信息
	* 便于理解分析
3. **设计性**：
	* 精心设计布局
	* 突出问题实质
	* 便于洞察规律

## 制图常见误区

### 数据展示误区

1. **非零起始点**：
	* 纵轴不从零开始
	* 夸大数据差异
	* 误导读者判断
2. **比例失调**：
	* 图片比例不当
	* 坐标轴比例失衡
	* 影响数据解读

### 视觉设计误区

1. **过度装饰**：
	* 过多视觉元素
	* 无关背景图案
	* 分散注意力
2. **视觉误导**：
	* 不恰当的图形选择
	* 误导性的视觉效果
	* 模糊数据本质

### 信息传达误区

1. **信息缺失**：
	* 缺少必要说明
	* 数据来源不明
	* 单位标注不清
2. **信息冗余**：
	* 重复信息过多
	* 无关细节过多
	* 影响重点突出

## 图表设计要点

### 图表类型选择

1. **饼图VS条形图**：
	* 饼图适合展示整体构成
	* 条形图更适合比较大小
	* 市场份额比较建议使用条形图
2. **线形图VS柱状图**：
	* 线形图适合展示时间序列趋势
	* 柱状图适合比较不同类别
	* 时间数据优先选择线形图
3. **平面图VS立体图**：
	* 二维图更适合观察
	* 三维图容易造成视觉干扰
	* 避免使用不必要的三维效果

### 图表元素设计

1. **颜色使用**：
	* 避免过度使用颜色
	* 选择同一色系或类比色
	* 确保颜色不影响数据解读
2. **饼图设计**：
	* 分割块数不超过5块
	* 最大块从12点钟开始
	* 按顺时针降序排列
3. **线图设计**：
	* 优先使用实线
	* 使用颜色区分不同线条
	* 避免过多虚线
4. **条形图设计**：
	* 按逻辑顺序排列数据
	* 保持适当的间距
	* 间距应为柱宽的一半

### 数据展示优化

1. **面积图**：
	* 添加透明度
	* 确保数据可见
	* 避免数据重叠
2. **散点图**：
	* 添加趋势线
	* 突出数据趋势
	* 便于理解关系
3. **气泡图**：
	* 气泡大小基于面积
	* 合理设置比例
	* 避免视觉误导
4. **热力图**：
	* 使用同一色系
	* 避免过多颜色
	* 突出数据差异

### 表格设计要点

1. **表格结构**：
	* 优先使用开放式三线表
	* 减少不必要的网格线
	* 保持简洁清晰
2. **表格格式**：
	* 合理使用条件格式
	* 统一字体和对齐方式
	* 突出重要数据