Chapter 1
Measure of Central tendency and Dispersion

数据特征的描述

数据特征的描述过程

- 数据收集
- 整理
- 显示
- 描述

第一节、统计数据的搜集

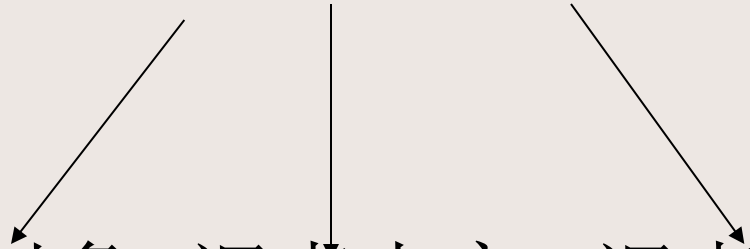
- **两种数据来源：**
- 原始数据
- 次级数据
- **两种数据形式**
- 横截面数据
- 时间数列数据

统计资料可利用组织：

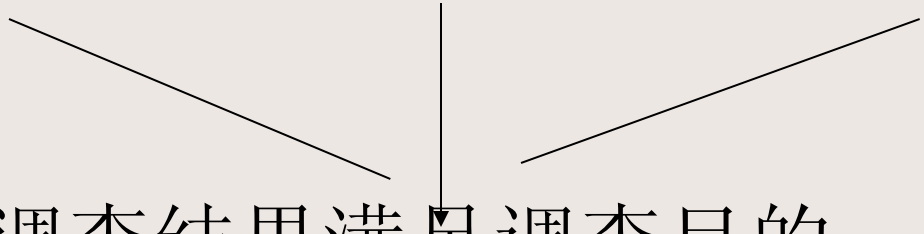
- 国际劳工组织统计局：劳动力、就业、工资
  社会保险、工会等
- 联合国教科文组织：教育、科学、文化、技
  术等
- 联合国棉农组织、卫生组织、国际货币基金
  会、世界银行等

一、基本 内容

- 调查目的



- 调查对象_ 调查内容 _ 调查方法



- 调查结果满足调查目的

二、调查方法

方法	对象	特点	适用条件
普查	全部单位	一次性、周期性、数据准确、全面、使用面窄	掌握总体情况有限总体
抽样调查	样本单位	经济、实用、准确、适应面广	掌握总体情况、有限总体与无限总体
重点调查	重点单位	非随机性	掌握趋势 存在重点单位
典型调查	典型单位	非随机性	用于定性分析
统计报表	全部单位与非全部单位	统一性、准确性	

三、调查对象

一、全面调查

不重复、不遗漏

二、非全面调查

代表性、选择偏性

引例：1936年罗斯福与兰登的总统竞选

《文学摘要》罗斯福（43%）兰登（57%）1千万

实际结果：罗斯福（62%）兰登（38%）

盖洛普：罗斯福（56%）5万人

泛法航空

四、调查内容-问卷设计

- （一）问卷结构：
- 说明词、填写要求、问卷正文及结尾
- 说明词：主办单位及调查员身份、调查的目的和意义、承诺及感谢
- 问卷正文：需要调查的问题及答案、被调查者的 背景资料
- 结尾：说明

(二)、问卷的措辞

- 清楚定义内容：5w
- “您使用什么品牌的化妆品”
- 用词通俗、词义明确
- “您经常收看电视节目吗？”“1、从来不看；2、偶尔看；3、有时看；4、经常看；5、天天看”
- 避免隐含的选择（乘车、牛仔裤）
- 避免否定形式的提问
- 避免诱导性或倾向性的词汇、避免重叠、答案详尽

（三）、问题的顺序

- 1、先易后难
- 2、封闭型问题置前，敏感性、开放性问题置后
- 3、注意对后继问题的影响：
- （1、您在选择购物时，哪些因素是重要的？
- 2、您在选择购物时，售后服务这个因素的重要性如何？）
- 4、逻辑思路保持一致

第二节、数据的整理

- 审核 → 分组 （品质数据、数
- 量数据） → 计算频数与频率 →

数据的表现

一、数据资料的可用性

- **方法错误**
- 引：时间，空间，口径等
- **逻辑错误**
- 引：产值与销售值，年龄与工作年限
- **主观错误**
- 敏感性、政治性等
- 引：失业率与平均每周申请失业保险人数

二、数据的分组与频率的计算

（一）品质数据的分组 与计算

频数：每组数据值出现的次数

表 2-5 购买 50台计算机的样本数据

IBM IBM 帕科特 • 贝尔 康柏 IBM

帕科特 • 贝尔 苹果 苹果 盖威特 -2000 帕科特 • 贝尔

康柏 康柏 苹果

…….....

表 2-6　购买计算机数据的频数分布表

按公司分组	频数
苹果机	13
康柏机	12
盖威特-2000	5
IBM	9
帕科特·贝尔	11
合计	50

（二）、数量数据的整理

数量数据频数分布的分组需要 3 个步骤： 1 、确定组数； 2、确定组距； 3 、确定组限。

引例：

表： 年终审计时间（天）

12	14	19	18	15	15	18	17	20	2
22	23	22	21	33	28	14	18	16	1

1、确定分组数目

$$K = 1 + \log_{10}^{N} \div \log_{10}^{2}$$

本例组数 = $1 + \log_{10}^{20} \div \log_{10}^{2} = 5.3$

我们确定分 5 个组。

2、确定组距

近似的组距 = $\dfrac{\text{最大数据值} - \text{最小数据值}}{\text{组数}}$

本例组距 = $\dfrac{33 - 12}{5} = 4.2$

取整数 5 天。

4、计算频数与频率

表 2-9 审计时间数据频数分布

按审计时间分组（天）	频　数
10-14	4
15-19	8
20-24	5
25-29	2
30-34	1
合　计	20

表 2-10　审计时间数据的相对频数和百分比频数分布

按审计时间分组（天）	相对频数	百分比频数
10-14	0.20	20
15-19	0.40	40
20-24	0.25	25
25-29	0.10	10
30-34	0.05	5
合　计	1.00	100

表 2-11　审计时间数据的累积频数分布

按审计时间分组（天）	频 数	向上累计频数分布	向下累计频数分布
10-14	4	4	20
15-19	8	12	16
20-24	5	17	8
25-29	2	19	3
30-34	1	20	1
合　计	20	—	—

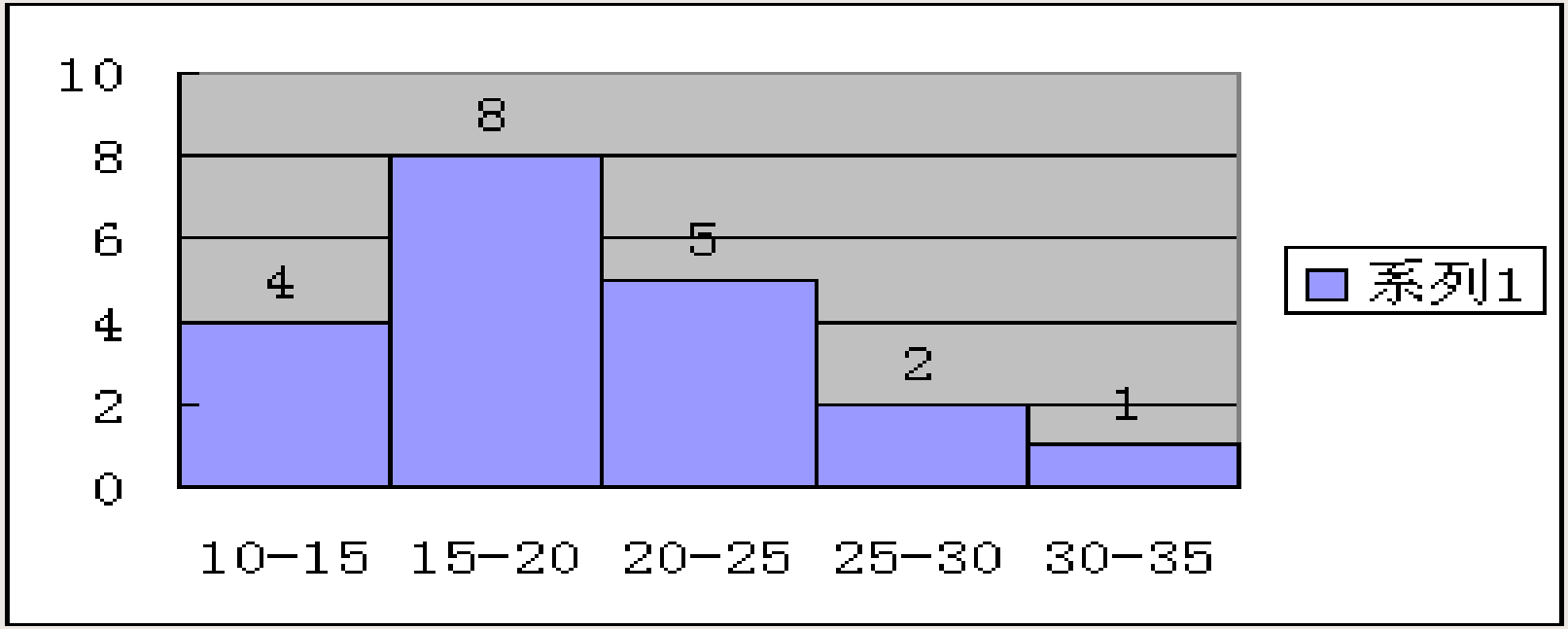
在数量数据整理中要注意的问题有：

- 1、在一些应用中，我们需要知道各分组的中点，也就是组中值。
- 2、开口组（即只有上限或只有下限的组），其组中值用邻组的组距计算。
- 3、在数据较少的情况下，可用品质数据整理的方式，采取单变量值分组。
- 4、连续变量与离散变量的组限问题 上组限不在内
- 5、等距与不等距分组

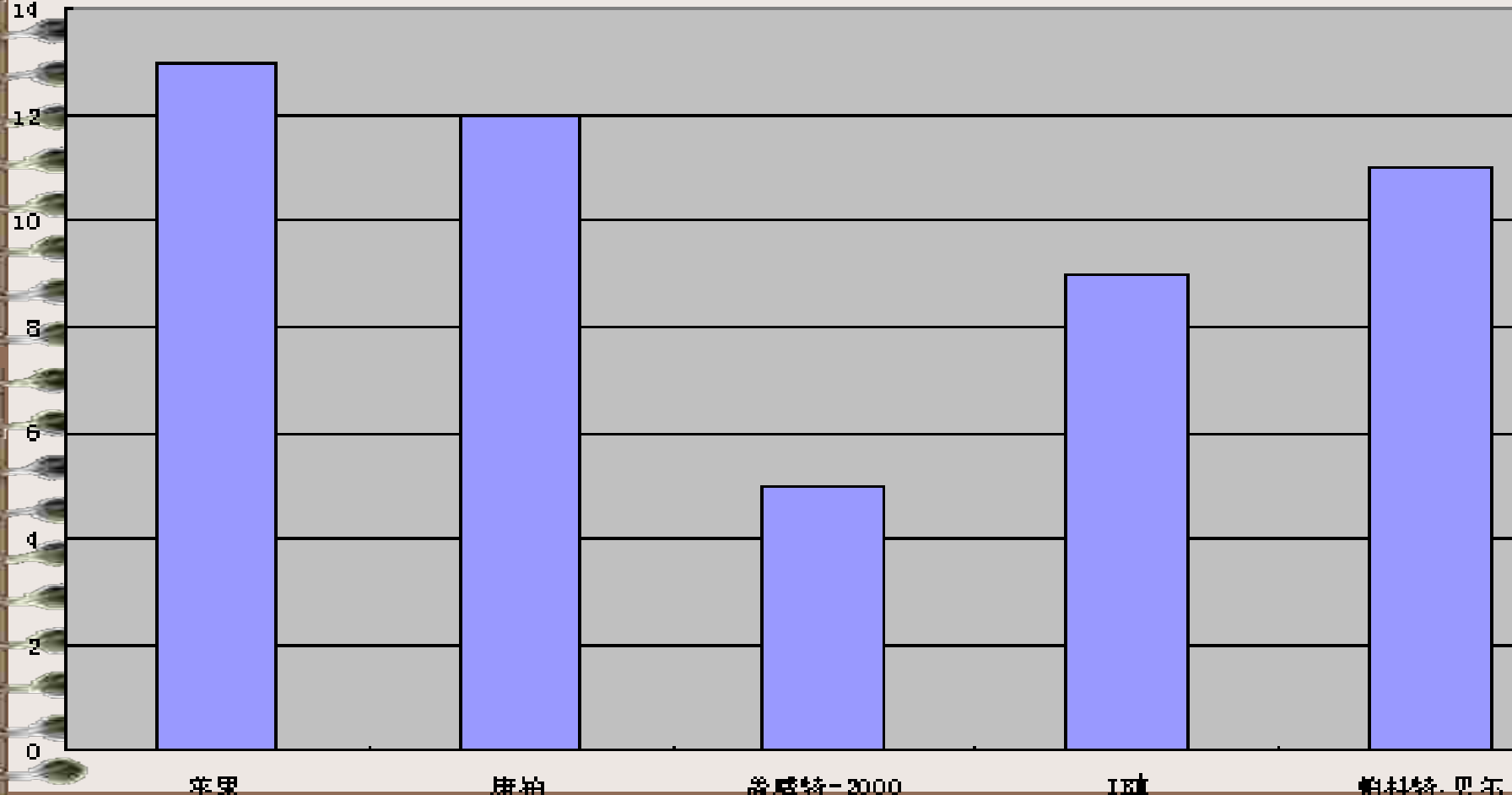
三、数据的表现

- 统计图
- 统计表
- 统计指标

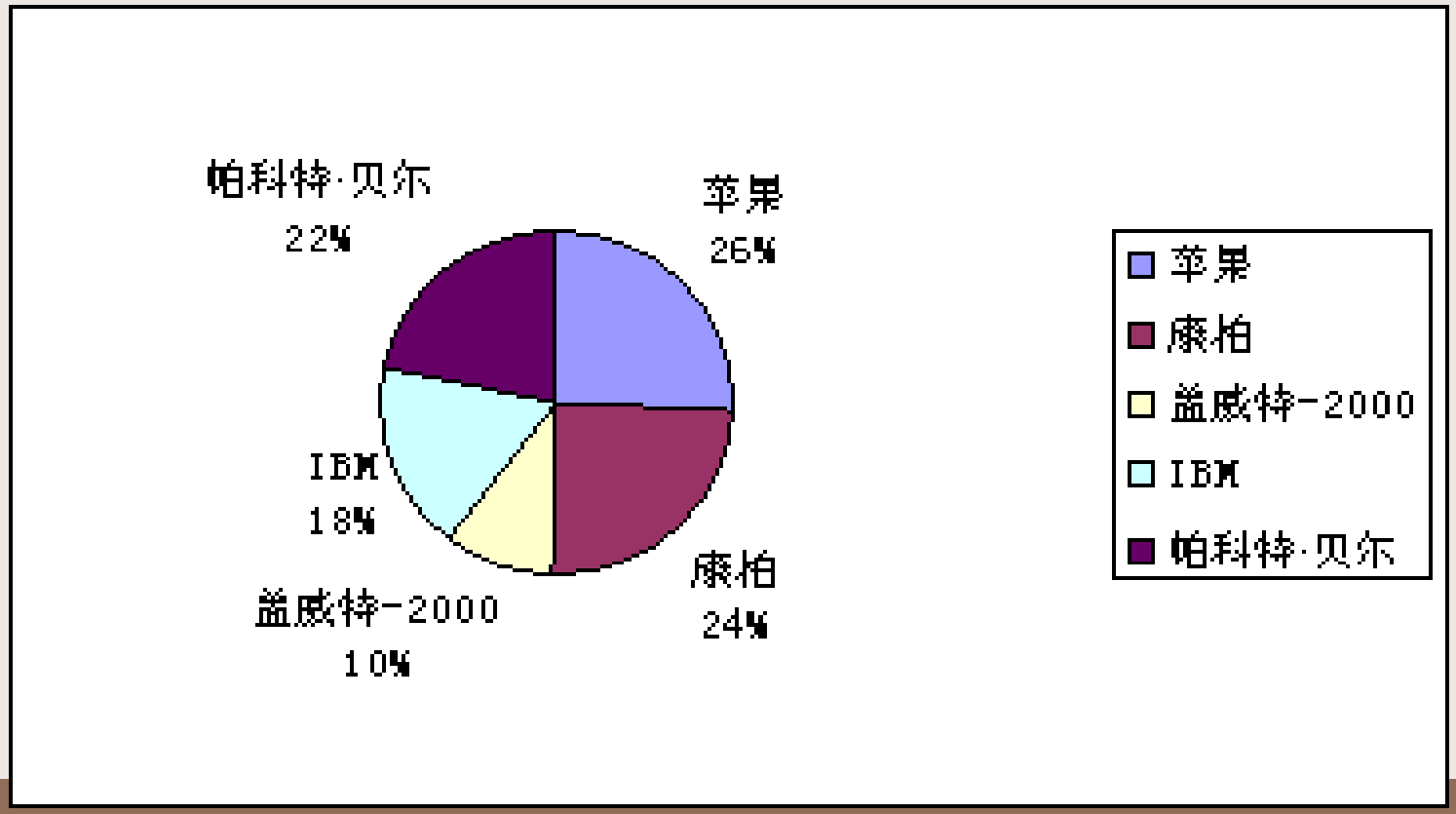
（一）统计图



2、条形图 是用图的方式描述已概括成频数、相对频数或百分比频数分布的品质数据的图形



3、饼图 是用圆的各部分面积来呈现品质数据的常用方法。本
所有各组计算机购买的百分比频数总和为 100，一个圆有 360
则饼图中苹果机的部分为 26%′360°=93.6°，其他组的部分
类推算出



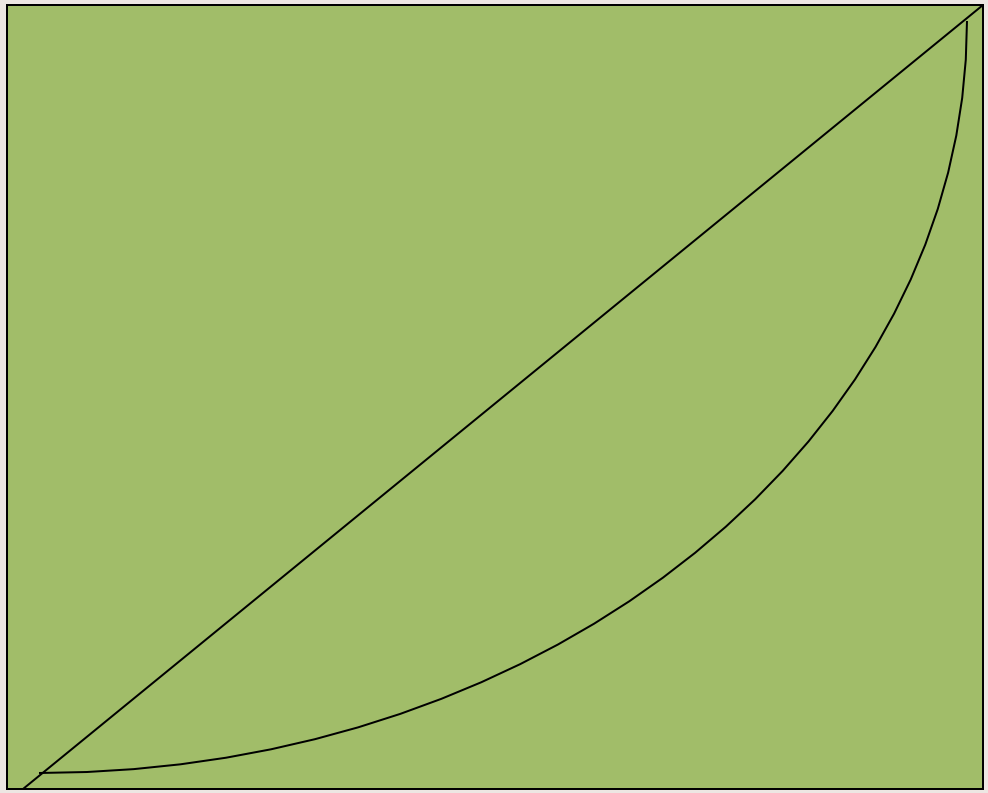
4、统计折线图与曲线图

- 洛伦茨曲线
- 生命曲线
- 投机需求曲线
- 质量曲线

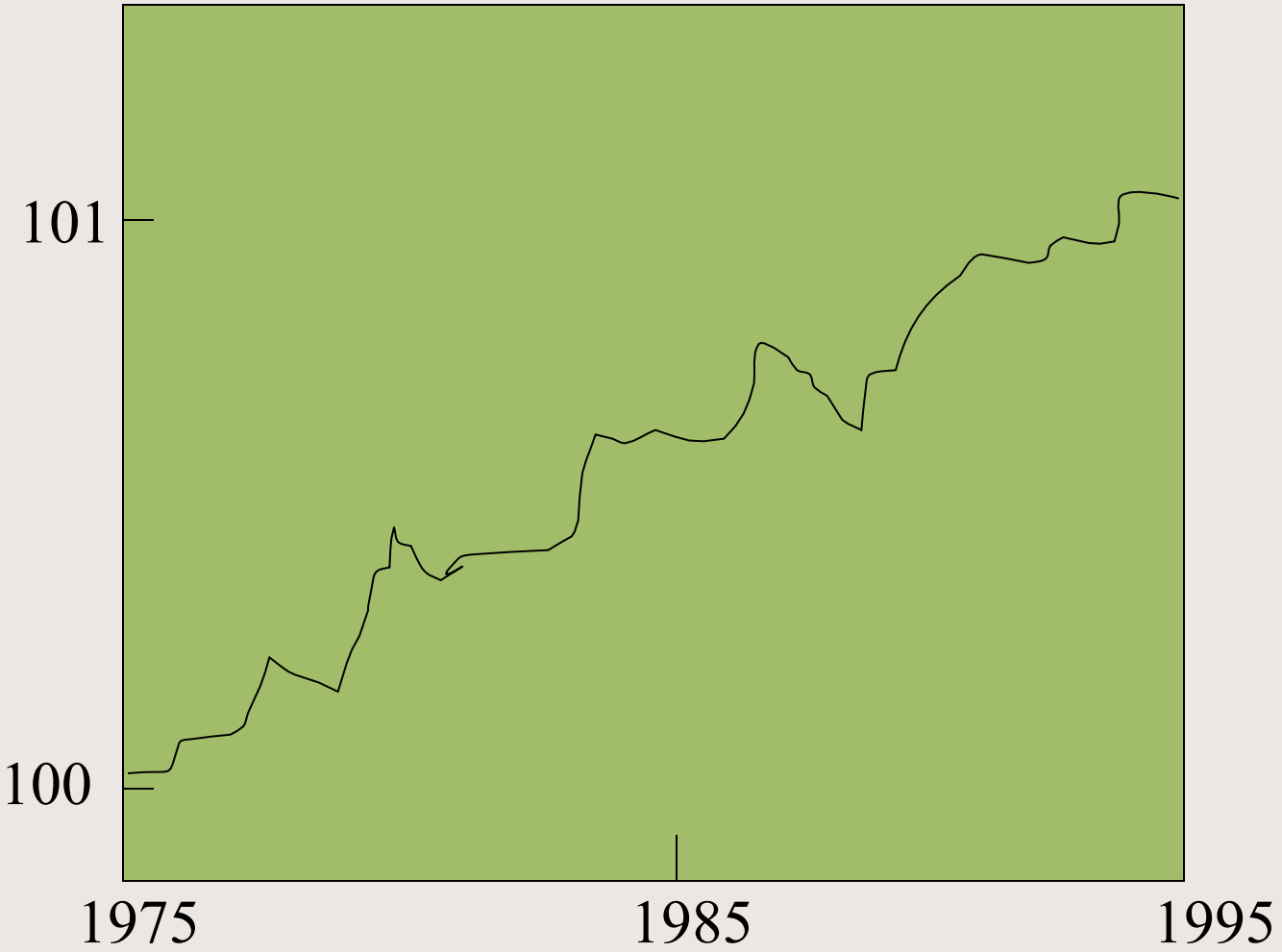
5、象形图

洛伦茨曲线

按收入大小顺序排列的家庭数	占总收入的%	累计家庭数%	累计收的%
最低的20%	4.7	20	4.7
第二个20%	11	40	15.7
第三个20%	17	60	32.4
第四个20%	24.4	80	56.8
最高的20%	43.2	100	100


累计收入的%
累计家庭的%

Sales of Chicago Carpet World since 1975



Sales u

（二）、统计表

年份	国内生产总值	最终消费	最终消费率	年末人口）
1989				112704
1990	18319.5	11365.2	61.3	114333
1991	21280.4	13145.9	60.8	115838
1992	25863.6	15952.1	59.9	117171
1993	34500.6	20182.1	58.3	118517
1994	47110.9	27216.2	58.2	119850
1995	59404.9	34529.4	59.0	121121
1996	68498.2	40171.7	58.6	122389
合计	274978.1	162562.6		

第三节、数据特征的描述

- 绝对数与相对数
- 集中趋势：
- 众数、中位数、平均数
- 离散趋势：
- 全距
- 方差、标准差
- 方差系数、标准差系数

一、绝对数与相对数

- （一）绝对数
- 反映社会现象整体规模和水平
- 时期数
- 时点数
- （二）相对数
- 结构、比较、计划等

（三）绝对数与相对数的应用

- 1、指标内涵和可比性：GNP、工业增加值
- 2、指标的结合运用
- 引例：在美国，1985年有19893人遭谋杀，与1970年16848人遭到谋杀相比，增加了20%。这些数字揭示了在1970-1985年期间美国变成一个更多暴力的社会
- 中国的国民生产总值增长了8%，美国的为1%

二、数据集中趋势的描述

- （一）、 众数 Mode
- 众数是总体数据中出现次数最多的变量值。
- 例 3-1：有 10 名大学生的年龄： 18， 18， 19， 19， 19， 19， 19， 20， 20， 21，在这里 19 岁的人数最多，所以 19 岁是众数。
- 例 3-2：有 10 名职工的年龄： 20， 21， 22， 23， 24， 25， 26， 27， 28， 29，由于各年龄的人数相同，没有明显集中趋势点的数值，所以这里没有众数。

注意：

- 1、是位置平均数，不受极端值的影响
- 2、假定各单位在组内是均匀分布的
- 3、信息量小，缺乏敏感性，不适合代数运算
- 4、用于非对称的次数数列、特别是品质标志数列
- 5、用于数列中有较多的数值向某一数值集中
- 6、有时会存在多个众数

（二）、中位数 Medium

- 中位数就是把计算对象的数据按大小顺序排列后，处于中间位置上的变量值。

=