第三章 数据整理

#一、标志与数据

1. 定类标志与定类数据
   定类标志：**说明实物类别的名称**。如性别，民族，种类等
   定类数据：只能归为某一类别的非数值型数据，对事物进行分类的结果，数据表现为类别，用文字、代码和其他符号来表述。例如人口按照性别分为男女两类
2. 定序标志与定序数据
   定序标志：说明事物**有序类别的名称**。如受教育程度、产业、等级等
   定序数据：只能归于某一有序类别的非数字型数据，对事物类别顺序的测度，数据表现为类别，用文字来表述。如产品分为一等二等三等品和次品等
3. 定距标志和定距数据
   定距标志：**说明事物有序类别及差距的名称**。如分数，温度等
   定距数据：按数字尺度测量的观察值，可进行**加减运算**，但不能进行乘除运算，表现为具体的数值，对事物的精确测度。例如温度的具体数值
4. 定比标志与定比数据
   定比标志：**比定距标志更高一级的数量标志**。如年龄、体重等
   定比数据：按数字尺度测量的观察值，**可进行加减乘除运算**；有一个绝对固定的，非任意性的零点，结果表现为具体的数值，对事物的精确测度。如职工人数，身高，体重等
5. 上述数据按照从低到高排序，等级越高适用范围就越广泛，并且等级高的数据可以兼有等级低的数据的功能

#二、数据分组

1. 分组的意义
   分组是将总体中所有单位按照一定的标准区分为若干部分。
   分组的目的是：概括数据，清晰条理
   分组的时候注意将有共性的个体归为同一组，将总体内部个体的差异通过组别区分开
   分组的原则是保证不重不漏
   统计分组的关键是分组标志的选择
2. 分组标准的选择
   根据研究目的和任务来确定分组的标准：
   - 按照分组标志的性质不同分类：
     品质分组：按照品质标志进行分组，比如人口总体按照性别分组，高校教师按照职称分组
     数量分组：按照数量标志进行分组，需要**确定分组的个数**，每一组的数量界限和组限的表示等问题

- 按照分组的标志的多少不同分类：
  简单分组：分组仅按照一个标志来进行
  复合分组：分组按照两个或两个以上的标志进行，并且层叠在一起

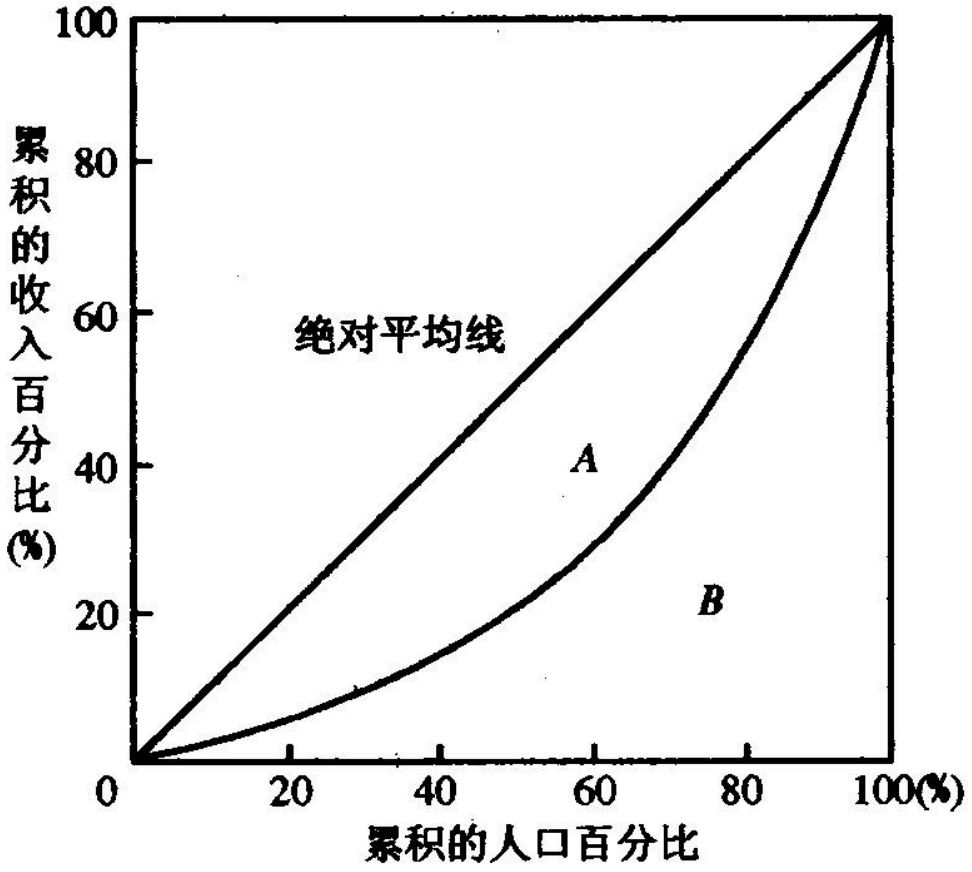
3. 次数分配
   将总体资料按照某个标志分成若干组，并且统计出各组数据个数，称这种分组结果为次数分配、次数分布或频数分布
   频数指的是各组数据的个数
   频率指的是各组次数与总次数之比
   分配方式：
   - 单项变量次数分配：依次将每一个变量值作为一组
   - 组距次数分配：将整个变量区间划分为几个子区间，各个变量值按照大小确定所归并的区间，每组区间的宽度叫做组距，区间的界限叫做组限，分为上限和下限
     包括：等距分组和不等距分组
     - 上下限
       上组限不计入原则：**遇到某单位的标志值刚好等于相邻两组上下限的时候，一般把此值归并到作为下限的那一组**
     - 全距：最大值与最小值的差
     - 组中值=(上限+下限)/2
       当最大值与最小值相差悬殊的时候设置开口组
       缺下限的开口组，出现“....以下”
       组中值=上限-邻组组距/2
       缺上限的开口组，出现“....以上”
       组中值=下限+邻组组距/2

#三、品质次数分配的编制

1. 品质次数分配的图示
   - 条形图：使用宽度相同的条形的高度或者长短来表示各类数据的图形。有单式条形图、复式条形图等形式**主要用于反映品质数据的频数分布**
   - 饼图：也称为圆形图，是用圆内扇形来表示数值大小的图形
     主要用于表示总体或样本中组成部分所占的比例，对于**研究结构性问题**十分有作用
   - 帕累托图：按照类别数据出现的频数进行排序后绘制柱形图，主要用于**展示分类数据的分布**
   - 环形图：总体的每一部分数据都可以用环中数据的一段来表示，环形图可以同时绘制多个总体的数据系列，每一个总体的数据系列为一个环，环形图可以**用于结构比较研究**

#四、变量次数分配的编制

1. 关键步骤：
   将原始资料按照顺序排序
   确定组数和组距
   当n较大的时候，k取10~20，n<50的时候，k取5~6
   如果数据分布比较均匀对称，中间数值次数多，大小极端值次数小，考虑使用**经验公式**确定组数组数 $= 1 + lgn/lg2 = 1 + 3.322lgn$
   然后组距 $=$ (观察值中的最大数值 $-$ 观察值中的最小数值)/组数
2. 分组注意事项
   分组太粗或者太细会反映不出观察值的分布特征
   绘制直方图的时候需要将频数除以所在组的组距得到频数密度，绘制分组和的密度由此得到分布图
3. 累积次数分配
   累积次数分配有向上积累和向下积累两种，分别指大于或者小于某值的累积次数
   洛伦茨曲线：绘制**累积的收入百分比与累积的人口百分比**的曲线，与绝对平均线



基尼系数 $= A/(A + B)$

#五、统计指标

1. 统计指标的概念
   统计指标是说明社会经济现象总体数量特征的名称和数值，是统计活动对客观存在的种种

社会经济现象，按照其具体的名称，在一定的空间时间条件下，进行科学计量的数字结果。

两个特征：可计量性(具体性)和总体性(综合性)

2. 统计指标的概念
   - 总量指标：反映**总体现象的规模水平，以绝对数形式表现**，故称为绝对指标，如总人口，国民生产总值等

     总量指标按其所说明的总体内容的不同，分为总体 单位总量和总体标志总量

     总体单位总量反映总体单位的总量指标如企业数目、职工人数等；

     总体标志总量反映总体各单位某一数量标志值总和的总量指标， 如商品销售额、总工资总额等；

     总量指标按其所反映的不同时间状况，分为时点总量和时期总量
   - 平均指标：将总体标志总量指标除以总体单位总量，得到平均指标

     平均指标=总体标志总量/总体单位总量
   - 相对指标：两个有联系的指标对比得到的指标都可以叫做相对指标
     - **结构**相对指标：将总体的**部分标志总量**与**总体的标志总量**相比较，或将总体 的部分单位总量与总体全部单位数相比较
     - **比例**相对指标：将总体内部的**部分与部分对比**所得到的指标
     - **动态**相对指标：将**同一内容**的指标在不同时间上的数值进行对比**。 说明现象在时间上的变化。
     - 强度相对指标：将**同一时期内容不同、但有一定联系的两个总量指标对比**。 强度相对指标常被用来说明现象的密度、普遍程度
3. 直方图与条形图的区别
   - 条形图是用条形的**长度**(横置时)表示**各类别频数**的多少，其**宽度(表示类别)**则是固定的
   - 直方图是用面积表示**各组频数**的多少，矩形的高度表示每一组的频数或百分比，宽度则表示各组的组距，其**高度与宽度均有意义**
   - 直方图的各矩形通常是连续排列，条形图则是分开排列
   - 条形图主要用于展示**分类数据**，直方图则主要用于展示**数值型数据**