

经济、管理类
基础课程

统计学

统计学

PowerPoint



作者：中国人民大学统计系

贾俊平

第一章 绪论

PowerPoint



第一章 绪论

第一节 统计与统计学

第二节 统计学的分科

第三节 统计学与其他学科的关系

第四节 统计学的产生与发展

学习目标

1. 理解统计与统计学的含义
2. 理解统计数据与统计学的关系
3. 区分描述统计与推断统计
4. 了解统计学与其他学科的关系
5. 了解统计学的产生与发展过程

第一节 统计与统计学

- 一. 统计与统计学的含义
- 二. 统计数据的规律与统计方法

什么是统计？



1. 统计工作

- 收集数据的活动

2. 统计数据

- 对现象计量的结果

3. 统计学

- 分析数据的方法与技术

什么是统计学？

- ➡ 统计学是一门收集、整理和分析数据的方法科学，其目的是探索数据的内在数量规律性，以达到对客观事物的科学认识



1. 数据搜集：例如，调查与试验
2. 数据整理：例如，分组
3. 数据展示：例如，图和表
4. 数据分析：例如，回归分析

Statistics的定义 (不列颠百科全书)

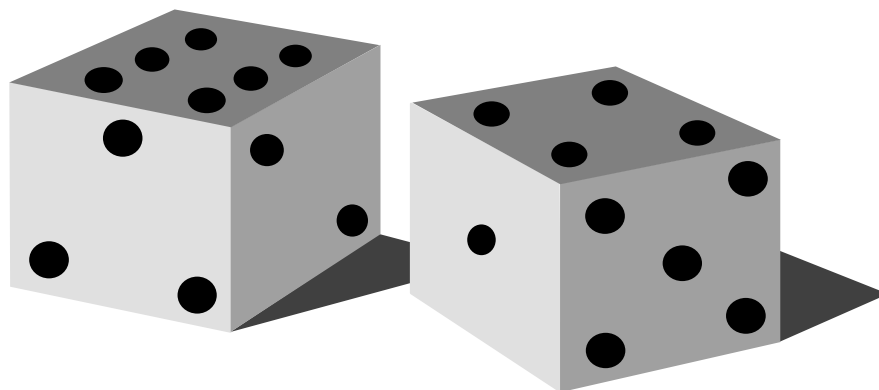
***Statistics: the science of
collecting, analyzing,
presenting, and
interpreting data.***

Copyright 1994-2000 Encyclopaedia Britannica, Inc.

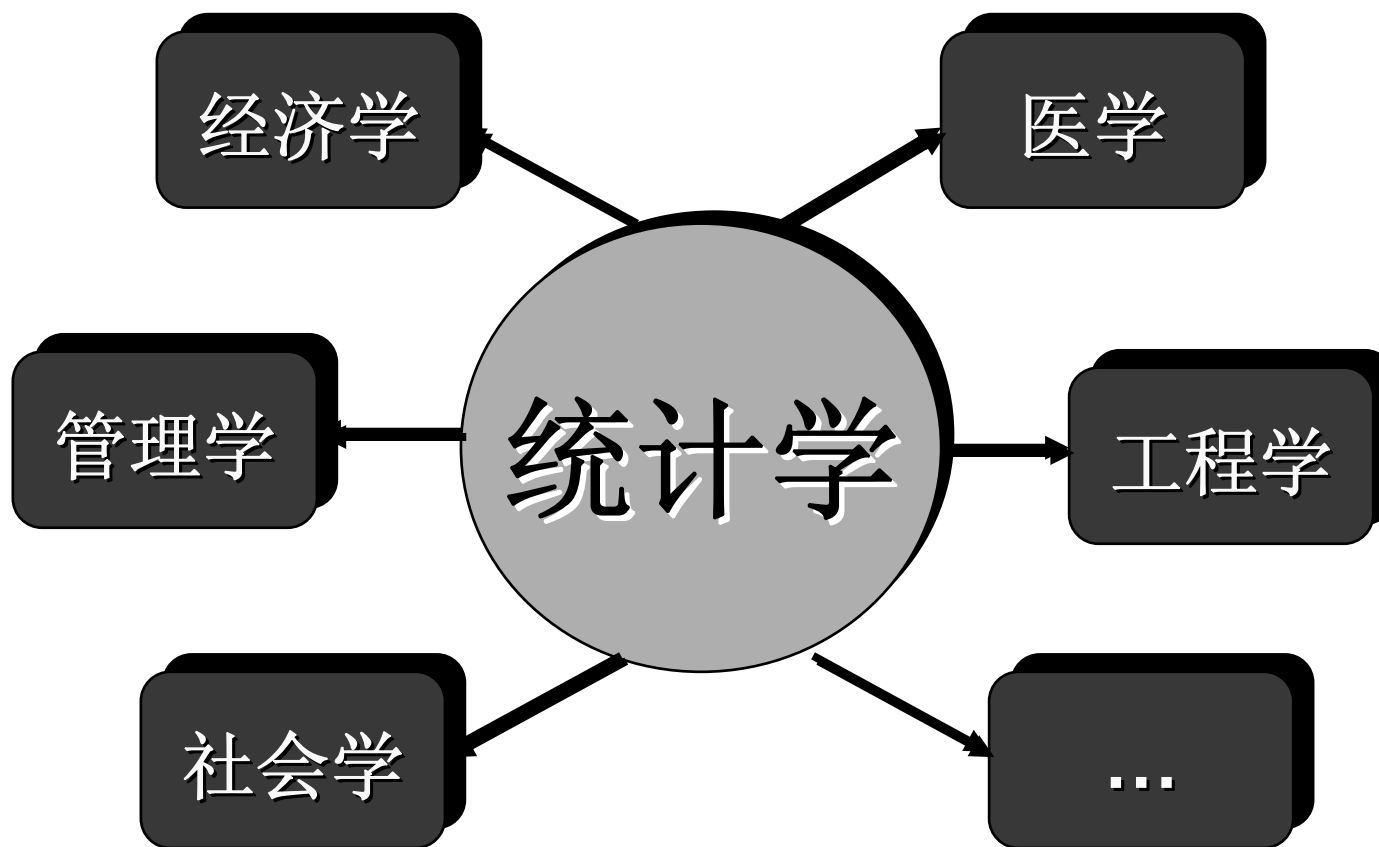
(不列颠百科全书)

统计数据的内在规律 (一些例子)

1. 正常条件下新生婴儿的性别比为107: 100
2. 投掷一枚均匀的硬币，出现正面和反面的频率各为 $1/2$ ；投掷一枚骰子出现1~6点的频率各为 $1/6$
3. 农作物的产量与施肥量之间存在相关关系



统计学的应用领域



应用统计的领域

actuarial work (精算)	agriculture (农业)
animal science (动物学)	anthropology (人类学)
archaeology (考古学)	auditing (审计学)
crystallography (晶体学)	demography (人口统计学)
dentistry (牙医学)	ecology (生态学)
econometrics (经济计量学)	education (教育学)
election forecasting and projection (选举预测和策划)	epidemiology (流行病学)
engineering (工程)	
finance (金融)	
fisheries research (水产渔业研究)	
gambling (赌博)	genetics (遗传学)
geography (地理学)	geology (地质学)
historical research (历史研究)	human genetics (人类遗传学)

应用统计的领域(续)

hydrology (水文学)

linguistics (语言学)

manpower planning (劳动力计划)

management science (管理科学)

marketing (市场营销学)

meteorology (气象学)

nuclear material safeguards (核材料安全管理)

ophthalmology (眼科学)

physics (物理学)

psychology (心理学)

quality control (质量控制)

sociology (社会学)

taxonomy (分类学)

Industry (工业)

literature (文学)

medical diagnosis (医学诊断)

military science (军事科学)

pharmaceutics (制药学)

political science (政治学)

psychophysics (心理物理学)

religious studies (宗教研究)

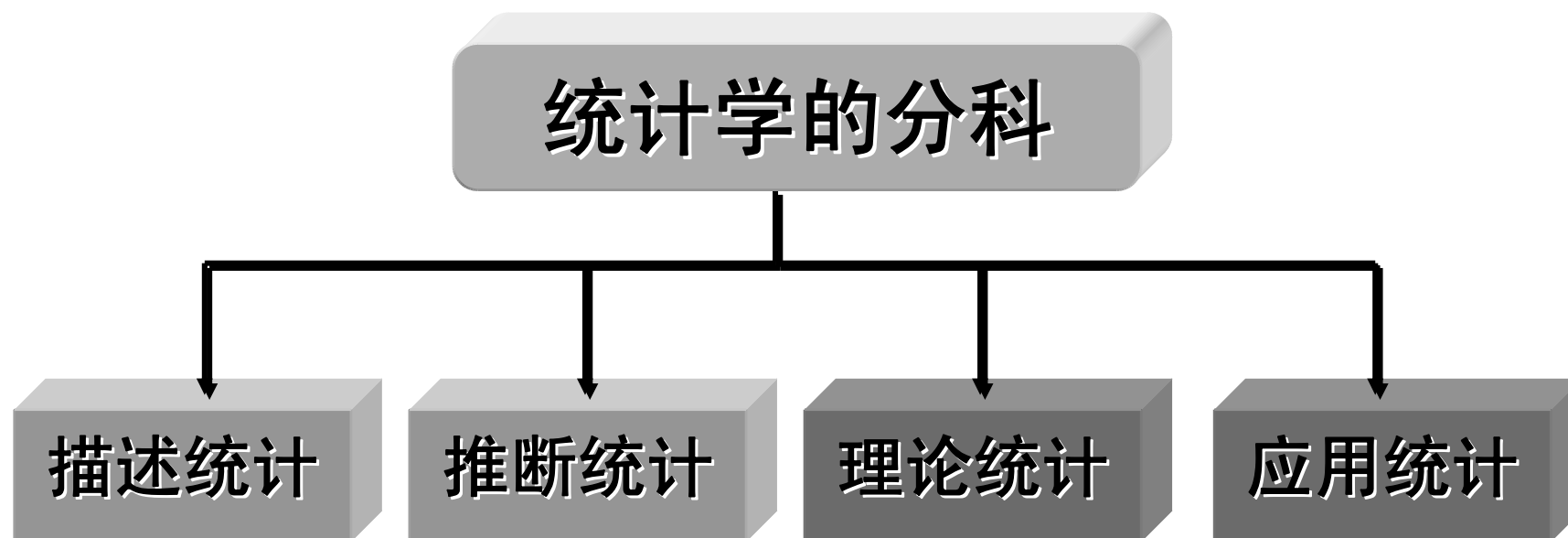
survey sampling (调查抽样)

weather modification (气象改善)

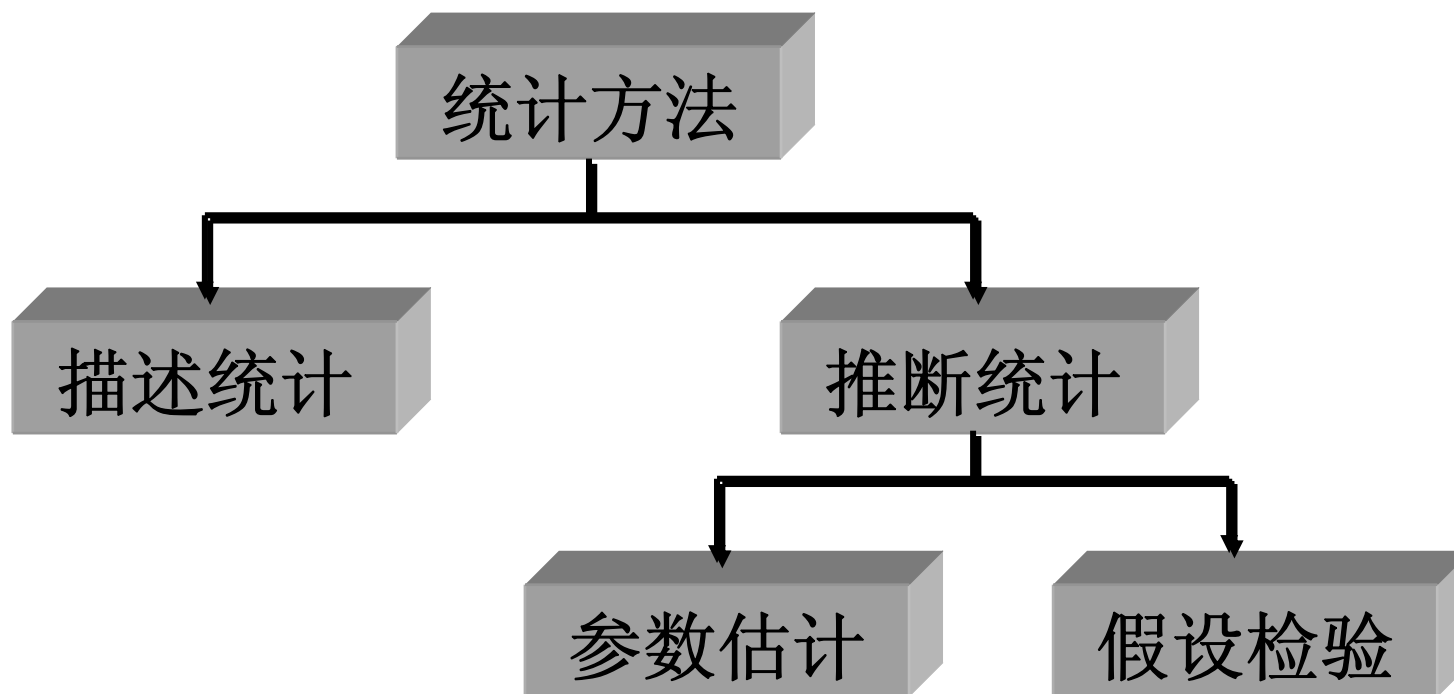
第二节 统计学的分科

- 一. 描述统计学和推断统计学
- 二. 理论统计学和应用统计学

统计学的分科



统计方法



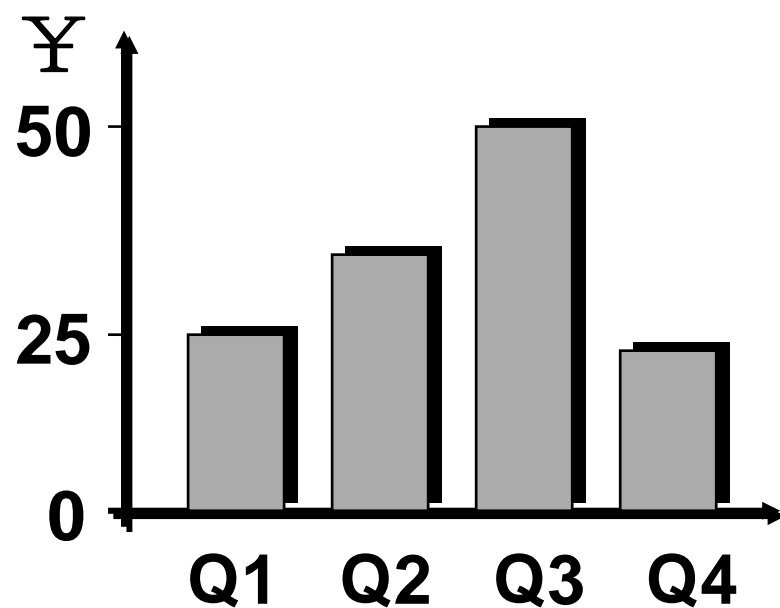
描述统计

1. 内容

- 搜集数据
- 整理数据
- 展示数据

2. 目的

- 描述数据特征
- 找出数据的基本规律



$$\bar{x} = 30 \quad s^2 = 105$$

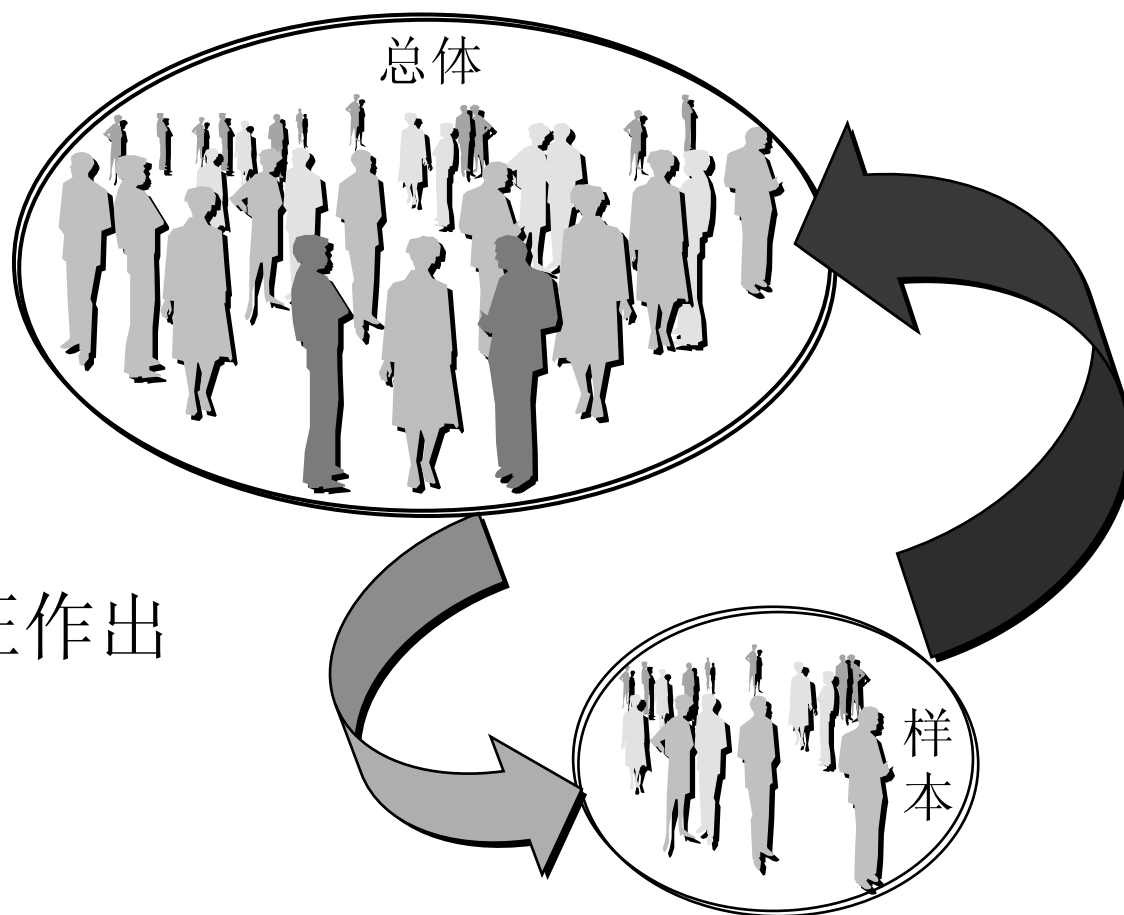
推断统计

1. 内容

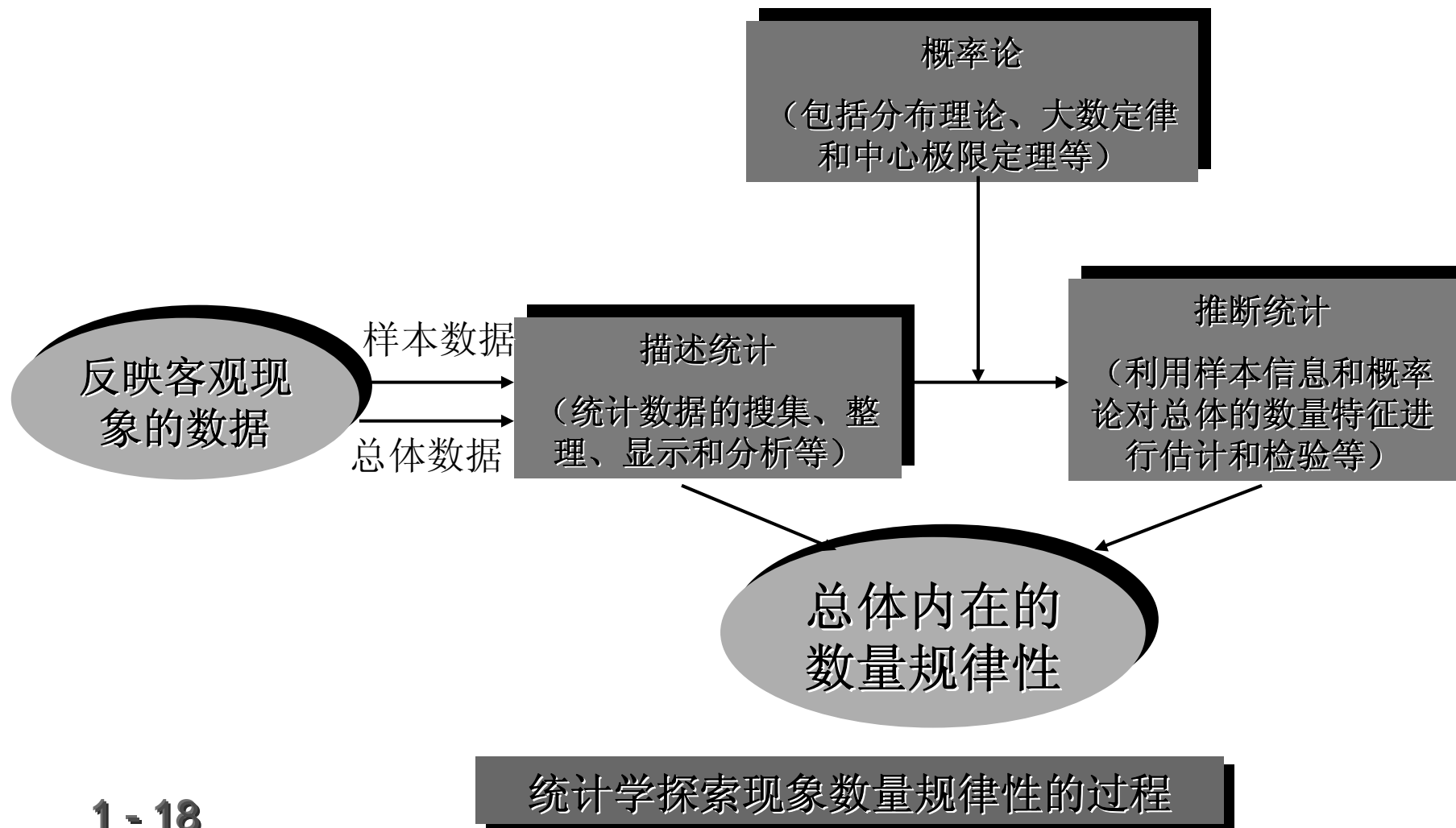
- 参数估计
- 假设检验

2. 目的

- 对总体特征作出推断



描述统计与推断统计的关系



理论统计与应用统计

1. 理论统计

- 研究统计学的一般理论
- 研究统计方法的数学原理

2. 应用统计

- 研究统计学在各领域的具体应用

第三节

统计学与其他学科的关系

- 一. 统计学与数学的关系
- 二. 统计学与其他学科的关系

统计学与数学的关系 (联系)

1. 统计学运用到大量的数学知识
2. 数学为统计理论和统计方法的发展提供基础
3. 不能将统计学等同于数学

统计学与数学的关系 (区别)

1. 数学研究的是抽象的数量规律，统计学则是研究具体的、实际现象的数量规律
2. 数学研究的是没有量纲或单位的抽象的数，统计学研究的是有具体实物或计量单位的数据
3. 统计学与数学研究中所使用的逻辑方法不同
 - 数学研究所使用的主要是演绎
 - 统计学则是演绎与归纳相结合，占主导地位的是归纳

统计学与其他学科的关系

1. 统计学可以用到几乎所有的学科领域
2. 统计学可以帮助其他学科探索学科内在的数量规律性
3. 统计学不能解决各学科领域的所有问题
4. 对统计分析结果的解释需要各学科领域的专业人员

第四节 统计学的产生与发展

- 一. 政治算术—社会经济统计
- 二. 概率论—数理统计

统计学家是科学家

统计学家是科学家



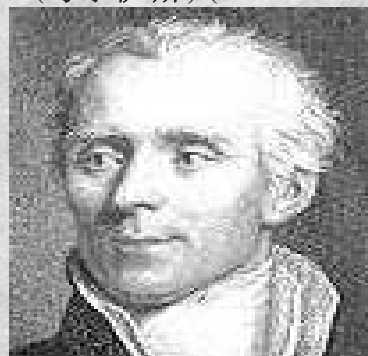
Thomas Robert Malthus (马尔萨斯) (1766-1834)



Leonhard Euler (欧拉) (1707-1783)



Friedrich Gauss (高斯) (1777-1855)



Pierre Simon Laplace (拉普拉斯) (1749-1827)



Johann Gregor Mendel (孟德尔) (1822-1884)

历史上著名的统计学家

Jacob Bernoulli (伯努利) (1654-1705)
Edmond Halley (哈雷) (1656-1742)
De Moivre (棣美佛) (1667-1754)
Thomas Bayes (贝叶斯) (1702-1761)
Leonhard Euler (欧拉) (1707-1783)
Pierre Simon Laplace (拉普拉斯) (1749-1827)
Adrien Marie Legendre (勒让德) (1752-1833)
Thomas Robert Malthus (马尔萨斯) (1766-1834)
Friedrich Gauss (高斯) (1777-1855)
Johann Gregor Mendel (孟德尔) (1822-1884)
Karl Pearson (皮尔森) (1857-1936)
Ronald Aylmer Fisher (费歇) (1890-1962)
Jerzy Neyman (内曼)(1894-1981)
Egon Sharpe Pearson (皮尔森) (1895-1980)
William Feller (费勒)(1906-1970)

统计学发展的历史线索

1. 一般认为，统计学产生于17世纪中叶
2. 统计学的发展过程基本上沿着两条主线展开
 - 以“政治算术学派”为开端形成和发展起来的、以社会经济问题为主要研究对象的社会经济统计
 - 以概率论的研究为开端、并以概率论为基础形成和发展起来的、以方法和应用研究为主的数理统计
3. 今天，社会经济统计和数理统计仍然在以各自不同的方式发展着

政治算术—社会经济统计

1. 政治算术学派产生于17世纪中叶的英国，代表人物主要是威廉·配第(William Petty, 1623—1687)和约翰·格朗特(John Graunt, 1620—1674)
2. 17世纪中叶的政治算术学派可看作是统计学的开端
3. 19世纪，沿着约翰·格朗特所开创的人口统计以及沿着威廉·配第所开创的经济统计有了进一步的发展
4. 威廉·配第为以后经济统计的发展开拓了道路；约翰·格朗特为人口统计的发展开拓了道路
5. 政治算术学派则为后来的社会经济统计的发展奠定了基础

概率论—数理统计

1. 概率论研究起源于意大利文艺复兴时代
2. 概率论的真正历史是从17世纪中叶开始的
3. 古典统计时期的概率论基本上是独立发展的，它与统计学(主要是指政治算术)没有太多的联系
4. 从19世纪中叶到20世纪中叶，概率论的进一步发展为数理统计学的形成和发展奠定了基础
5. 本世纪50年代以后，统计理论、方法和应用进入了一个全面发展的阶段

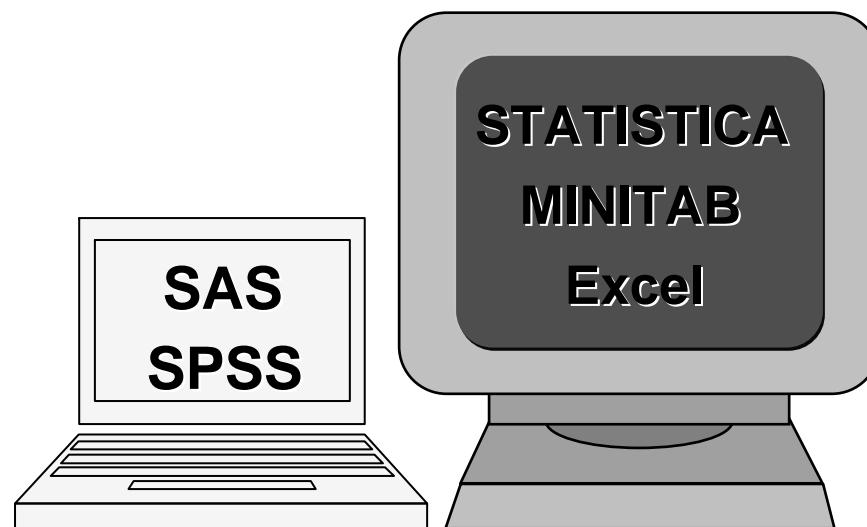
统计学中的几个主要术语

1. 总体(Population)
 - 所关心的所有元素的集合
2. 样本(Sample)
 - 总体的一部分
3. 参数(Parameter)
 - 总体的数字特征
4. 统计量(Statistic)
 - 样本的概括性测度值

几种常用的统计软件 (Software)

👉 典型的统计软件

- SAS
- SPSS
- MINITAB
- STATISTICA
- Excel



本章小节

1. 统计与统计学
2. 统计学的用途
3. 描述统计学与推断统计学
4. 统计学与其他学科的关系
5. 统计学的产生与发展

结 束



第二章 统计数据的搜集

PowerPoint



第二章 统计数据的搜集

第一节 数据的计量与类型

第二节 统计数据的来源

第三节 调查方案设计

第四节 统计数据的质量

学习目标

1. 了解数据的计量尺度与数据的类型
2. 了解统计调查方式
3. 了解数据的搜集方法
4. 掌握调查方案的设计
5. 了解数据误差及对数据的质量要求

第一节 数据的计量与类型

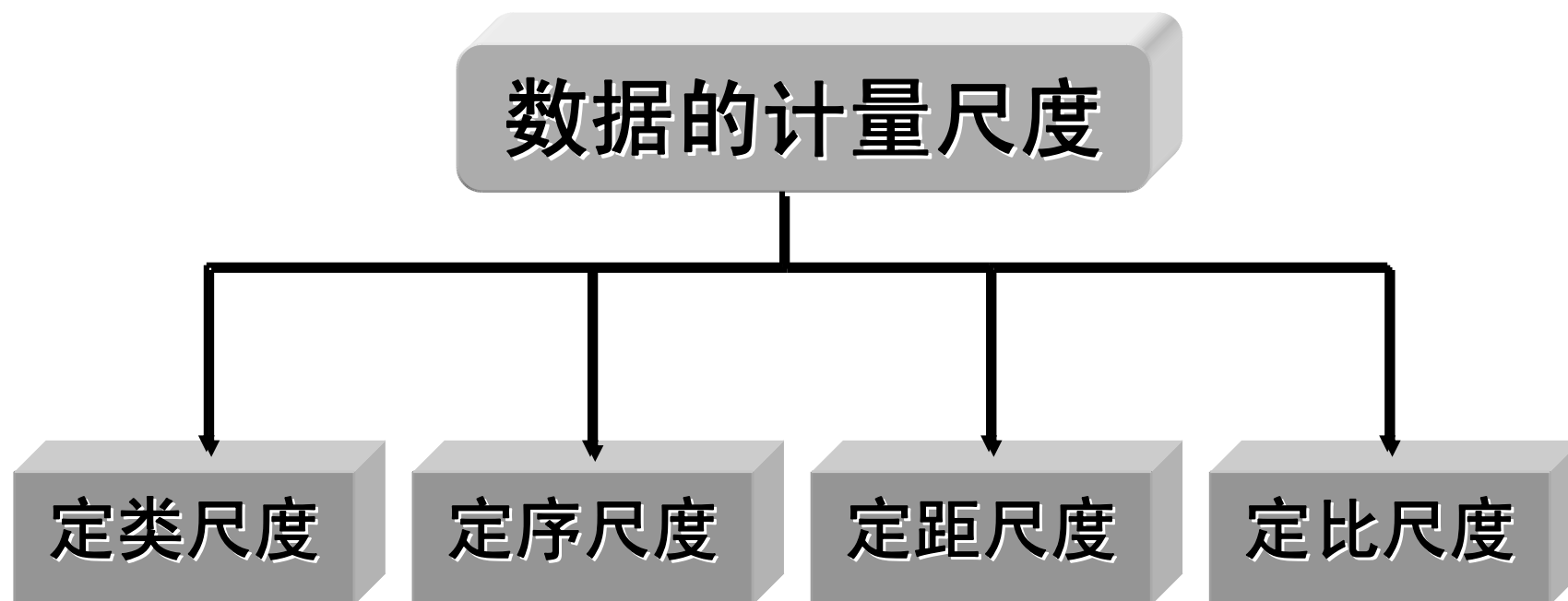
- 一. 数据的计量尺度
- 二. 数据的类型和分析方法
- 三. 统计指标及其类型

经济、管理类
基础课程

统计学

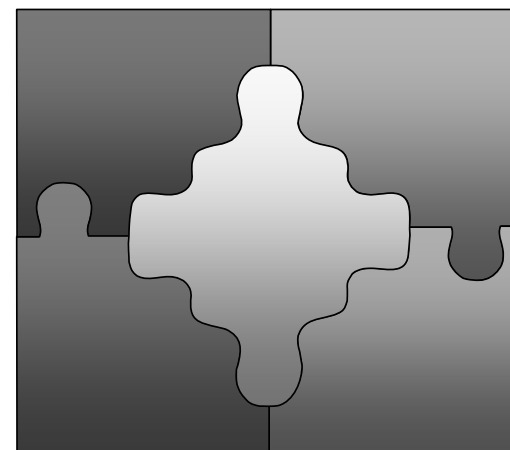
数据的计量尺度

四种计量尺度



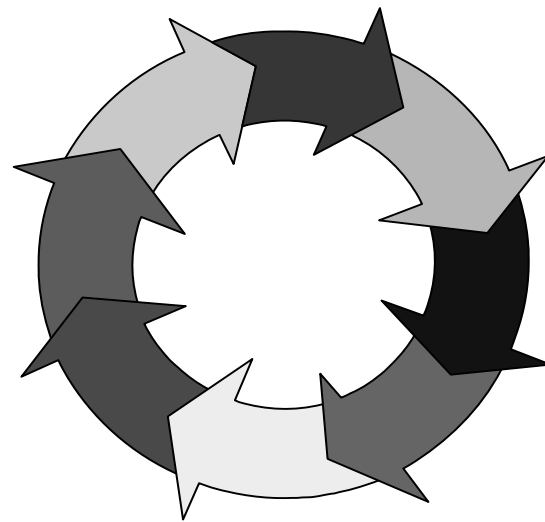
定类尺度 (概念要点)

1. 计量层次最低
2. 对事物进行平行的分类
3. 各类别可以指定数字代码表示
4. 使用时必须符合类别穷尽和互斥的要求
5. 数据表现为“类别”
6. 具有 $=$ 或 \neq 的数学特性



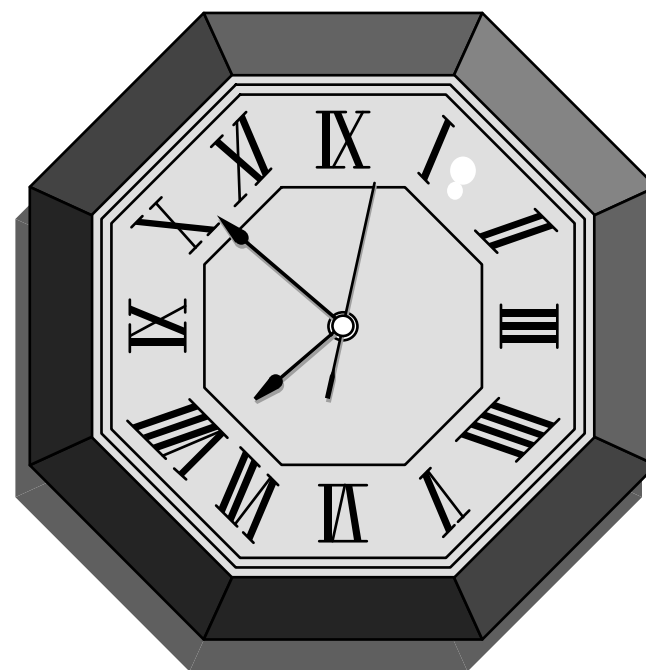
定序尺度 (概念要点)

1. 对事物分类的同时给出各类别的顺序
2. 比定类尺度精确
3. 未测量出类别之间的准确差值
4. 数据表现为“类别”，但有序
5. 具有 $>$ 或 $<$ 的数学特性



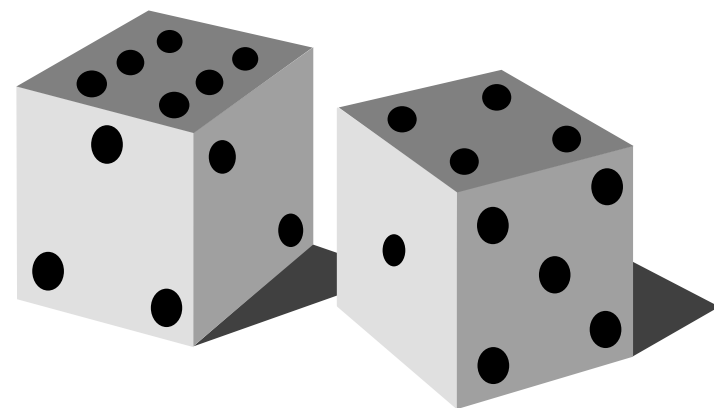
定距尺度 (概念要点)

1. 对事物的准确测度
2. 比定序尺度精确
3. 数据表现为“数值”
4. 没有绝对零点
5. 具有 + 或 - 的数学特性



定比尺度 (概念要点)

1. 对事物的准确测度
2. 与定距尺度处于同一层次
3. 数据表现为“数值”
4. 有绝对零点
5. 具有 \times 或 \div 的数学特性



四种计量尺度的比较

四种计量尺度的比较				
计量尺度 数学特性	定类尺度	定序尺度	定距尺度	定比尺度
分类 (= , ≠)	√	√	√	√
排序 (< , >)		√	√	√
间距 (+ , -)			√	√
比值 (× , ÷)				√

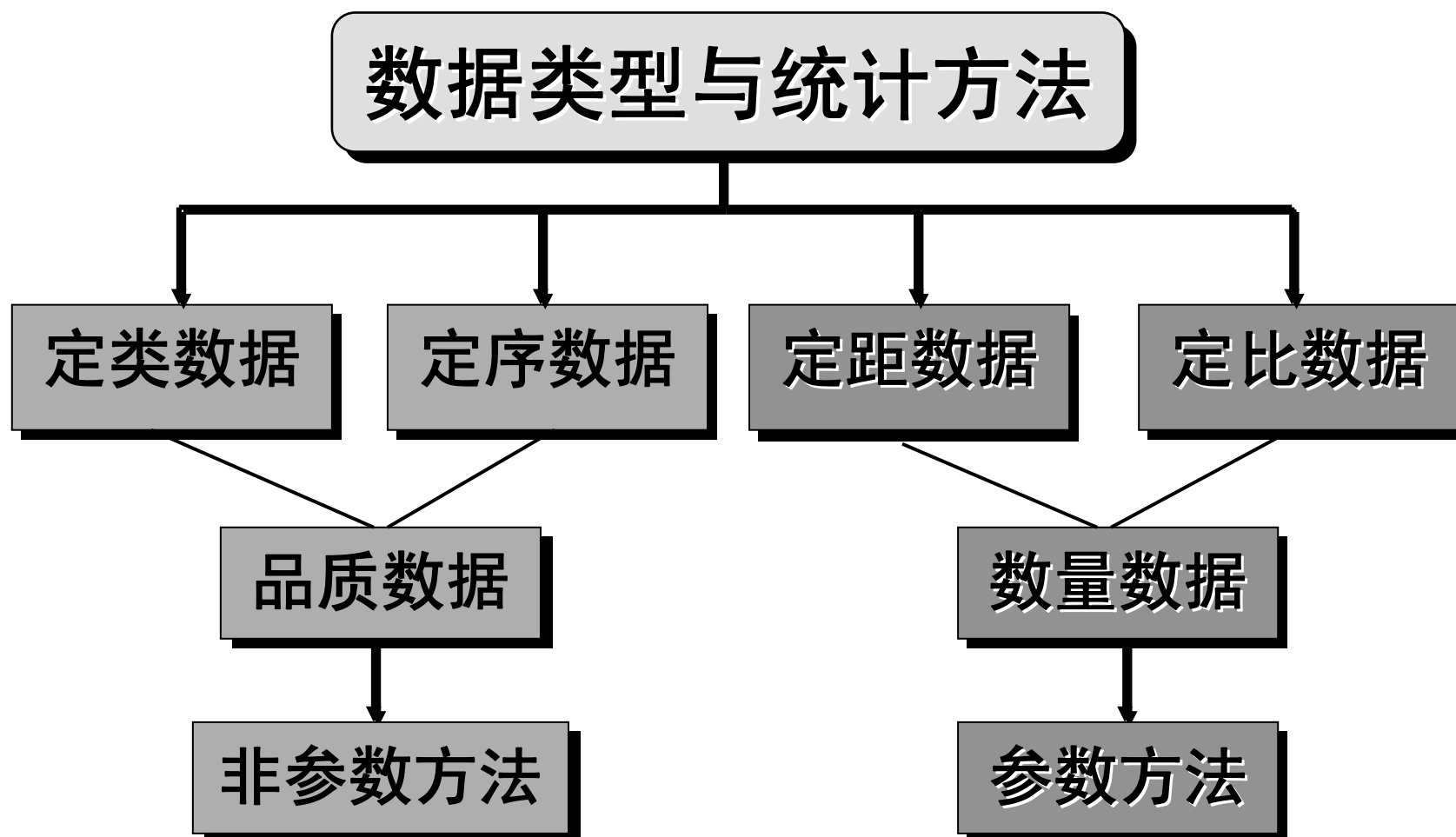
“√”表示该尺度所具有的特性

经济、管理类
基础课程

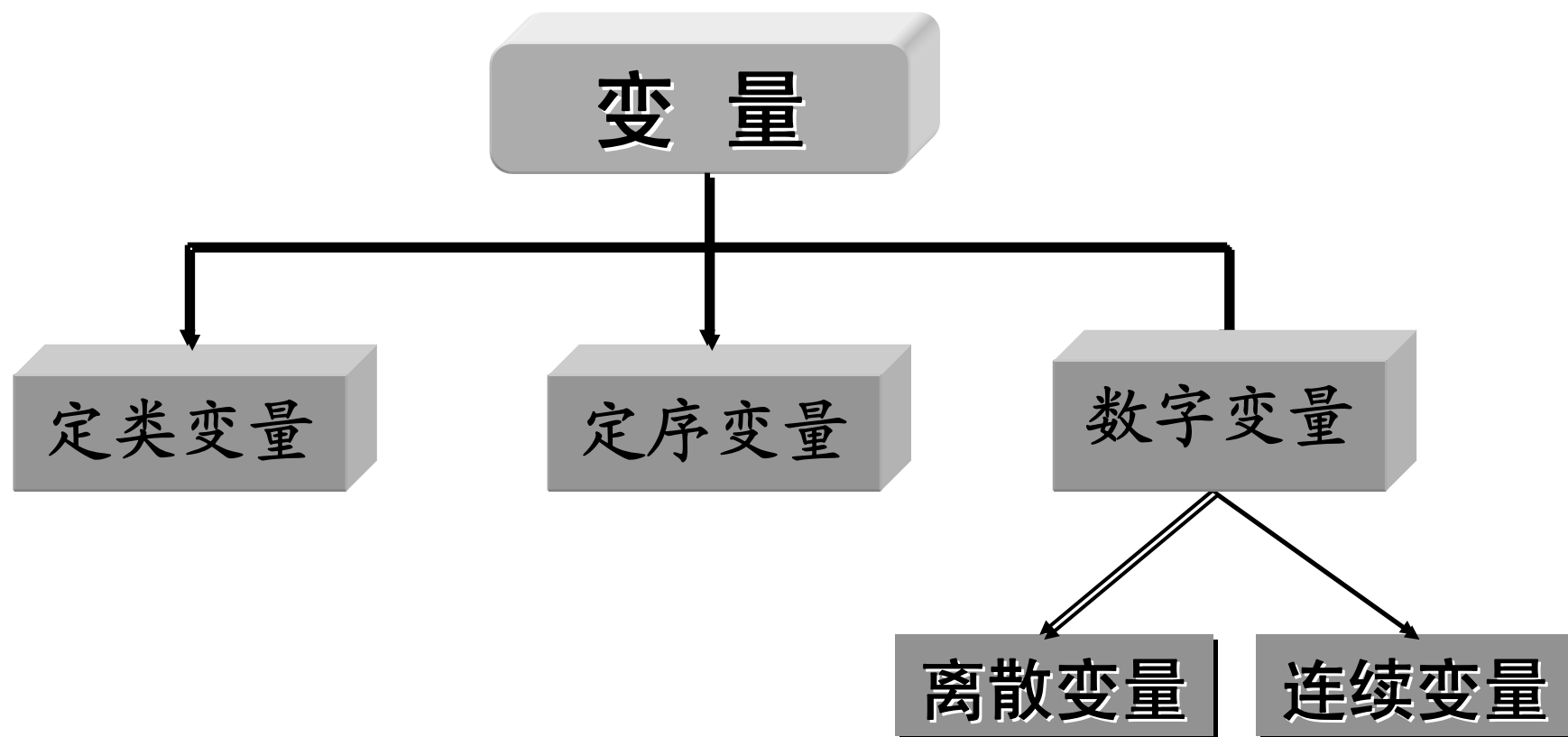
统计学

数据类型和分析方法

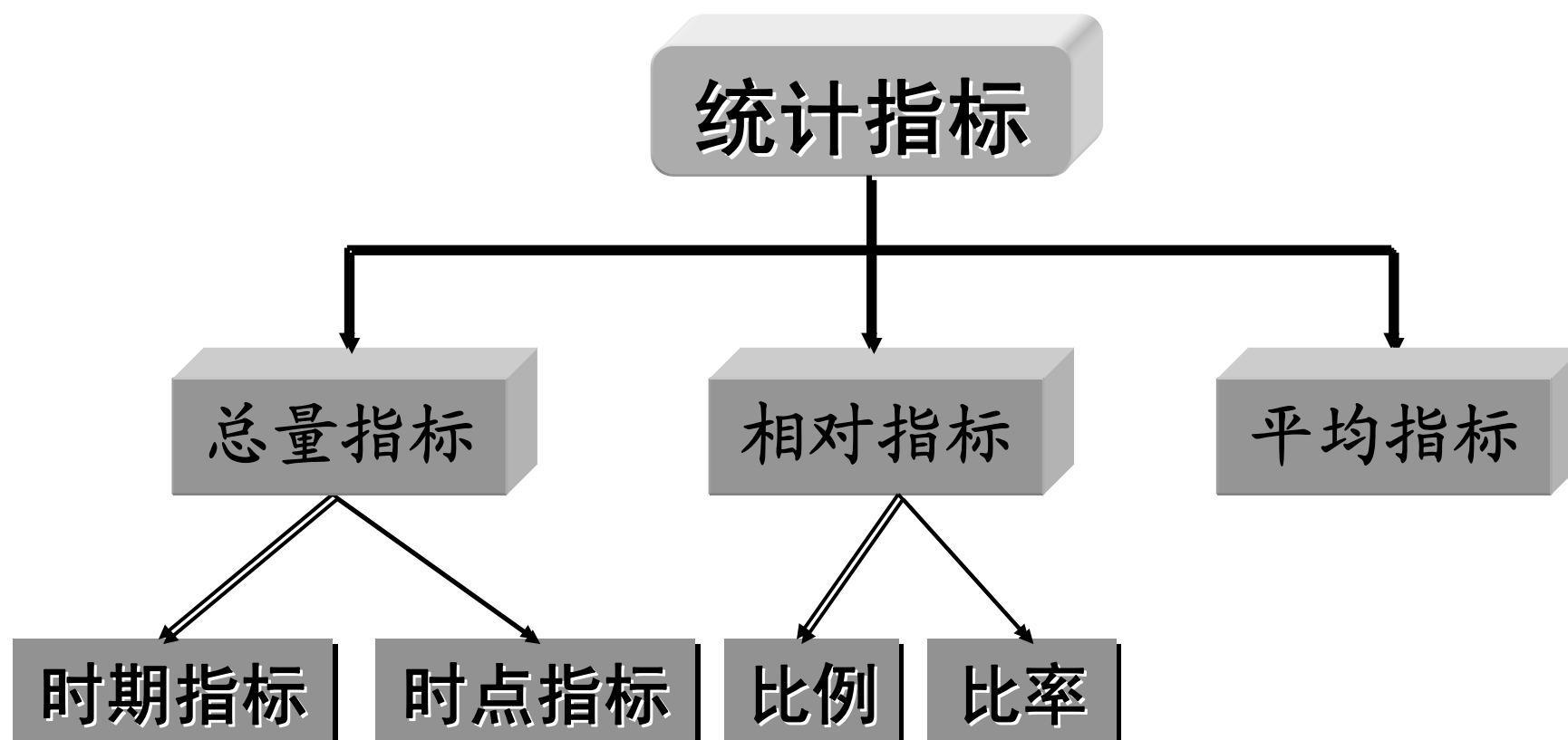
数据类型与统计方法



变量及其类型



统计指标及其类型



第二节 统计数据的来源

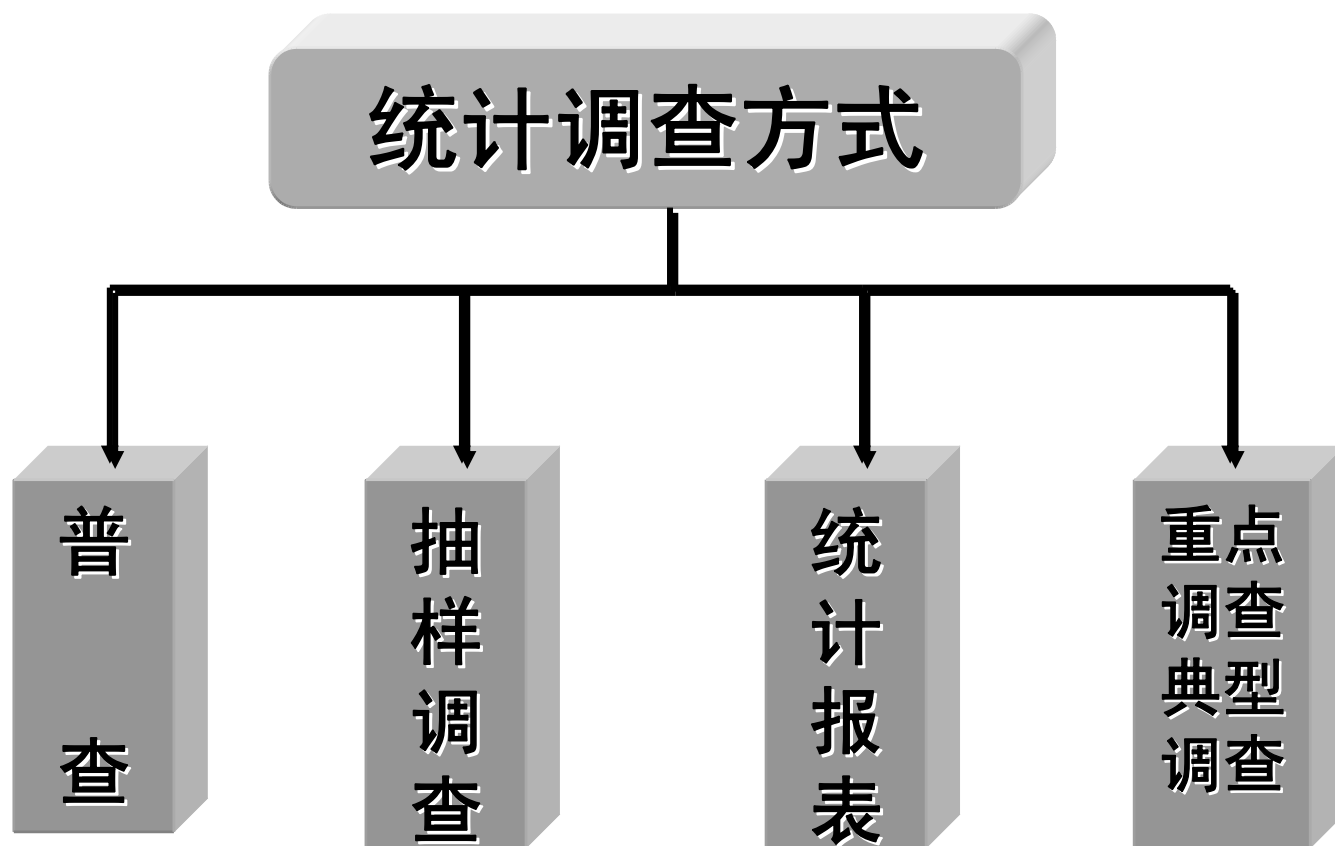
- 一. 统计数据的直接来源
- 二. 统计数据的间接来源

经济、管理类
基础课程

统计学

统计调查方式

统计调查方式



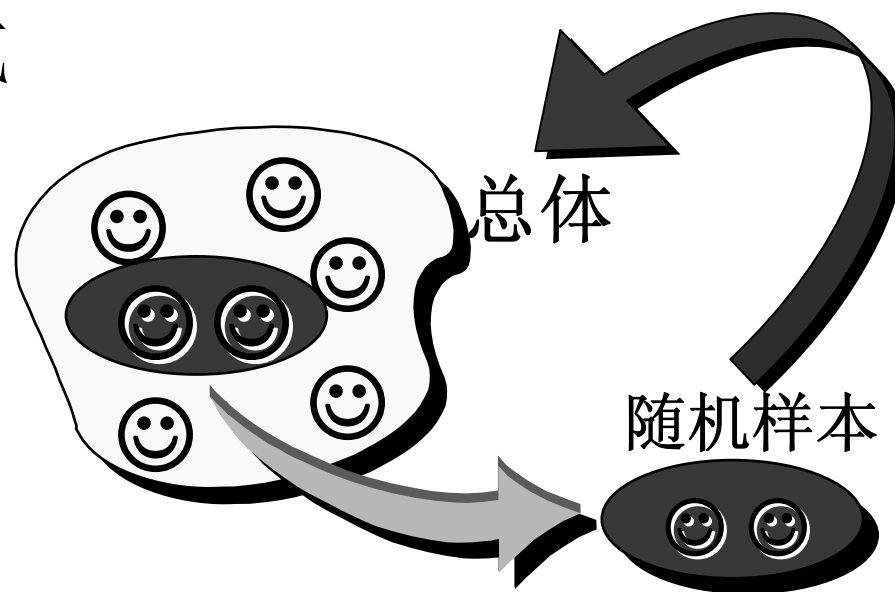
普查 (概念要点)

1. 为特定目的专门组织的非经常性全面调查
2. 通常是一次性或周期性的
3. 一般需要规定统一的标准调查时间
4. 数据的规范化程度较高
5. 应用范围比较狭窄



抽样调查 (概念要点)

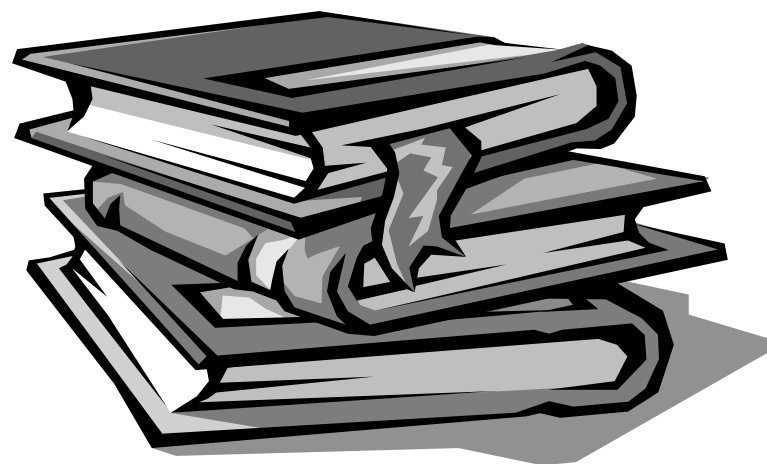
1. 从总体中随机抽取一部分单位(样本)进行调查
2. 目的是推断总体的未知数字特征
3. 最常用的调查方式
4. 具有经济性、时效性强、适应面广、准确性高等特点



一次失败的抽样调查

统计报表 (概念要点)

1. 统计调查方式之一
2. 过去曾经是我国主要的数据收集方式
3. 按照国家有关法规的规定、自上而下地统一布置、自下而上地逐级提供基本统计数据
4. 有各种各样的类型



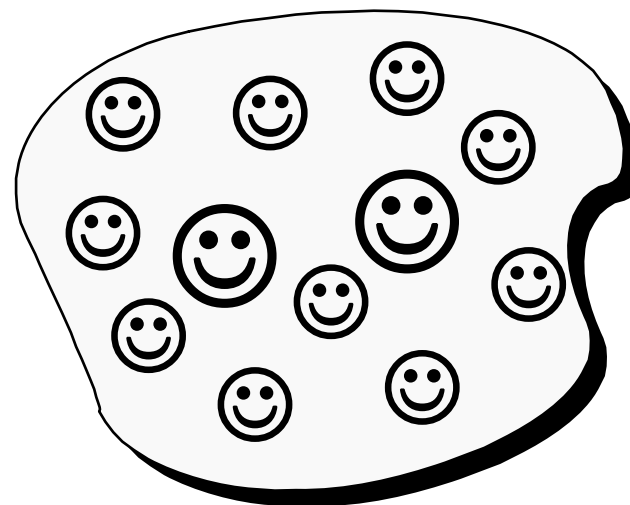
重点调查和典型调查 (概念要点)

1. 重点调查

- 从调查对象的全部单位中选择少数重点单位进行调查
- 调查结果不能用于推断总体

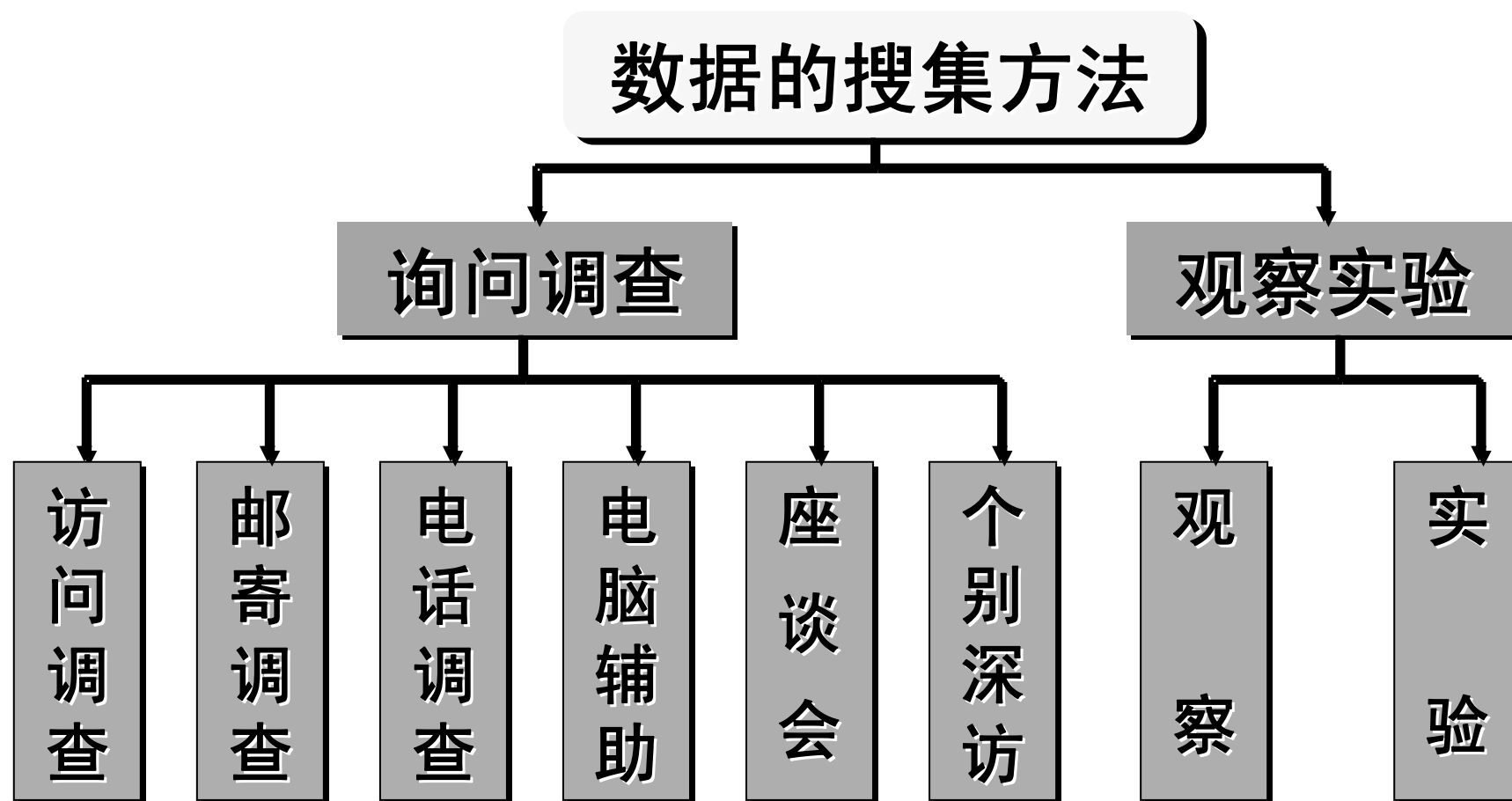
2. 典型调查

- 从调查对象的全部单位中选择少数典型单位进行调查
- 目的是描述和揭示事物的本质特征和规律
- 调查结果不能用于推断总体



数据的搜集方法

数据的搜集方法



访问调查 (概念要点)

1. 调查者与被调查者通过面对面地交谈而获得资料
2. 有标准式访问和非标准式访问
 - 标准式访问通常按事先设计好的问卷进行
 - 非标准式访问事先一般不制作问卷

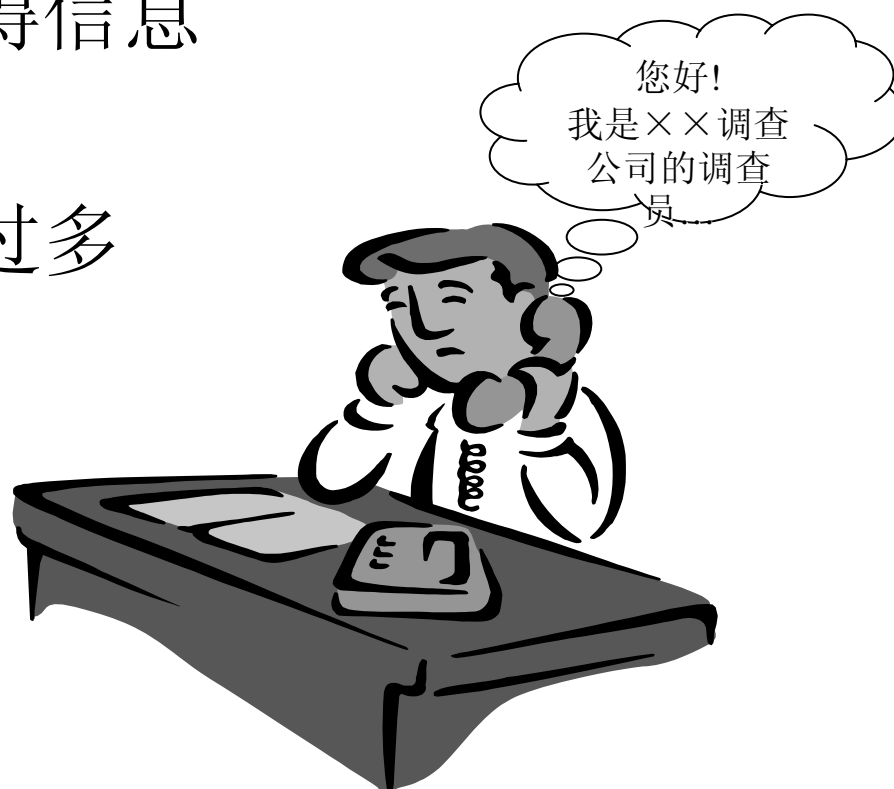


邮寄调查 (概念要点)

1. 也称邮寄问卷调查
2. 是一种标准化调查
3. 调查者与被调查者没有直接的语言交流，信息的传递依赖于问卷
4. 通过某种方式将调查表或问卷送至某调查者手中，由被调查者填写，然后将问卷寄回指定收集点
5. 问卷或表格的发放方式有邮寄、宣传媒介传送、专门场所分发三种

电话调查 (概念要点)

1. 调查者利用电话与被调查者进行语言交流以获得信息
2. 时效快、成本低
3. 问题的数量不宜过多



电脑辅助调查 (概念要点)

1. 又称电脑辅助电话调查
2. 电脑与电话相结合完成调查的全过程
3. 一般需借助专门的软件进行
4. 硬件设备要求较高



座谈会 (概念要点)

1. 也称集体访谈
2. 将一组被调查者集中在调查现场，让他们对调查的主题发表意见以获得资料
3. 参加座谈会的人数不宜过多，一般为6~10人
4. 侧重于定性研究



个别深度访问 (概念要点)

1. 一次只有一名受访者参加、
针对特殊问题的调查
2. 适合于较隐秘的问题，如个
人隐私问题；或较敏感的问题，如政治方面的问题
3. 侧重于定性研究



观察法 (概念要点)

1. 就调查对象的行动和意识，调查人员边观察边记录以收集所需信息
2. 调查人员不是强行介入
3. 能够在被调查者不察觉的情况下获得资料



实验法 (概念要点)

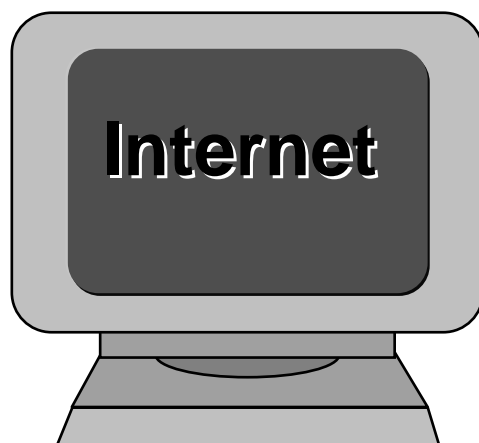
1. 在设定的特殊实验场所、特殊状态下，对调查对象进行实验以获得所需资料
2. 有室内实验法和市场实验法



统计数据的间接来源

1. 公开出版物：《中国统计年鉴》、《中国统计摘要》、《中国社会统计年鉴》、《中国工业经济统计年鉴》、《中国农村统计年鉴》、《中国人口统计年鉴》、《中国市场统计年鉴》、《世界经济年鉴》、《国外经济统计资料》、《世界发展报告》……

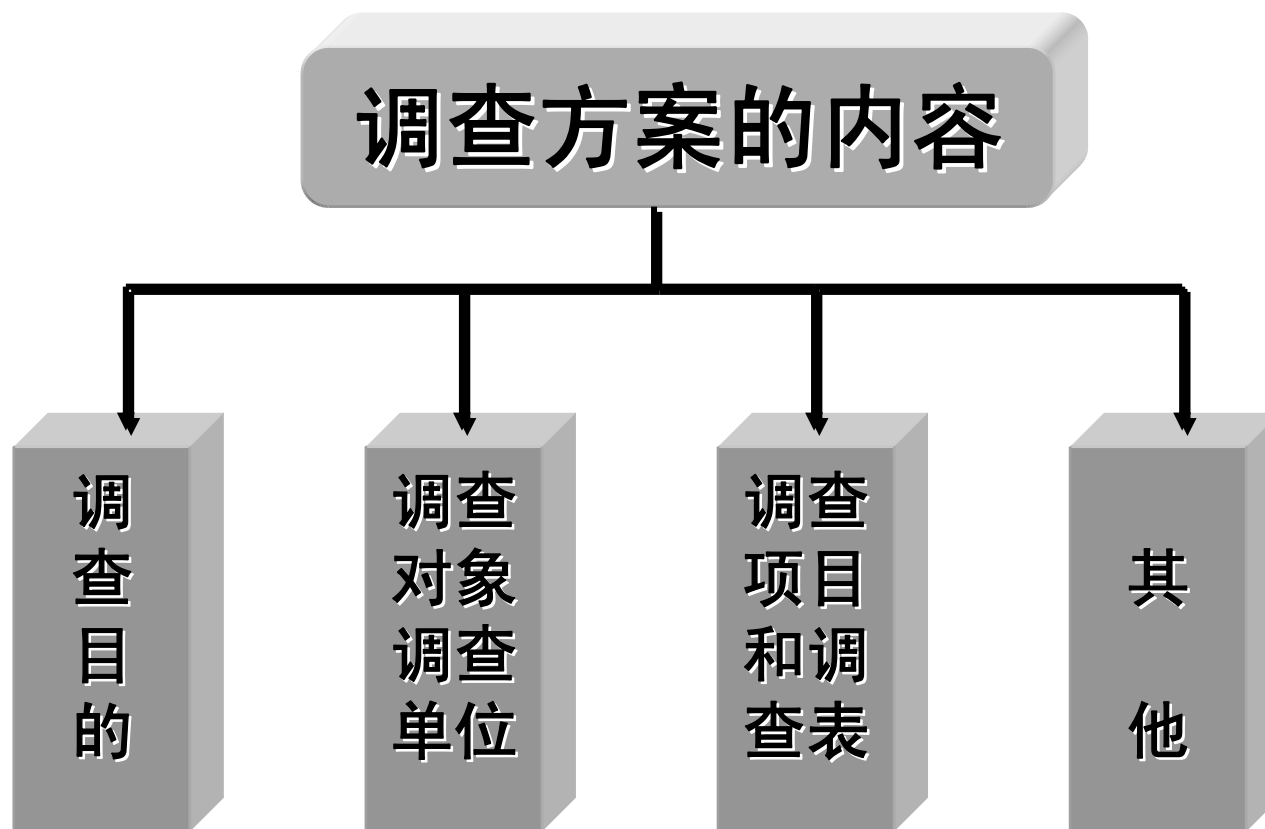
2. 网络



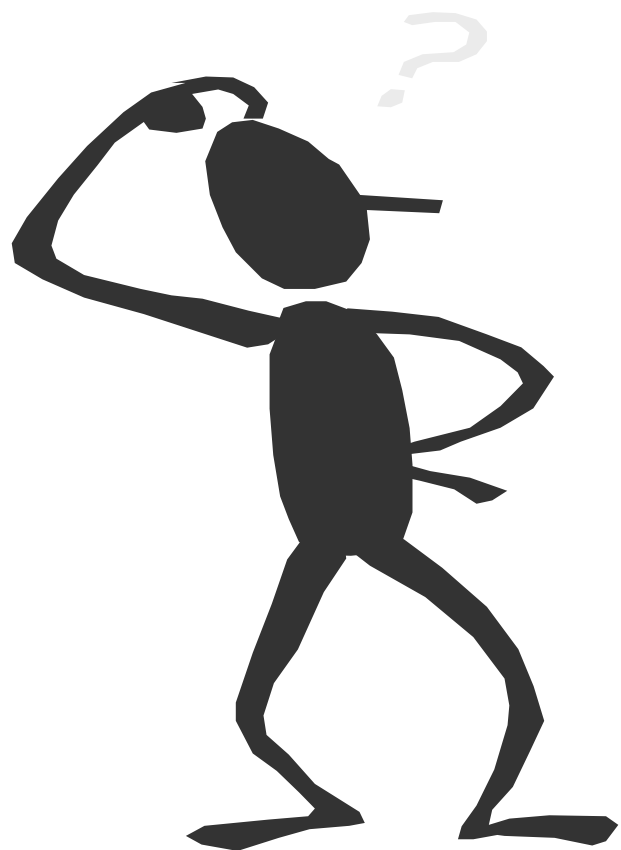
第三节 调查方案设计

- 一. 确定调查目的
- 二. 确定调查对象和调查单位
- 三. 设计调查项目和调查表
- 四. 方案设计中的其他内容

调查方案设计



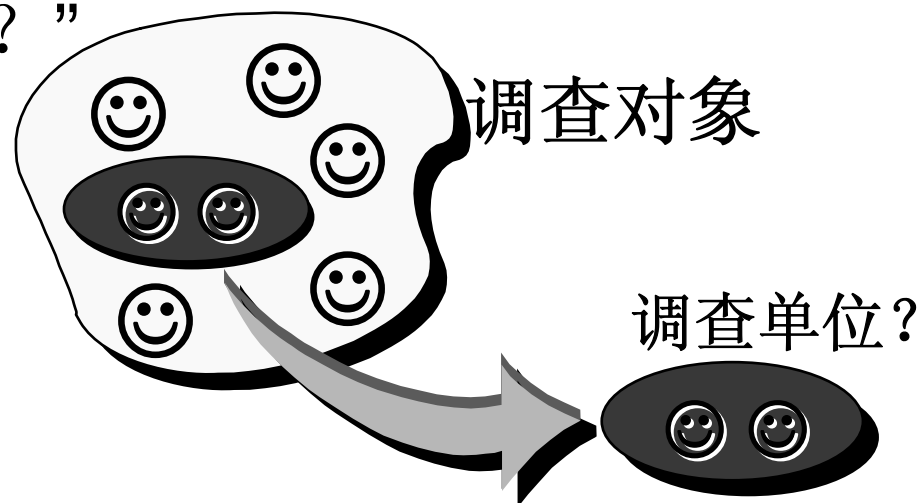
调查目的



1. 调查要达到的具体目标
2. 回答“为什么调查？”
3. 调查之前必须明确

调查对象和调查单位

1. 调查对象：调查研究的总体或调查范围
2. 调查单位：需要对之进行调查的单位。可以是调查对象的全部单位（全面调查），也可以是调查对象中的一部分单位（非全面调查）
3. 回答“向谁调查？”



调查项目和调查表

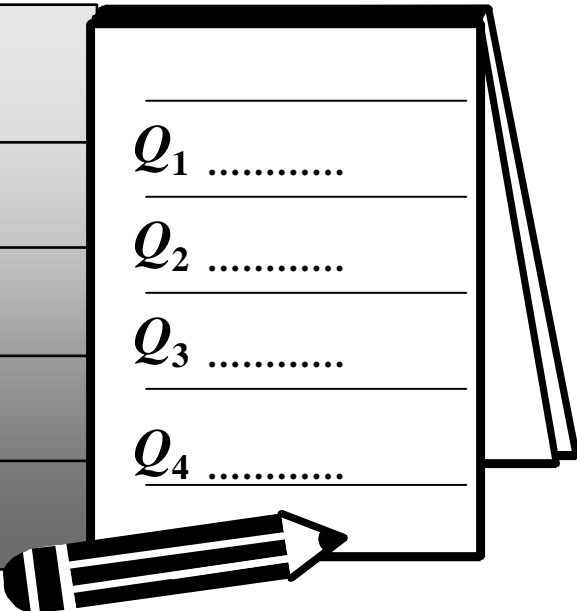
1. 调查项目：调查的具体内容
2. 调查表：表现调查项目的表格或问卷
3. 回答“调查什么？”

Q_1

Q_2

Q_3

Q_4



方案设计中的其他问题

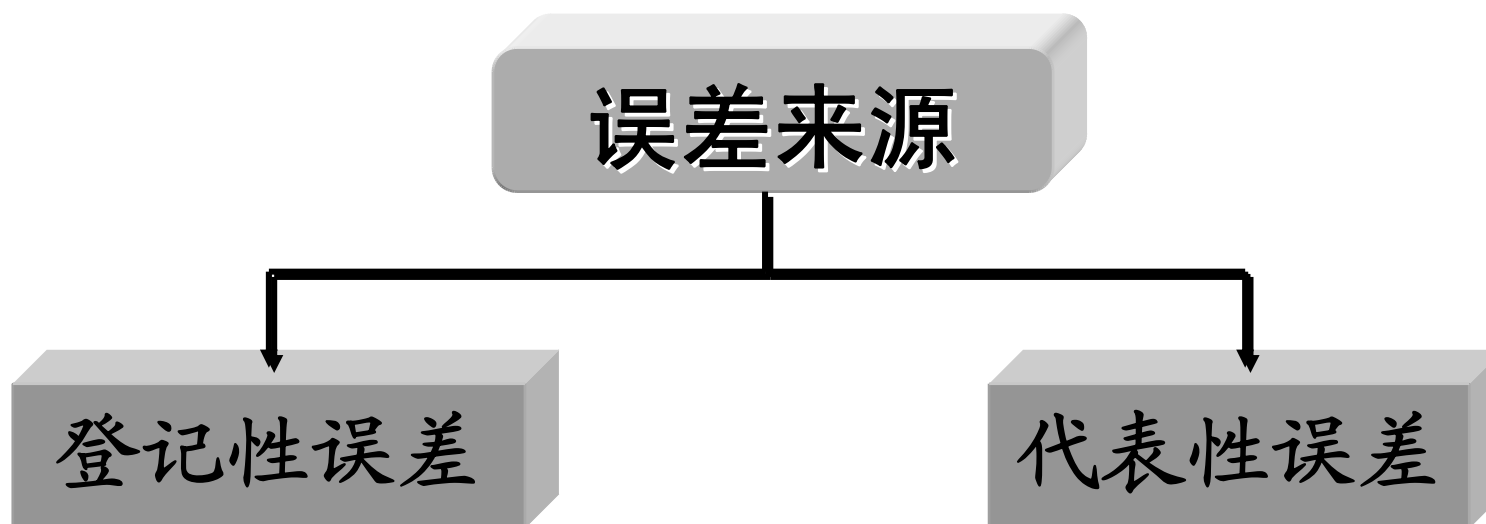
1. 明确调查所采用的方法
2. 确定调查资料的所属时间和调查工作的期限
3. 调查的组织与实施细则



第四节 统计数据的质量

- 一. 统计数据的误差
- 二. 统计数据的质量要求

数据误差的来源



统计数据的误差

1. 统计数据与客观现实之间的差距
2. 有登记性误差和代表性误差两类
 - 登记性误差：由于调查者或被调查者的人为因素所造成的误差。理论上讲可以消除
 - 代表性误差：用样本数据进行推断时所产生的误差。通常无法消除，但事先可以进行控制和计算

统计数据的质量要求

1. 精 度：最低的抽样误差或随机误差
2. 准 确 性：最小的非抽样误差或偏差
3. 关 联 性：满足用户决策、管理和研究的需要
4. 及 时 性：在最短的时间里取得并公布数据
5. 一 致 性：保持时间序列的可比性
6. 最低成本：以最经济的方式取得数据

本章小结

1. 数据的计量尺度与数据的类型
2. 统计调查方式
3. 数据的搜集方法
4. 调查方案的设计
5. 数据误差及对数据的质量要求

结 束



第三章 统计数据的整理与显示

PowerPoint



第三章 统计数据的整理与显示

第一节 数据的预处理

第二节 品质数据的整理与显示

第三节 数值型数据的整理与显示

第四节 统计表

学习目标

1. 了解数据预处理的内容和目的
2. 掌握品质数据整理与显示的方法
3. 掌握数值型数据整理与显示的方法
4. 用Excel作频数分布表和形图
5. 合理使用统计表

第一节 数据的预处理

- 一. 数据的审核与筛选
- 二. 数据的排序

数据的审核、筛选与排序



1. 数据的审核
 - 发现数据中的错误
2. 数据的筛选
 - 找出符合条件的数据
3. 数据排序
 - 发现数据的基本特征
 - 升序和降序

数据的审核 (原始数据)

➡ 审核的内容

1. 完整性审核

- 检查应调查的单位或个体是否有遗漏
- 所有的调查项目或指标是否填写齐全

2. 准确性审核

- 检查数据是否真实反映客观实际情况，内容是否符合实际
- 检查数据是否有错误，计算是否正确等

数据的审核 (原始数据)

➡审核数据准确性的方法

1. 逻辑检查

- 从定性角度，审核数据是否符合逻辑，内容是否合理，各项目或数字之间有无相互矛盾的现象
- 主要用于对定类数据和定序数据的审核

2. 计算检查

- 检查调查表中的各项数据在计算结果和计算方法上有无错误
- 主要用于对定距和定比数据的审核

数据的审核 (第二手数据)

1. 适用性审核

- 弄清楚数据的来源、数据的口径以及有关的背景材料
- 确定这些数据是否符合自己分析研究的需要

2. 时效性审核

- 应尽可能使用最新的统计数据

3. 确认是否必要做进一步的加工整理

数据的筛选

1. 对审核过程中发现的错误应尽可能予以纠正
2. 当发现数据中的错误不能予以纠正，或者有些数据不符合调查的要求而又无法弥补时，需要对数据进行筛选
3. 数据筛选的内容包括：
 - 将某些不符合要求的数据或有明显错误的数
据予以剔除
 - 将符合某种特定条件的数据筛选出来，而不
符合特定条件的数据予以剔除

数据的排序 (要点)

1. 按一定顺序将数据排列，以发现一些明显的特征或趋势，找到解决问题的线索
2. 排序有助于对数据检查纠错，以及为重新归类或分组等提供依据
3. 在某些场合，排序本身就是分析的目的之一
4. 排序可借助于计算机完成

数据的排序 (方法)

1. 定类数据的排序

- 字母型数据，排序有升序降序之分，但习惯上用升序
- 汉字型数据，可按汉字的首位拼音字母排列，也可按笔画排序，其中也有笔画多少的升序降序之分

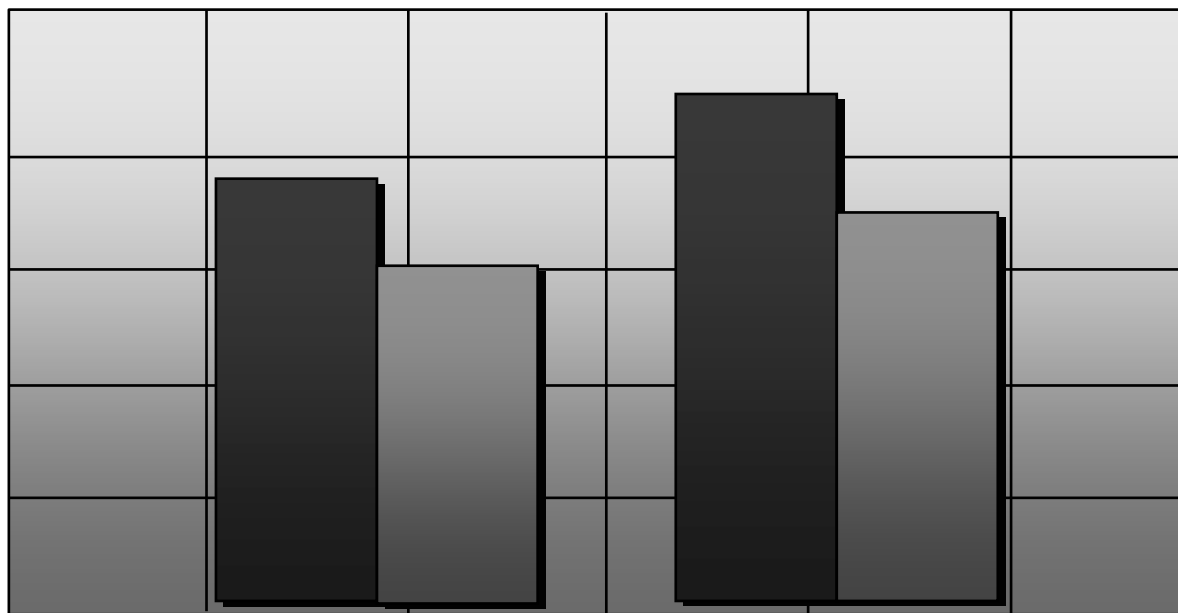
2. 定距和定比数据的排序

- 递增排序：设一组数据为 X_1, X_2, \dots, X_N ，递增排序后可表示为： $X_{(1)} < X_{(2)} < \dots < X_{(N)}$
- 递减排序可表示为： $X_{(1)} > X_{(2)} > \dots > X_{(N)}$

第二节 品质数据的整理与显示

- 一. 定类数据的整理与显示
- 二. 定序数据的整理与显示

定类数据的整理与显示



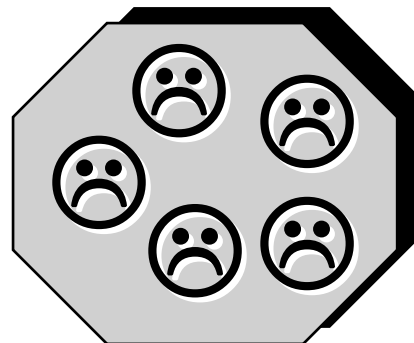
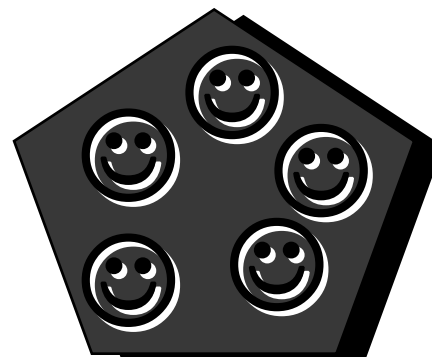
定类数据的整理与显示

（基本问题）

1. 要弄清所面对的数据类型，因为不同类型的数据，所采取的处理方式和方法是不同的
2. 对定类数据和定序数据主要是做分类整理
3. 对定距数据和定比数据则主要是做分组整理
4. 适合于低层次数据的整理和显示方法也适合于高层次的数据；但适合于高层次数据的整理和显示方法并不适合于低层次的数据

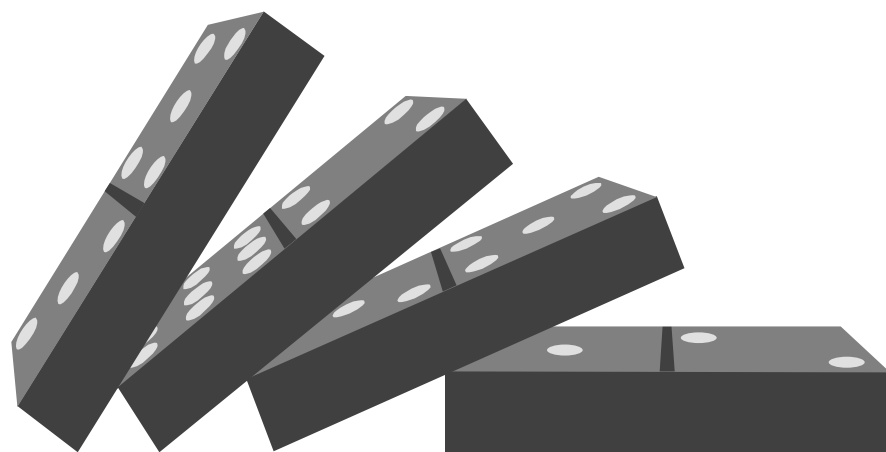
定类数据的整理 (基本过程)

1. 列出各类别
2. 计算各类别的频数
3. 制作频数分布表
4. 用图形显示数据



定类数据的整理 (可计算的指标)

1. 频 数：落在各类别中的数据个数
2. 比 例：某一类别数据占全部数据的比值
3. 百分比：将对比的基数作为100而计算的比值
4. 比 率：不同类别数值的比值



定类数据整理—频数分布表 (实例)

【例3.1】为研究广告市场的状况，一家广告公司在某城市随机抽取200人就广告问题做了邮寄问卷调查，其中的一个问题是“您比较关心下列哪一类广告？”

1. 商品广告；2. 服务广告；3. 金融广告；4. 房地产广告；5. 招生招聘广告；6. 其他广告。

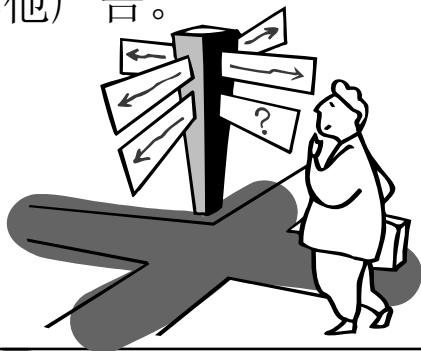


表3-1 某城市居民关注广告类型的频数分布

广告类型	人数(人)	比例	频率(%)
商品广告	112	0.560	56.0
服务广告	51	0.255	25.5
金融广告	9	0.045	4.5
房地产广告	16	0.080	8.0
招生招聘广告	10	0.050	5.0
其他广告	2	0.010	1.0
合计	200	1	100

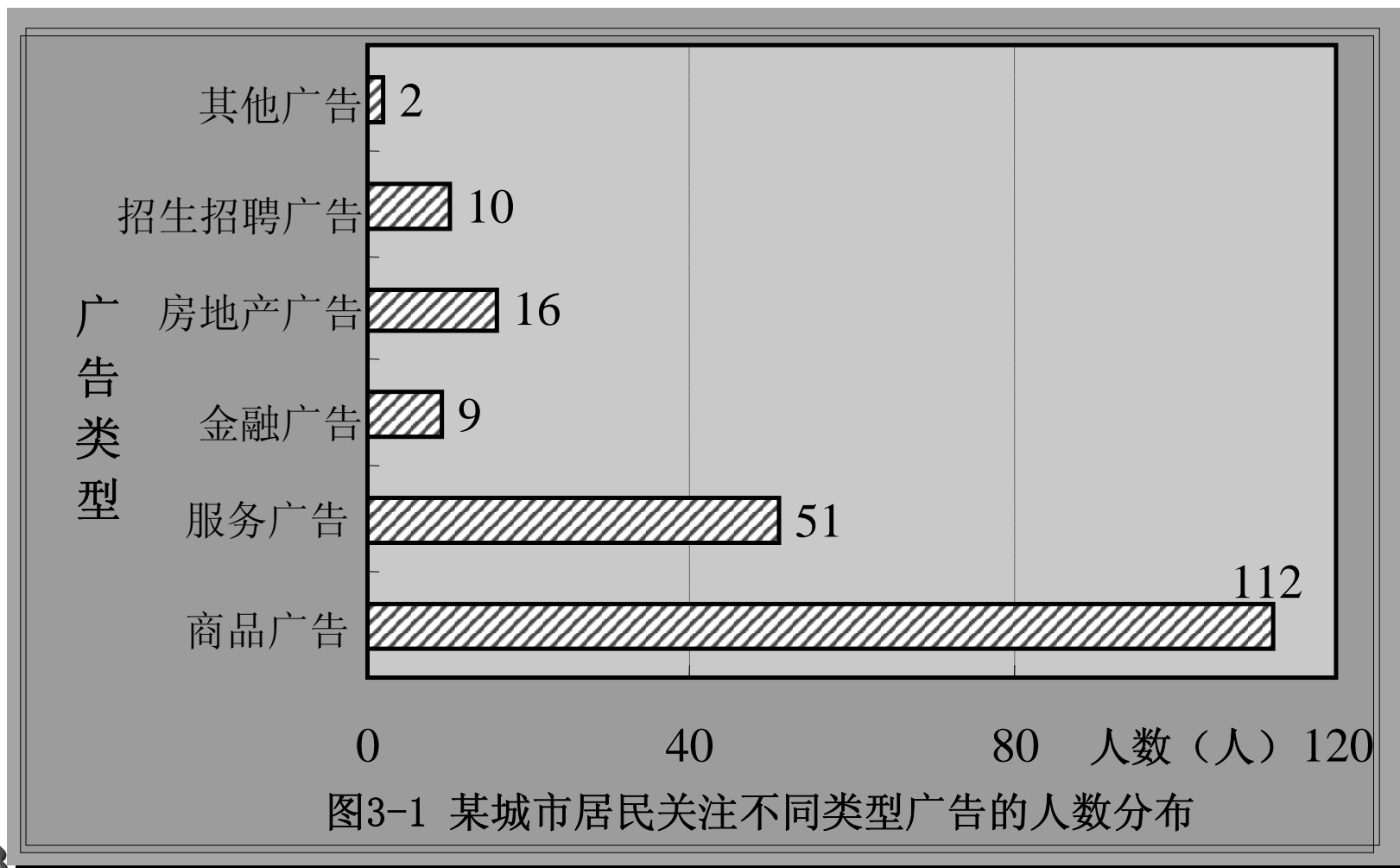
定类数据的图示—条形图

（条形图的制作）

1. 条形图是用宽度相同的条形的高度或长短来表示数据变动的图形
2. 条形图有单式、复式等形式
3. 在表示定类数据的分布时，是用条形图的高度来表示各类别数据的频数或频率
4. 绘制时，各类别可以放在纵轴，称为条形图，也可以放在横轴，称为柱形图

定类数据的图示—条形图

（由 Excel 绘制的条形图）



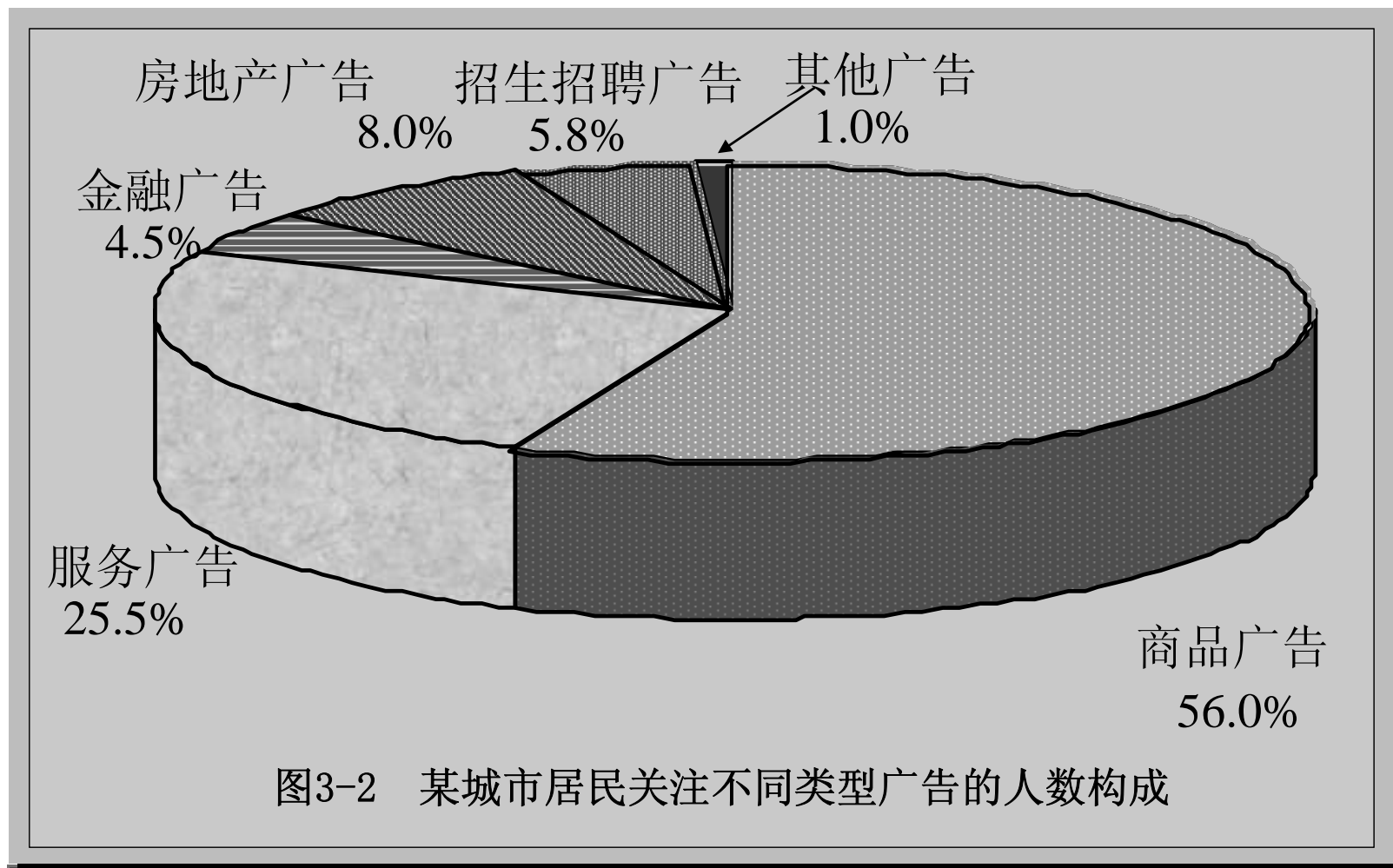
定类数据的图示—圆形图

（圆形图的制作）

1. 也称饼图，是用圆形及园内扇形的面积来表示数值大小的图形
2. 主要用于表示总体中各组成部分所占的比例，对于研究结构性问题十分有用
3. 在绘制圆形图时，总体中各部分所占的百分比用园内的各个扇形面积表示，这些扇形的中心角度，是按各部分百分比占 360^0 的相应比例确定的
4. 例如，关注服务广告的人数占总人数的百分比为25.5%，那么其扇形的中心角度就应为 $360^0 \times 25.5\% = 91.8^0$ ，其余类推

定类数据的图示—圆形图

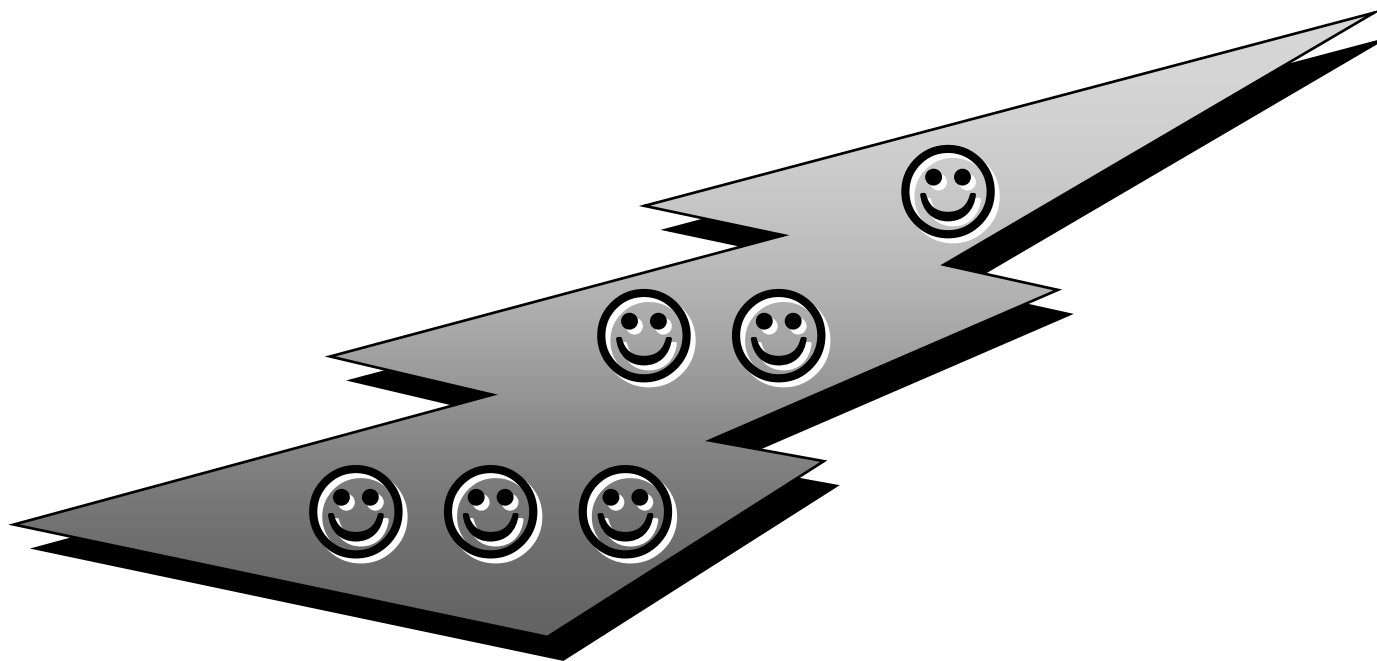
（由 Excel 绘制的圆形图）



定序数据的整理与显示

定序数据的整理 (可计算的指标)

1. 累计频数：将各类别的频数逐级累加
2. 累计频率：将各类别的频率(百分比)逐级累加



定序数据频数分布表 (实例)

【例3.2】 在一项城市住房问题的研究中，研究人员在甲乙两个城市各抽样调查300户，其中的一个问题是：“您对您家庭目前的住房状况是否满意？”

1. 非常不满意；
2. 不满意；
3. 一般；
4. 满意；
5. 非常满意。

表3-2 甲城市家庭对住房状况评价的频数分布

回答类别	甲城市					
	户数 (户)	百分比 (%)	向上累积		向下累积	
			户数 (户)	百分比 (%)	户数 (户)	百分比 (%)
非常不满意	24	8	24	8.0	300	100.0
不满意	108	36	132	44.0	276	92
一般	93	31	225	75.0	168	56
满意	45	15	270	90.0	75	25
非常满意	30	10	300	100.0	30	10
合计	300	100.0	—	—	—	—

定序数据频数分布表 (实例)

表3-3 乙城市家庭对住房状况评价的频数分布

回答类别	乙城市					
	户数 (户)	百分比 (%)	向上累积		向下累积	
			户数 (户)	百分比 (%)	户数 (户)	百分比 (%)
非常不满意	21	7.0	21	7.0	300	100.0
不满意	99	33.0	120	40.0	279	93.0
一般	78	26.0	198	66.0	180	60.0
满意	64	21.3	262	87.3	102	34.0
非常满意	38	12.7	300	100.0	38	12.7
合计	300	100.0	—	—	—	—



定序数据的图示—累计频数分布图 (由 Excel 绘制的累计频数分布图)

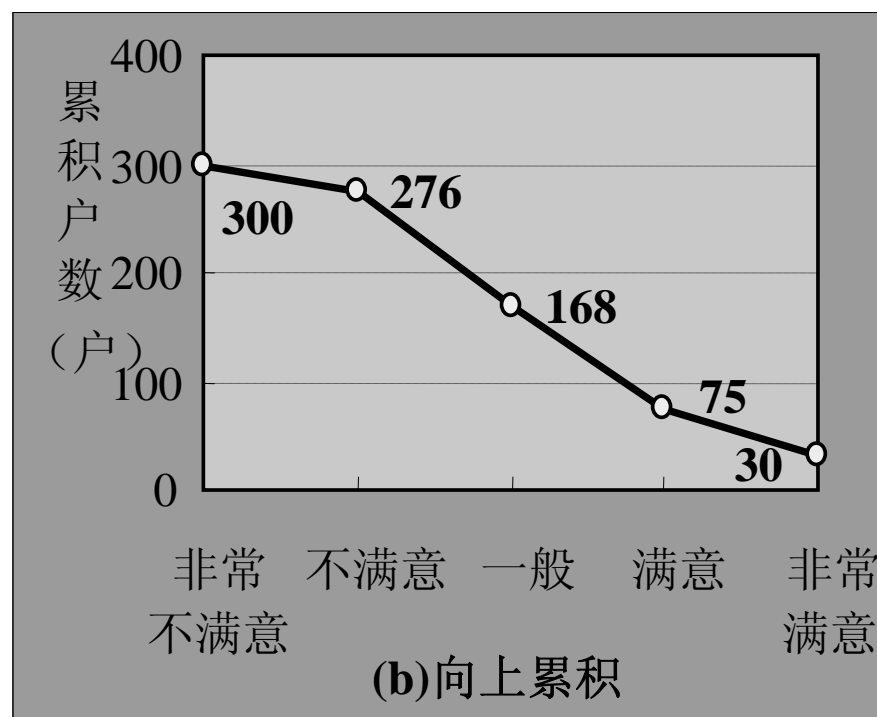
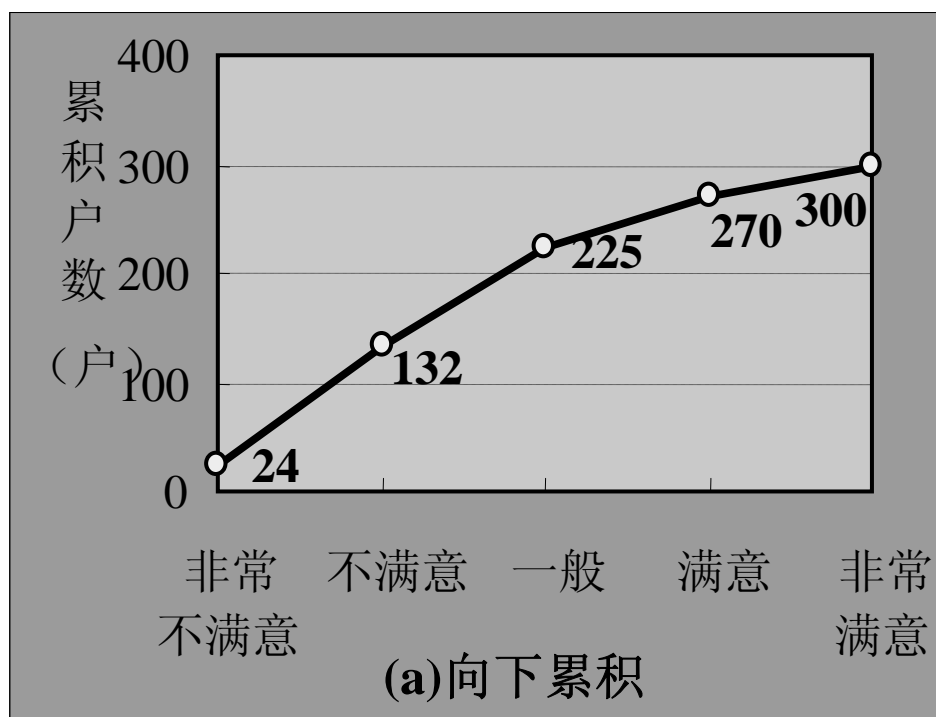


图3-3 甲城市家庭对住房状况评价的累积频数分布

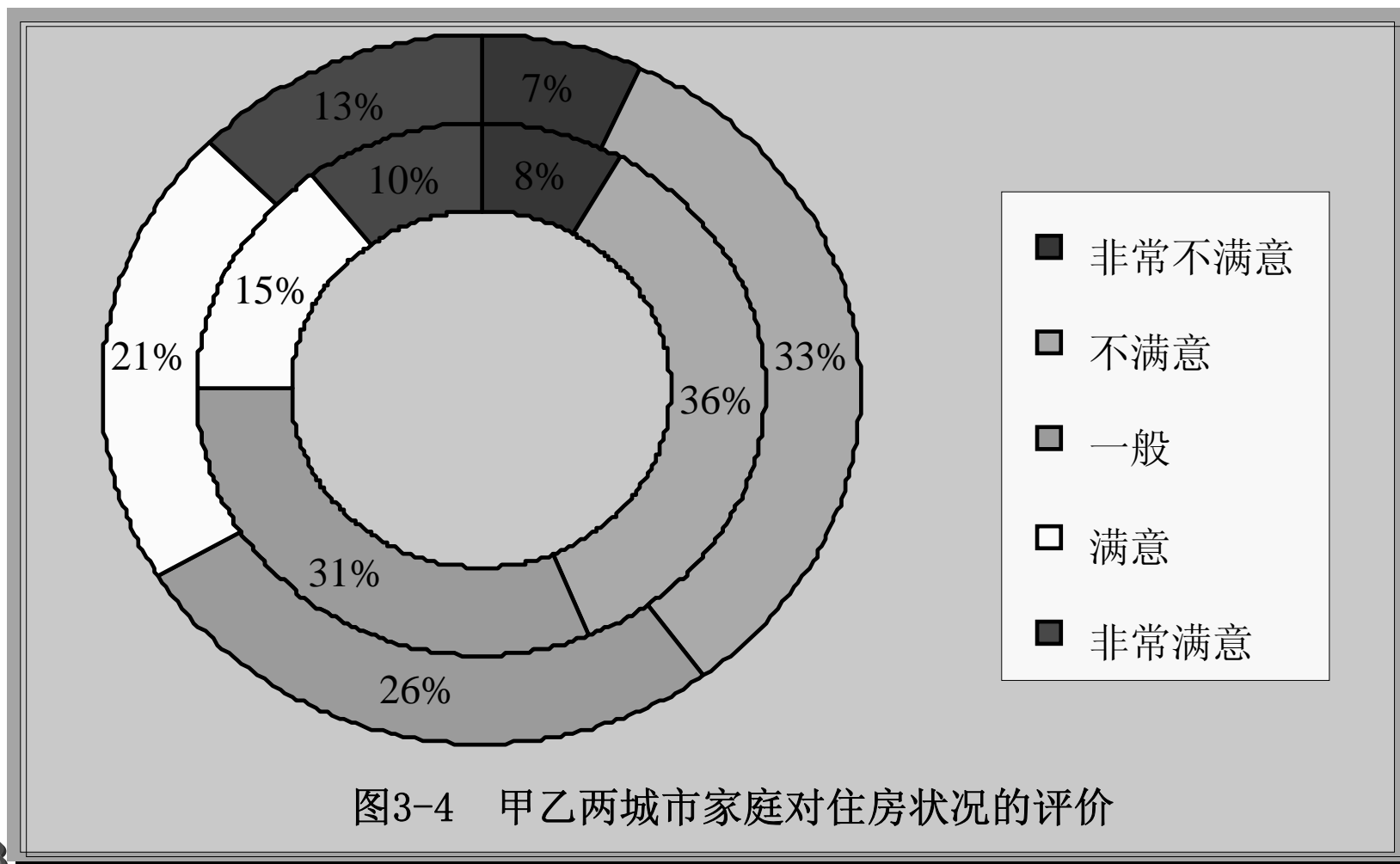
定类数据的图示—环形图

（环形图的制作）

1. 环形图中间有一个“空洞”，总体中的每一部分数据用环中的一段表示
2. 环形图与圆形图类似，但又有区别
 - 圆形图只能显示一个总体各部分所占的比例
 - 环形图则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
3. 环形图可用于进行比较研究
4. 环形图可用于展示定类和定序的数据

品质数据的图示—环形图

(由 Excel 绘制的环形图)

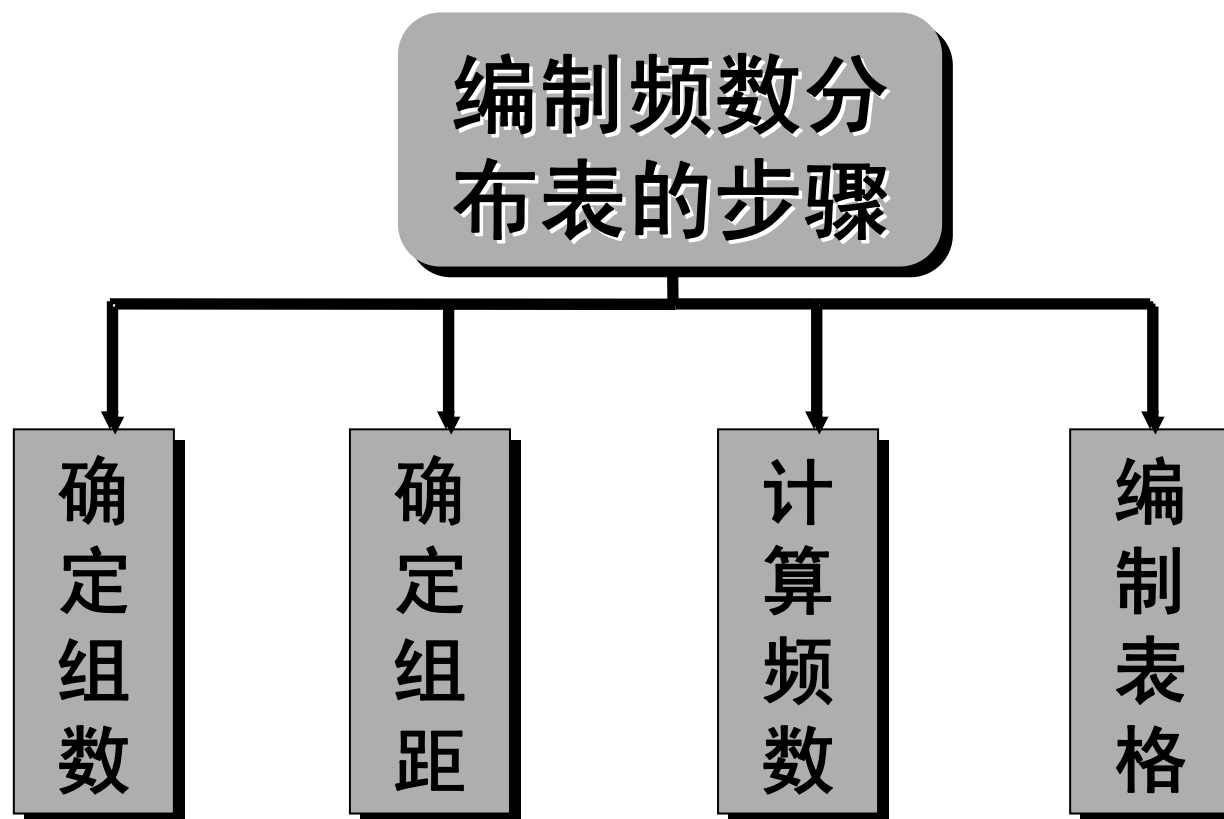


第三节 数值型数据的整理与显示

- 一. 数据的分组
- 二. 数值型数据的图示
- 三. 频数分布的类型

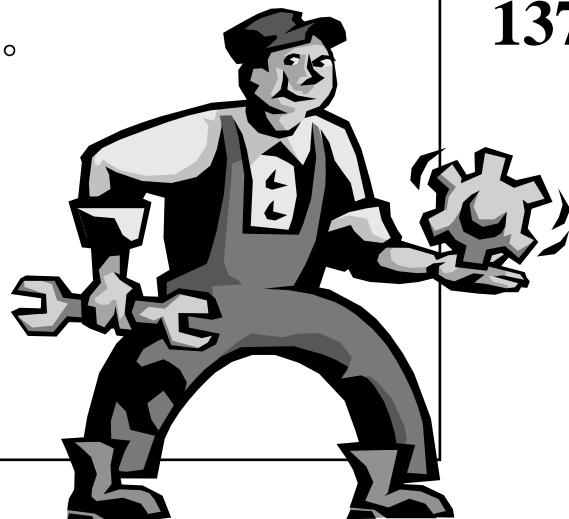
频数分布表的编制

编制频数分布表的步骤



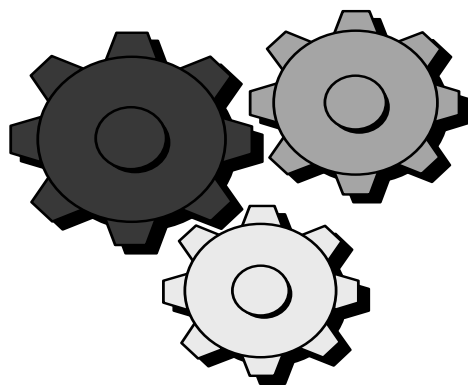
频数分布表的编制 (实例)

【例3.3】某生产车间50名工人日加工零件数如下（单位：个）。试采用单变量值对数据进行分组。

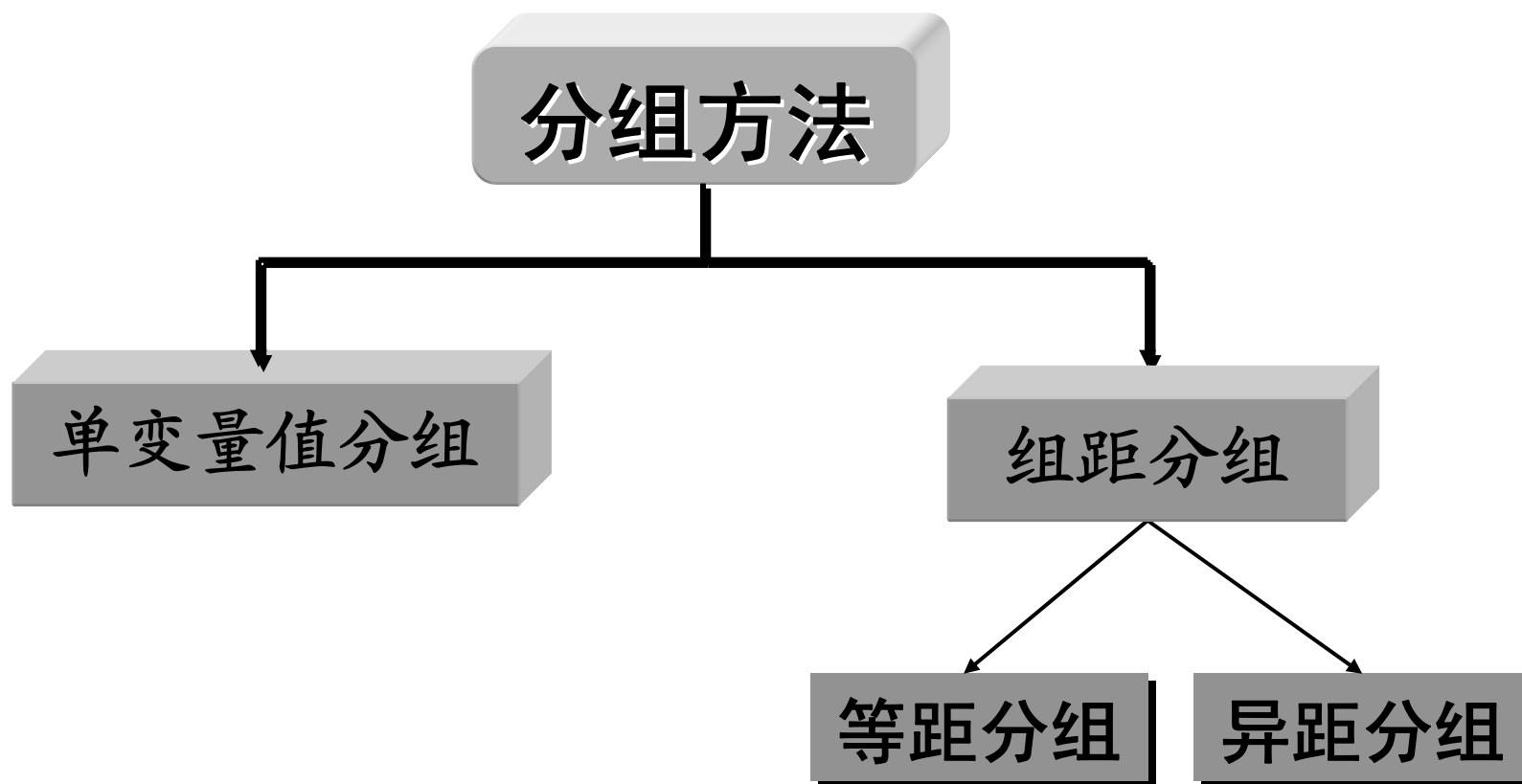


3 - 32

117	122	124	129	139	107	117	130	122	125
108	131	125	117	122	133	126	122	118	108
110	118	123	126	133	134	127	123	118	112
112	134	127	123	119	113	120	123	127	135
137	114	120	128	124	115	139	128	124	121



分组方法



单变量值分组 (要点)

1. 将一个变量值作为一组
2. 适合于离散变量
3. 适合于变量值较少的情况



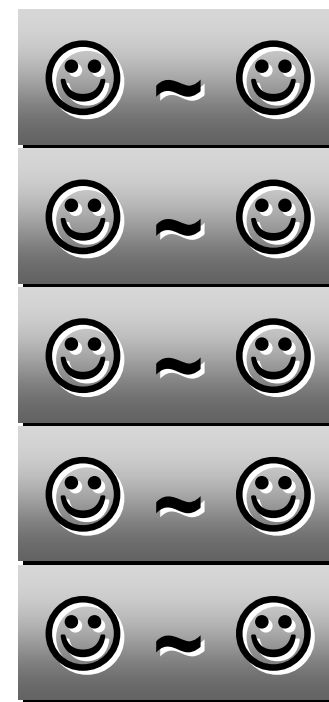
单变量值分组表 (实例)

表3-4 某车间50名工人日加工零件数分组表

零件数 (个)	频数 (人)	零件数 (个)	频数 (人)	零件数 (个)	频数 (人)
107	1	119	1	128	2
108	2	120	2	129	1
110	1	121	1	130	1
112	2	122	4	131	1
113	1	123	4	133	2
114	1	124	3	134	2
115	1	125	2	135	1
117	3	126	2	137	1
118	3	127	3	139	2

组距分组 (要点)

1. 将变量值的一个区间作为一组
2. 适合于连续变量
3. 适合于变量值较多的情况
4. 必须遵循“不重不漏”的原则
5. 可采用等距分组，也可采用不等距分组



组距分组 (步骤)

1. 确定组数：组数的确定应以能够显示数据的分布特征和规律为目的。在实际分组时，可以按 Sturges 提出的经验公式来确定组数 K

$$K = 1 + \frac{\lg(n)}{\lg(2)}$$

2. 确定各组的组距：组距(Class Width)是一个组的上限与下限之差，可根据全部数据的最大值和最小值及所分的组数来确定，即

$$\text{组距} = (\text{最大值} - \text{最小值}) \div \text{组数}$$

3. 根据分组整理成频数分布表

组距分组 (几个概念)

1. 下 限：一个组的最小值
2. 上 限：一个组的最大值
3. 组 距：上限与下限之差
4. 组中值：下限与上限之间的中点值

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

等距分组表 (上下组限重叠)

表3-5 某车间50名工人日加工零件数分组表		
按零件数分组	频数 (人)	频率 (%)
105~110	3	6
110~115	5	10
115~120	8	16
120~125	14	28
125~130	10	20
130~135	6	12
135~140	4	8
合计	50	100

等距分组表 (上下组限间断)

表3-6 某车间50名工人日加工零件数分组表		
按零件数分组	频数 (人)	频率 (%)
105~109	3	6
110~114	5	10
115~119	8	16
120~124	14	28
125~129	10	20
130~134	6	12
135~139	4	8
合计	50	100

等距分组表 (使用开口组)

表3-7 某车间50名工人日加工零件数分组表		
按零件数分组	频数 (人)	频率 (%)
110以下	3	6
110~114	5	10
115~119	8	16
120~124	14	28
125~129	10	20
130~134	6	12
135以上	4	8
合计	50	100

组距分组与不等距分组 (在表现频数分布上的差异)

1. 等距分组

- 各组频数的分布不受组距大小的影响
- 可直接根据绝对频数来观察频数分布的特征和规律

2. 不等距分组

- 各组频数的分布受组距大小不同的影响
- 各组绝对频数的多少不能反映频数分布的实际状况
- 需要用频数密度（ $\text{频数密度} = \text{频数} / \text{组距}$ ）反映频数分布的实际状况

数值型数据的图示 用Excel作图

以下图形均由
计算机绘制!



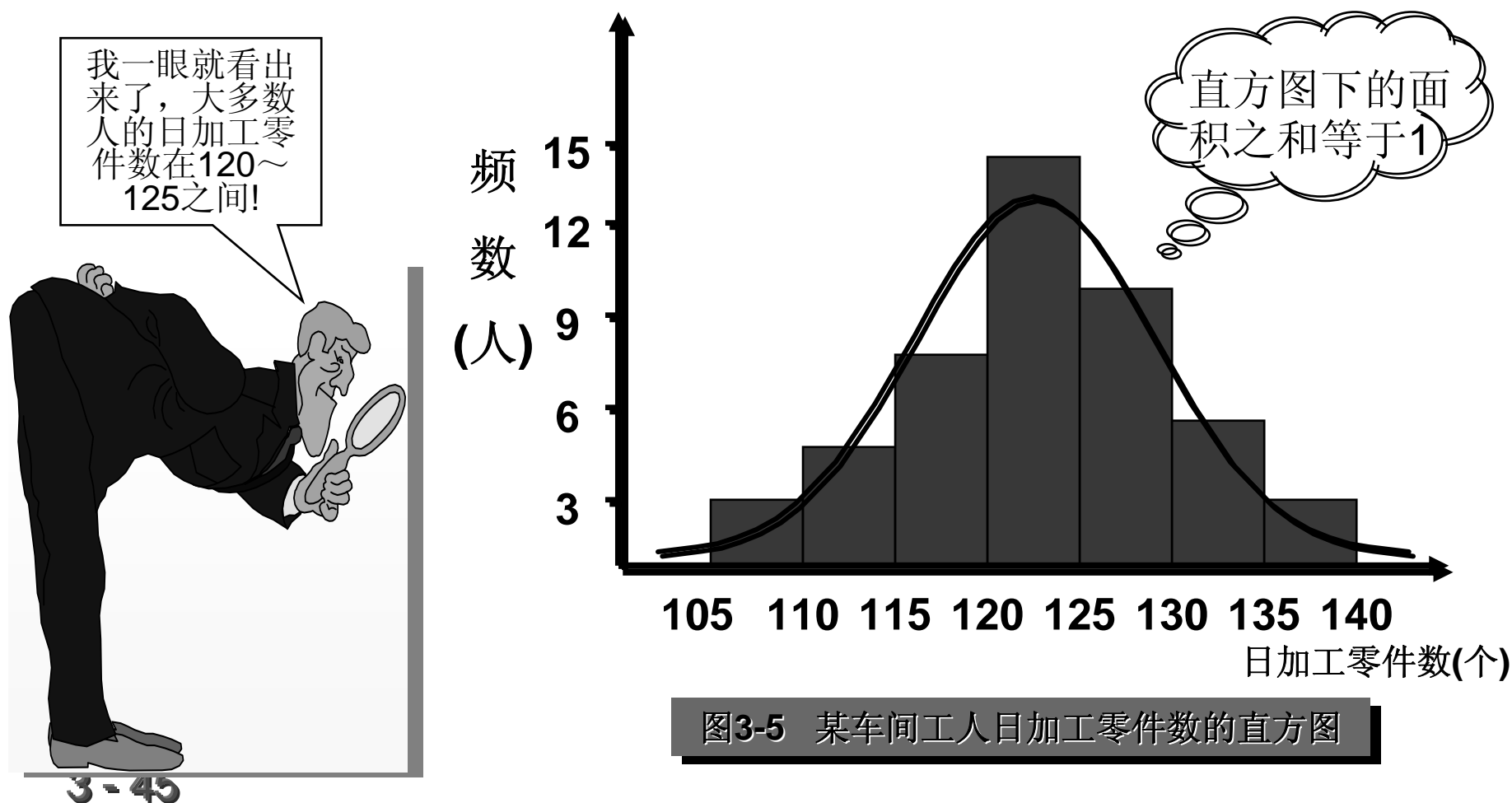
分组数据—直方图

（直方图的制作）

1. 用矩形的宽度和高度来表示频数分布的图形，实际上是用矩形的 **面积** 来表示各组的频数分布
2. 在直角坐标中，用横轴表示数据分组，纵轴表示频数或频率，各组与相应的频数就形成了一个矩形，即直方图(Histogram)
3. 直方图下的总面积等于1

分组数据—直方图

（直方图的绘制）



分组数据—直方图

(直方图与条形图的区别)

1. 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
2. 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
3. 直方图的各矩形通常是连续排列, 条形图则是分开排列

分组数据－折线图

（折线图的制作）

1. 折线图也称频数多边形图(Frequency polygon)
2. 是在直方图的基础上，把直方图顶部的中点(组中值)用直线连接起来，再把原来的直方图抹掉
3. 折线图的两个终点要与横轴相交，具体的做法是
 - 第一个矩形的顶部中点通过竖边中点（即该组频数一半的位置）连接到横轴，最后一个矩形顶部中点与其竖边中点连接到横轴
 - 折线图下所围成的面积与直方图的面积相等，二者所表示的频数分布是一致的

分组数据—折线图 (折线图的绘制)

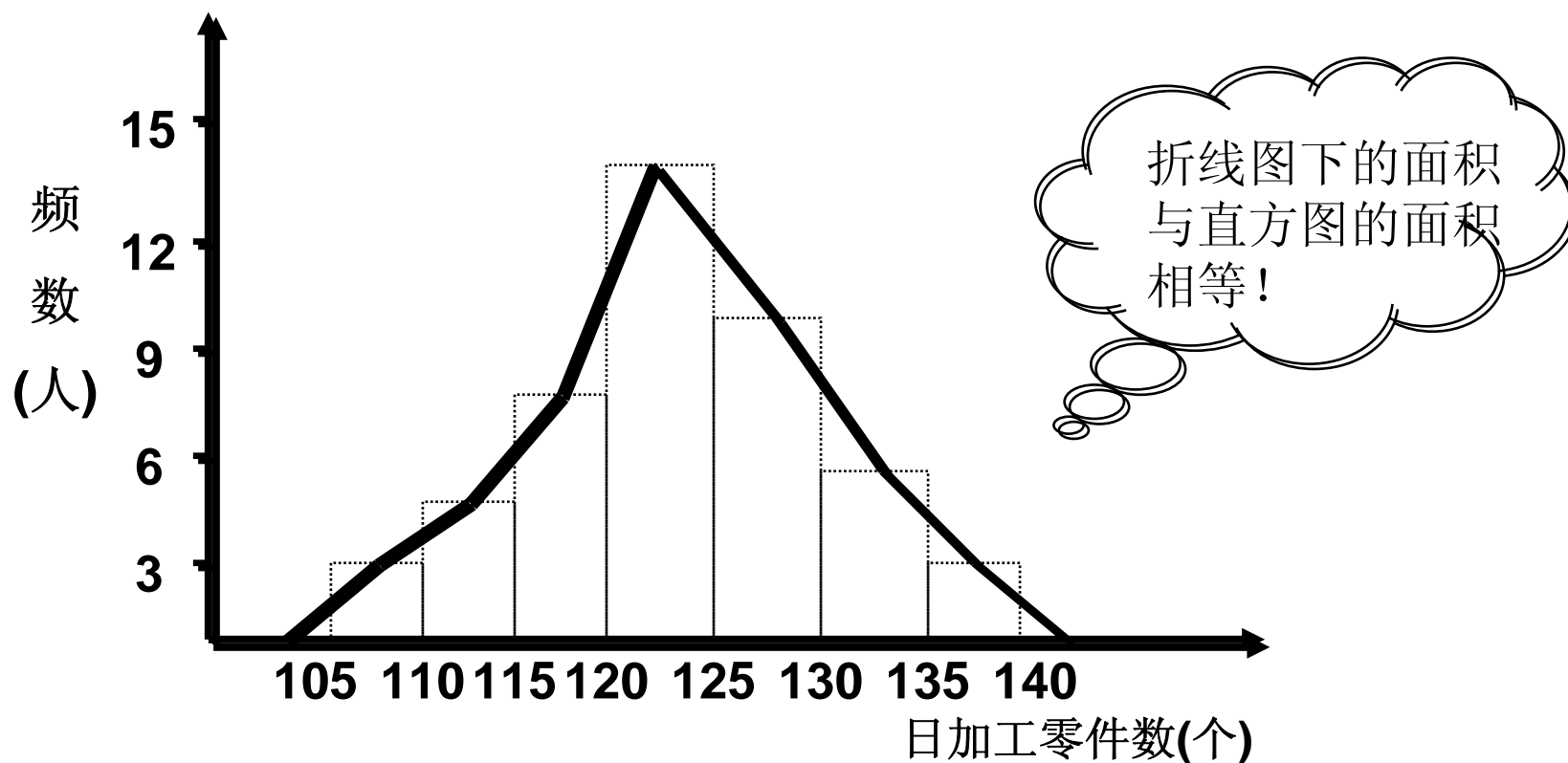


图3-6 某车间工人日加工零件数的折线图

未分组数据—茎叶图

（茎叶图的制作）

1. 用于显示未分组的原始数据的分布
2. 由“茎”和“叶”两部分构成，其图形是由数字组成的
3. 以该组数据的高位数值作树茎，低位数字作树叶
4. 对于 $n(20 \leq n \leq 300)$ 个数据，茎叶图最大行数不超过

$$L = [10 \times \log_{10} n]$$

5. 茎叶图类似于横置的直方图，但又有区别
 - 直方图可大体上看出一组数据的分布状况，但没有给出具体的数值
 - 茎叶图既能给出数据的分布状况，又能给出每一个原始数值，保留了原始数据的信息

未分组数据—茎叶图

（茎叶图的制作）



图3-7 某车间工人日加工零件数的茎叶图

未分组数据—茎叶图 (扩展的茎叶图)

树茎 | 树叶

10*

10. 7 8 8

11* 0 2 2 3 4

11. 5 7 7 7 8 8 8 9

12* 0 0 1 2 2 2 2 3 3 3 3 4 4 4

12. 5 5 6 6 7 7 7 8 8 9

13* 0 1 3 3 4 4

13. 5 7 9 9

树茎 | 树叶

10s 7

10. 8 8

11* 0

11t 2 2 3

11f 4 5

11s 7 7 7

11. 8 8 8 9

12* 0 0 1

12t 2 2 2 2 3 3 3 3

12f 4 4 4 5 5

12s 6 6 7 7 7

12. 8 8 9

13* 0 1

12t 3 3

13f 4 4 5

13s 7

13. 9 9

图3-8 图3.7扩展后的茎叶图

未分组数据—箱线图

（箱线图的制作）

1. 用于显示未分组的原始数据或分组数据的分布
2. 箱线图由一组数据的5个特征值绘制而成，它由一个箱子和两条线段组成
3. 其绘制方法是：
 - 首先找出一组数据的5个特征值，即最大值、最小值、中位数 Me 和两个四分位数(下四分位数 Q_L 和上四分位数 Q_U)
 - 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

未分组数据—单批数据箱线图 (箱线图的构成)

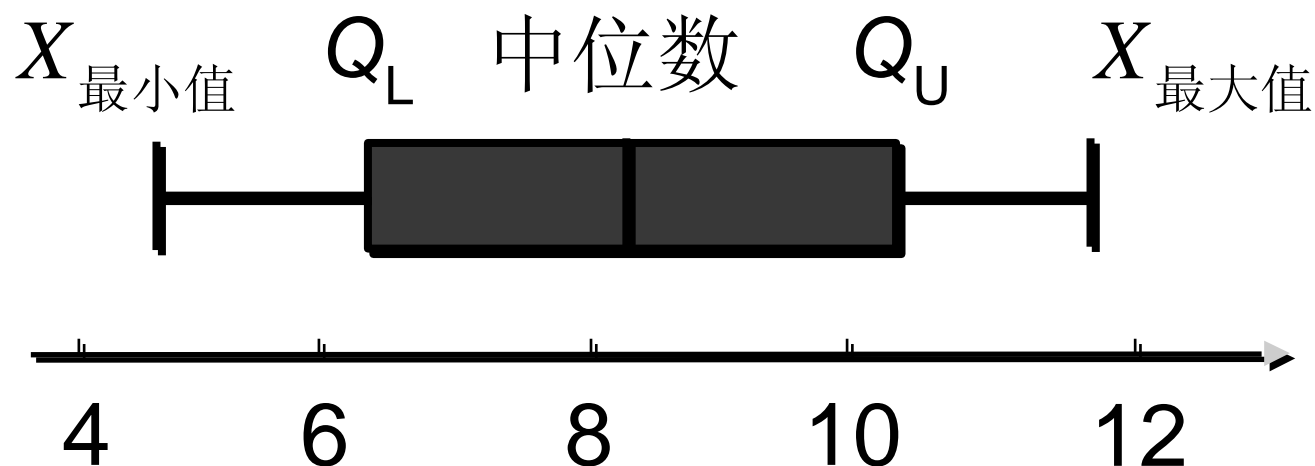


图3-9 简单箱线图

未分组数据—单批数据箱线图 (实例)

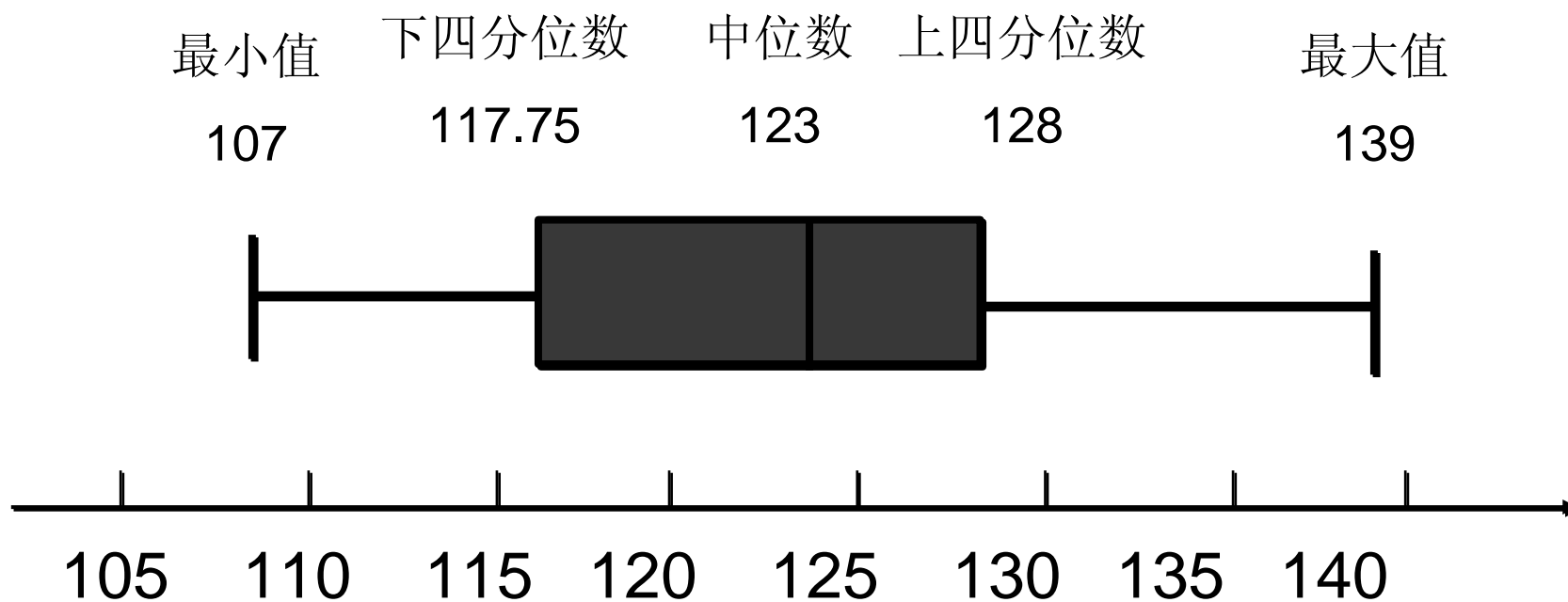
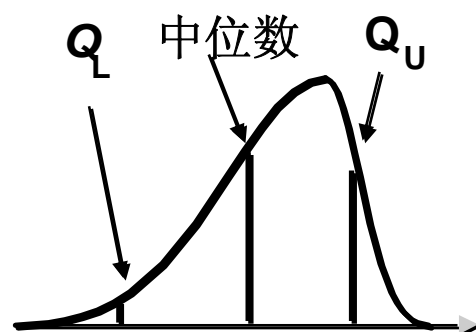
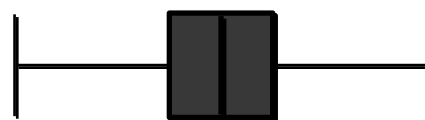
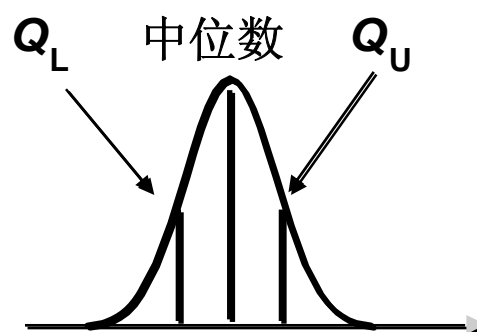


图3-10 50名工人日加工零件数的箱线图

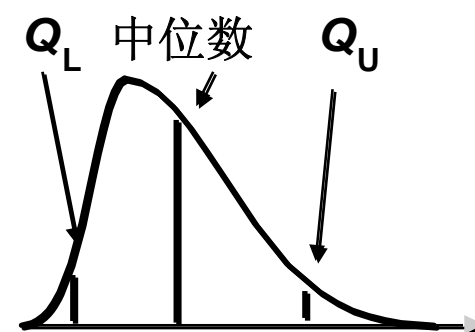
分布的形状与箱线图



左偏分布



对称分布



右偏分布

图3-11 不同分布的箱线图

未分组数据—多批数据箱线图 (实例)

【例 3.4】 从某大学经济管理学院二年级学生中随机抽取 11 人，对 8 门主要课程的考试成绩进行调查，所得结果如表 3-8。试绘制各科考试成绩的批比较箱线图，并分析各科考试成绩的分布特征

表 3-8 11 名学生各科的考试成绩数据

课程名称	学生编号										
	1	2	3	4	5	6	7	8	9	10	11
英语	76	90	97	71	70	93	86	83	78	85	81
经济数学	65	95	51	74	78	63	91	82	75	71	55
西方经济学	93	81	76	88	66	79	83	92	78	86	78
市场营销学	74	87	85	69	90	80	77	84	91	74	70
财务管理	68	75	70	84	73	60	76	81	88	68	75
基础会计学	70	73	92	65	78	87	90	70	66	79	68
统计学	55	91	68	73	84	81	70	69	94	62	71
计算机应用基础	85	78	81	95	70	67	82	72	80	81	77

未分组数据—多批数据箱线图

(由STATIATICA绘制的多批数据箱线图)

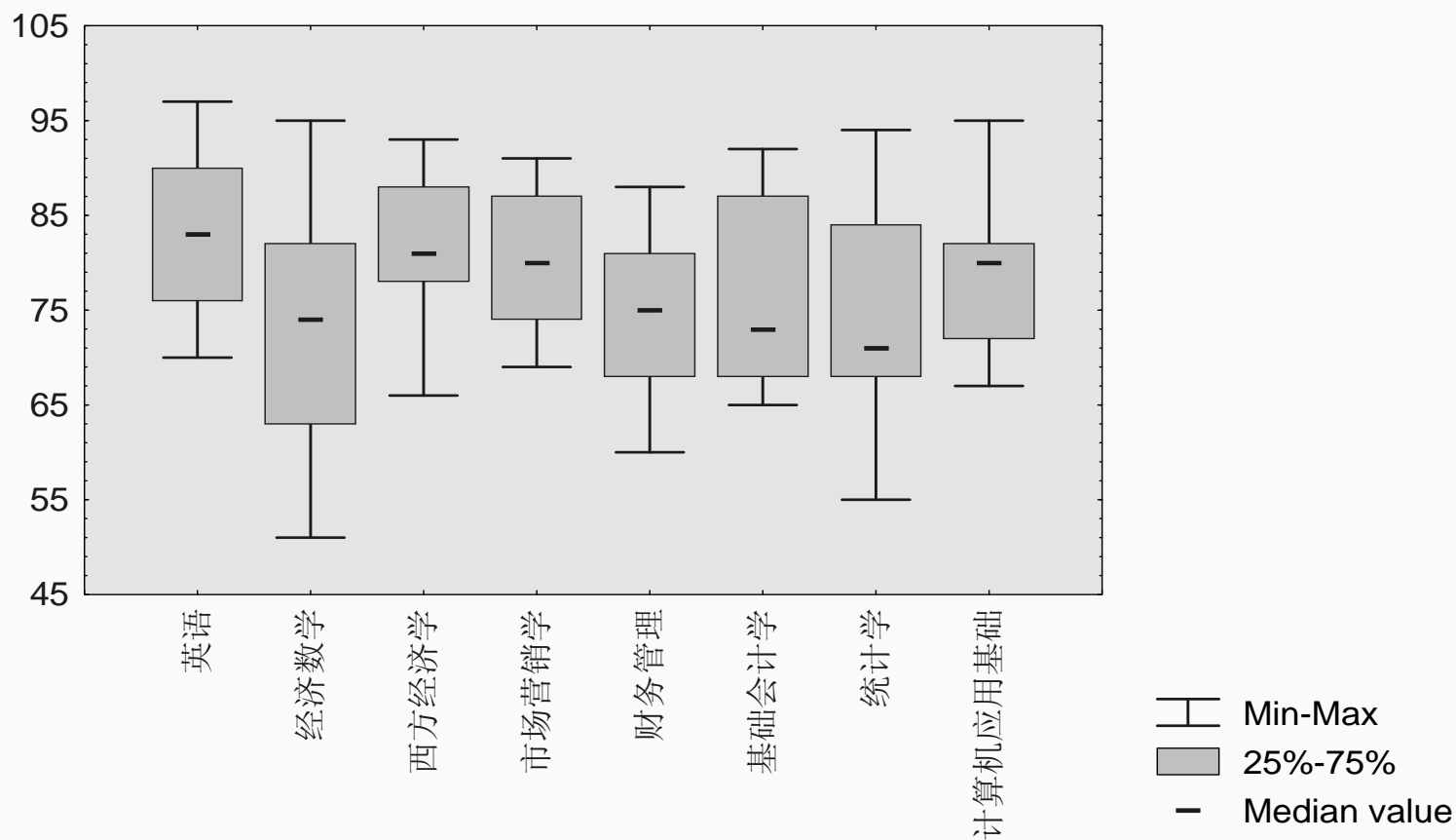


图3-12 8门课程考试成绩的箱线图

未分组数据—箱线图

(由STATIATICA绘制的多批数据箱线图)

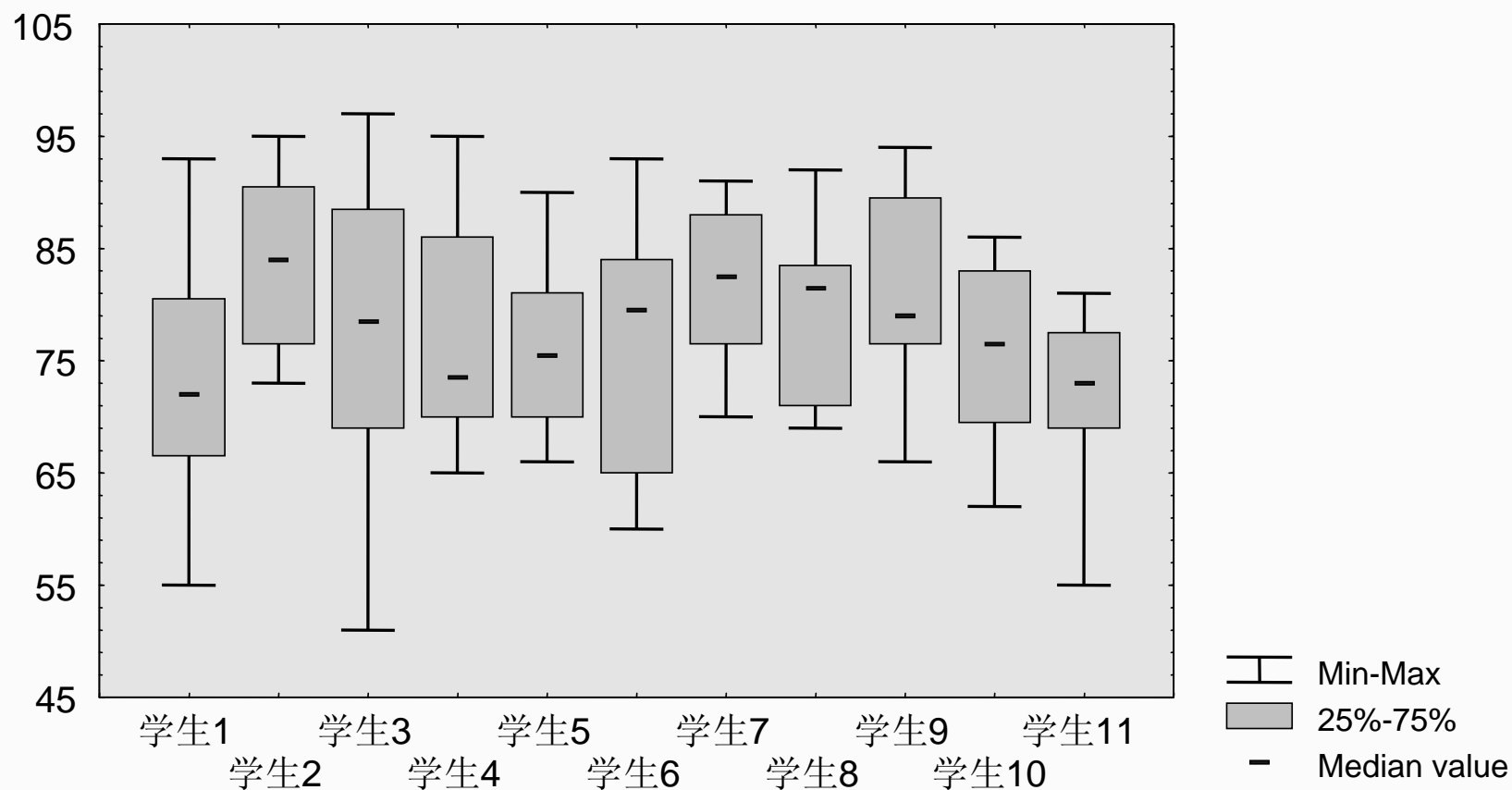


图3-13 11名学生8门课程考试成绩的箱线图

时间序列数据—线图

（线图的制作）

➡ 绘制线图时应注意以下几点

1. 时间一般绘在横轴，指标数据绘在纵轴
2. 图形的长宽比例要适当，其长宽比例大致为**10：7**
3. 一般情况下，纵轴数据下端应从“**0**”开始，以便于比较。数据与“**0**”之间的间距过大时，可以采取折断的符号将纵轴折断

时间序列数据—线图

(实例)

【例3.5】 已知1991~1998年我国城乡居民家庭的人均收入数据如表3-11。试绘制线图

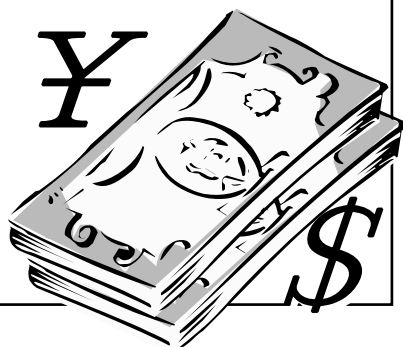
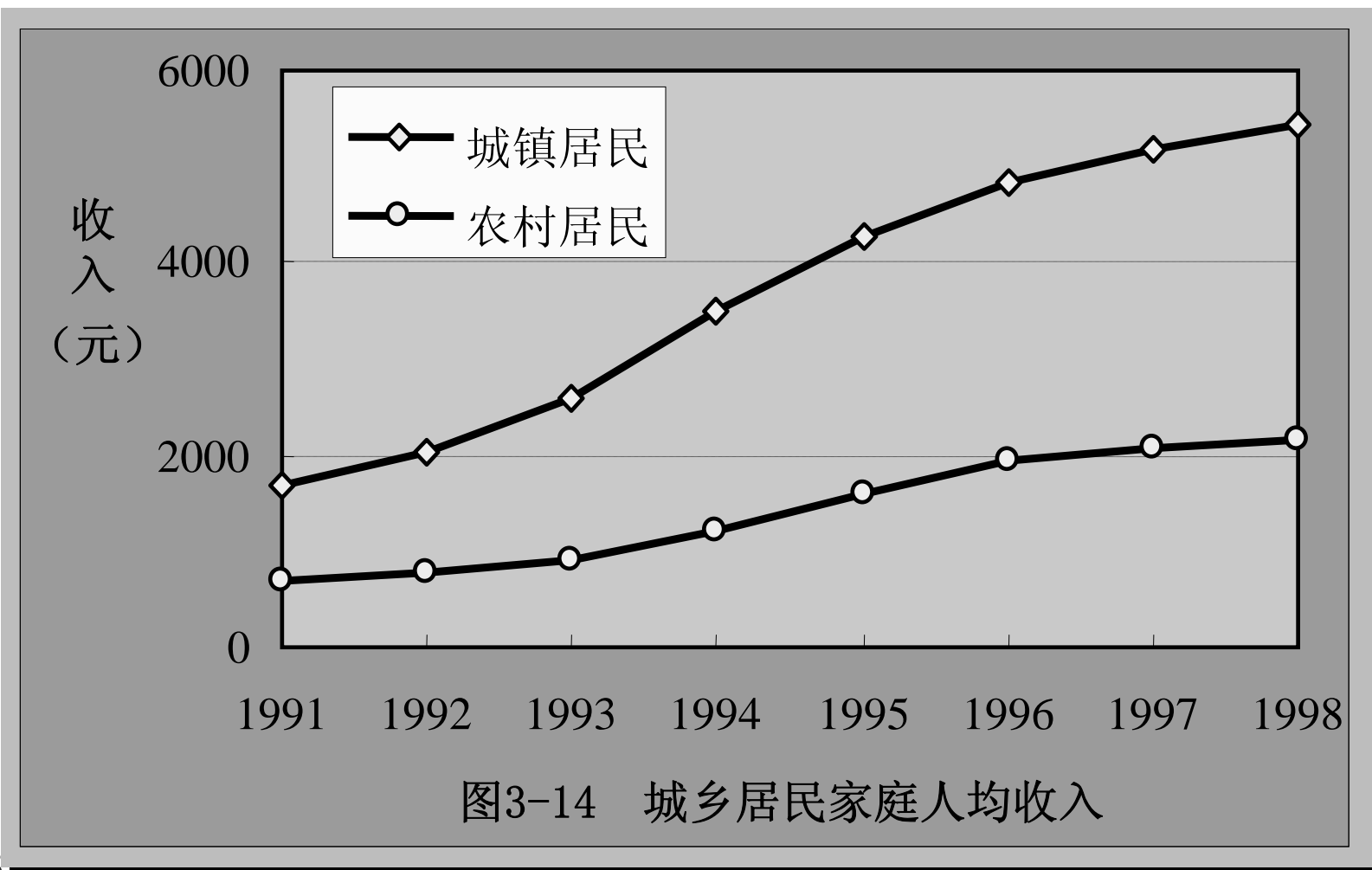


表3-11 1991~1998年城乡居民家庭人均收入

年份	城镇居民	农村居民
1991	1700.6	708.6
1992	2026.6	784.0
1993	2577.4	921.6
1994	3496.2	1221.0
1995	4283.0	1577.7
1996	4838.9	1926.1
1997	5160.3	2091.1
1998	5425.1	2162.0

时间序列数据 (由 Excel 绘制的线图)



多变量数据—雷达图

（要点）

1. 雷达图（Radar Chart）是显示多个变量的常用图示方法
3. 在显示或对比各变量的数值总和时十分有用
4. 假定各变量的取值具有相同的正负号，总的绝对值与图形所围成的区域成正比
5. 可用于研究多个样本之间的相似程度

多变量数据—雷达图

（雷达图的制作）

- ➡ 设有 n 组样本 S_1, S_2, \dots, S_n ，每个样本测得 P 个变量 X_1, X_2, \dots, X_p ，要绘制这 P 个变量的雷达图，其具体做法是
 - 先做一个圆，然后将圆 P 等分，得到 P 个点，令这 P 个点分别对应 P 个变量，在将这 P 个点与圆心连线，得到 P 个幅射状的半径，这 P 个半径分别作为 P 个变量的坐标轴，每个变量值的大小由半径上的点到圆心的距离表示
 - 再将同一样本的值在 P 个坐标上的点连线。这样， n 个样本形成的 n 个多边形就是一个雷达图

多变量数据—雷达图

（实例）

【例 3.6】1997 年我国城乡居民家庭平均每人各项生活消费支出数据如表3-12。试绘制雷达图。



今天的主
食是面包

表3-12 1997年城乡居民家庭平均每人生活消费支出

项 目	城镇居民	农村居民
食品	1942.59	890.28
衣着	520.91	109.41
家庭设备用品及服务	316.89	85.41
医疗保健	179.68	62.45
交通通讯	232.90	53.92
娱乐教育文化服务	448.38	148.18
居住	358.64	233.23
杂项商品与服务	185.65	34.27
合 计	4185.64	1617.15

多变量数据—雷达图 (由 Excel 绘制的雷达图)

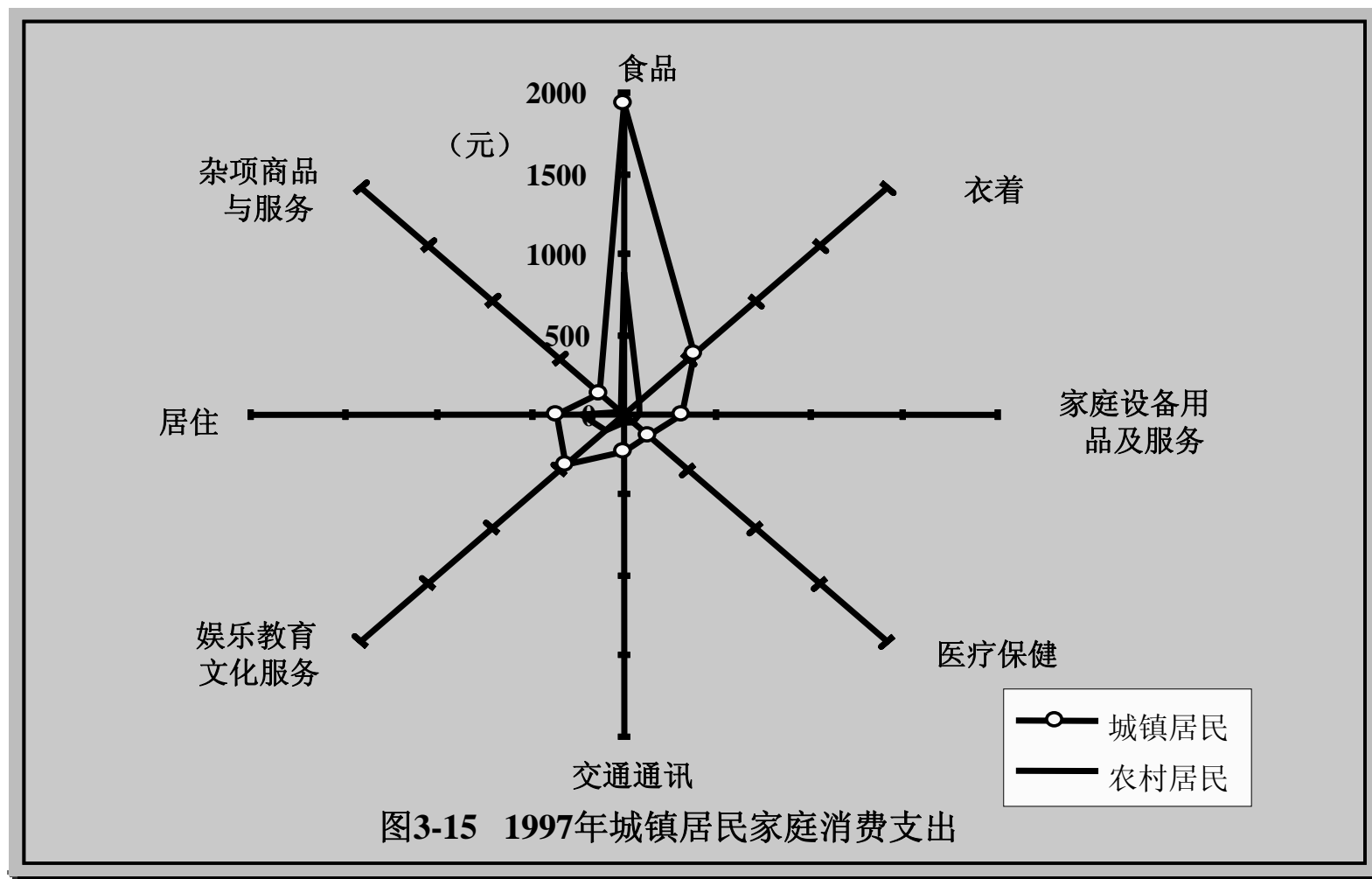


图3-15 1997年城镇居民家庭消费支出

多变量数据—雷达图

（实例）

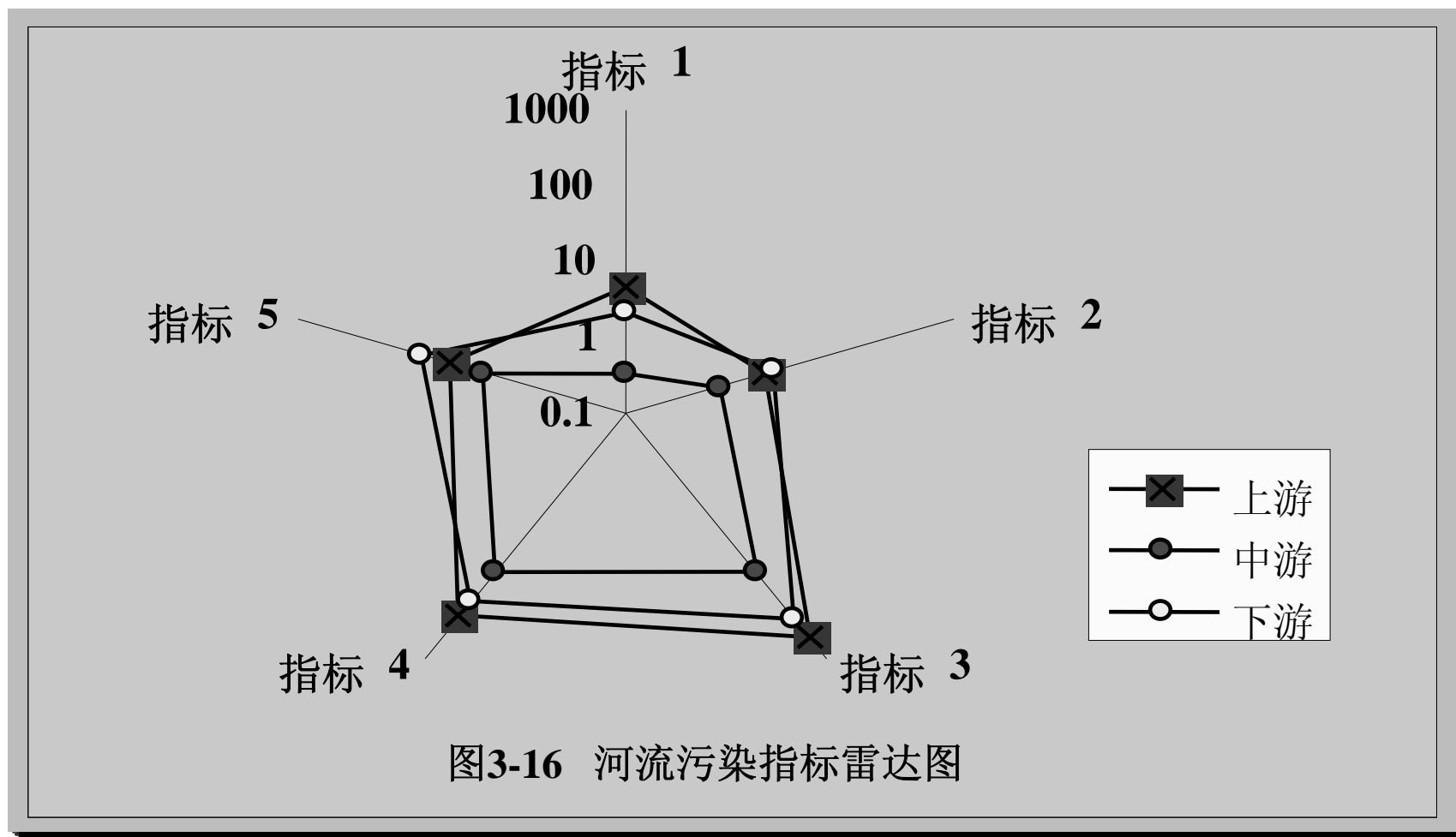
【例3.7】为研究某条河流的污染程度，环保局分别在上游、中游和下游设立取样点，每个取样点化验水中的五项污染指标，所得数据如表3-13。将各指标用雷达图表示出来，并分析该河流的主要污染源。

表3-13 不同样本点的化验指标

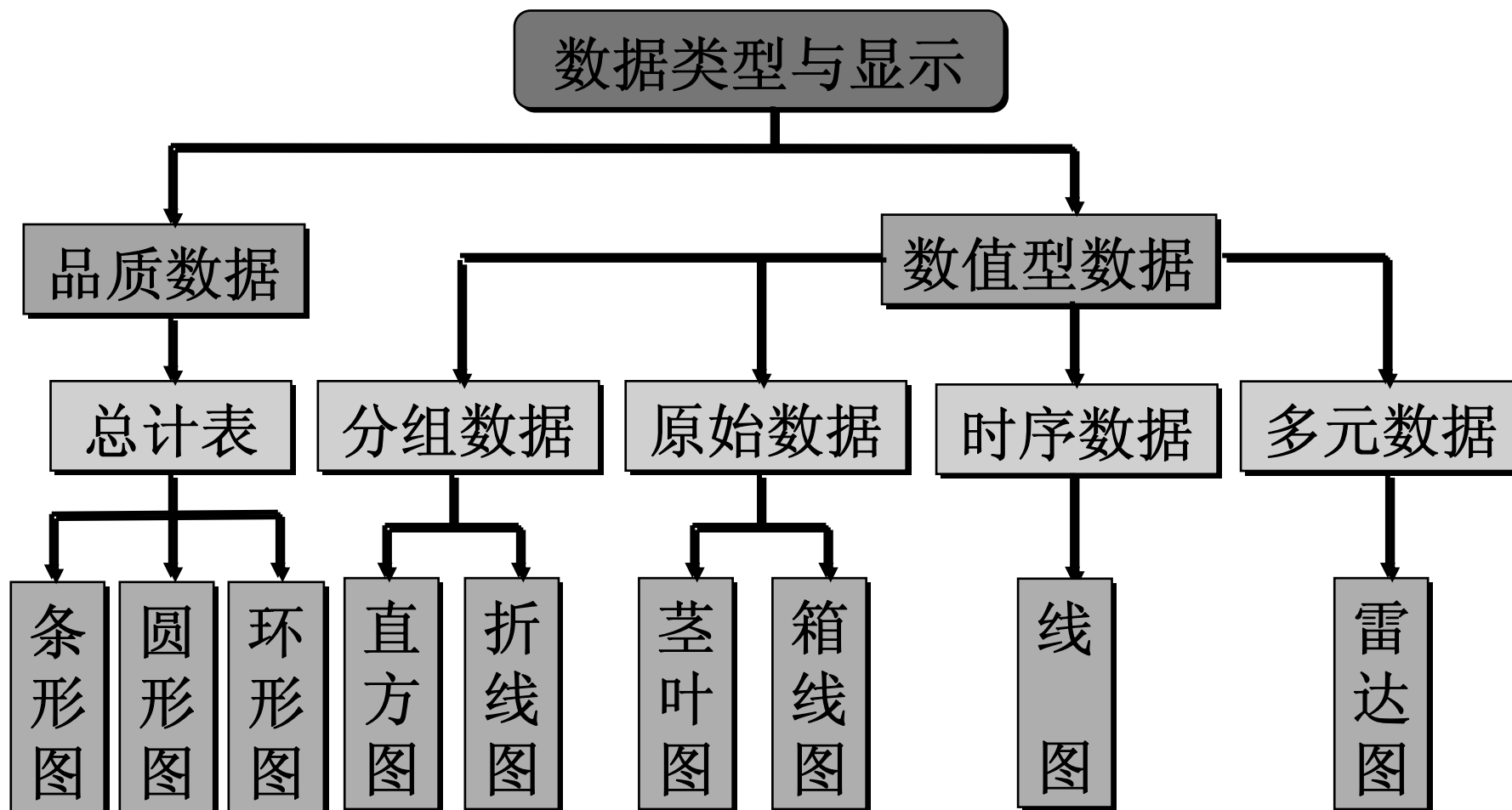
	指标1	指标2	指标3	指标4	指标5
上游	4.52	5.0	483	196	14
中游	0.34	1.4	36	41	6
下游	2.17	6.8	208	112	35

多变量数据—雷达图

(由 Excel 绘制的对数坐标雷达图)

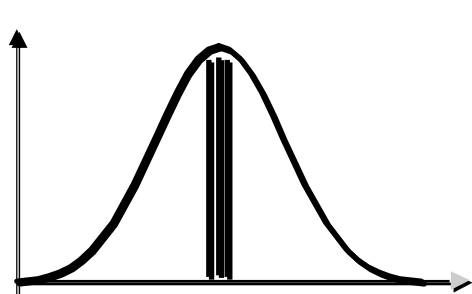


数据类型及图示 (小结)

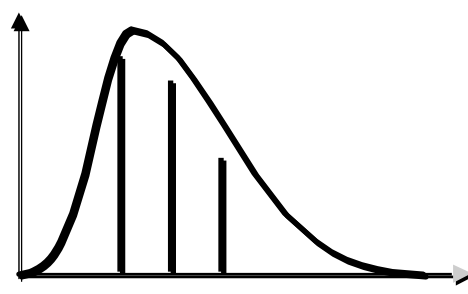


频数分布的类型

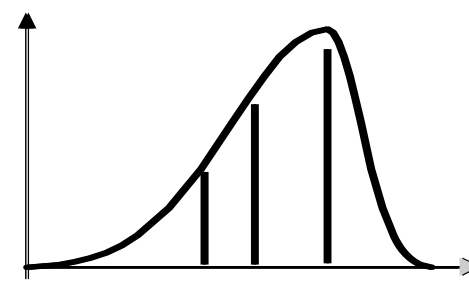
频数分布的类型



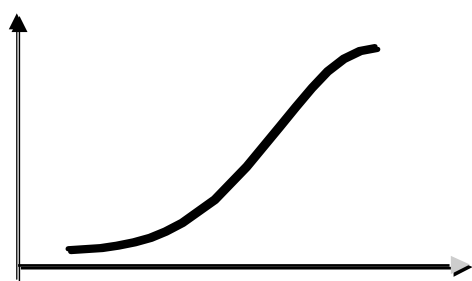
对称分布



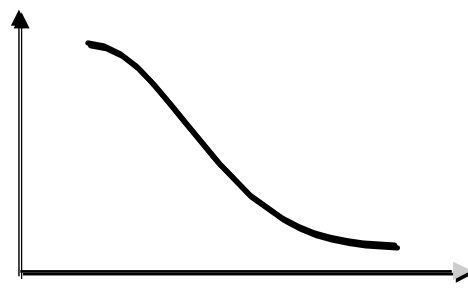
右偏分布



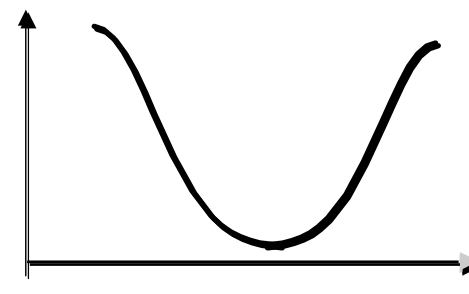
左偏分布



正J型分布



反J型分布



U型分布

图3-17 几种常见的频数分布

第四节 统计表

- 一. 统计表的构成
- 二. 统计表的设计

统计表的结构

表3-14 1997~1998年城镇居民家庭抽样调查资料				表头
项目	单位	1997年	1998年	列标题
行标题	一、调查户数	户	37890	数字资料
	二、平均每户家庭人口数	人	3.19	
	三、平均每户就业人口数	人	1.83	
	四、平均每人全部收入	元	5188.54	
	五、平均每人实际支出	元	4945.87	
	# 消费性支出	元	4185.64	
	非消费性支出	元	755.94	
	六、平均每人居住面积	平方米	11.90	12.40
资料来源：《中国统计摘要1999》，中国统计出版社，1999，第79页。 注：1. 本表为城市和县城的城镇居民家庭抽样调查材料。 2. 消费性支出项目包括：食品、衣着、家庭设备用品及服务、医疗保健、交通和通讯、娱乐教育文化服务、居住、杂项商品和服务。				附加

统计表的设计

1. 要合理安排统计表的结构
2. 总标题内容应满足3W要求
3. 数据计量单位相同时，可放在表的右上角标明，不同时应放在每个指标后或单列出一列标明
4. 表中的上下两条横线一般用粗线，其他线用细线
5. 通常情况下，统计表的左右两边不封口
6. 表中的数据一般是右对齐，有小数点时应以小数点对齐，而且小数点的位数应统一
7. 对于没有数字的表格单元，一般用“—”表示
8. 必要时可在表的下方加上注释

本章小结

1. 数据预处理的内容和目的
2. 品质数据整理与显示方法
3. 数值型数据整理与显示方法
4. 合理使用统计表
5. 用**Excel**作频数分布表和图形

结 束



第五章 概率与概率分布

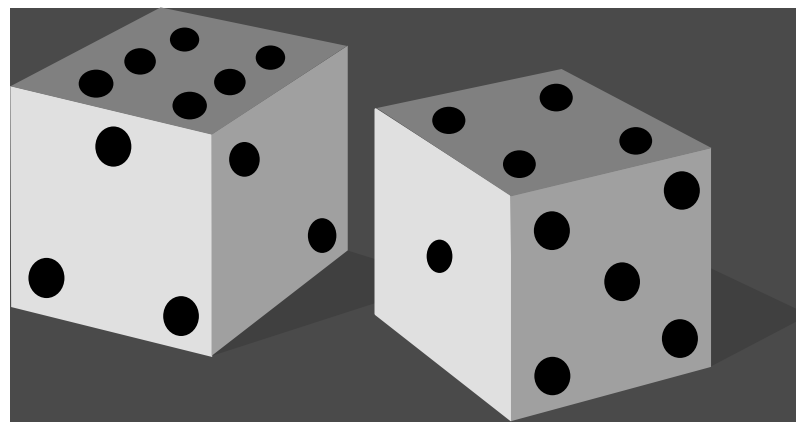
PowerPoint



第五章 概率与概率分布

第一节 概率基础

第二节 随机变量及其分布



学习目标

1. 了解随机事件的概念、事件的关系和运算
2. 理解概率的定义，掌握概率的性质和运算法则
3. 理解随机变量及其分布，计算各种分布的概率
4. 用**Excel**计算分布的概率

第一节 概率基础

- 一. 随机事件及其概率
- 二. 概率的性质与运算法则

随机事件的几个基本概念

试 验

1. 在相同条件下，对事物或现象所进行的观察
2. 例如：掷一枚骰子，观察其出现的点数
3. 试验具有以下特点
 - 可以在相同的条件下重复进行
 - 每次试验的可能结果可能不止一个，但试验的所有可能结果在试验之前是确切知道的
 - 在试验结束之前，不能确定该次试验的确切结果

事件的概念

1. 事件：随机试验的每一个可能结果(任何样本点集合)
 - 例如：掷一枚骰子出现的点数为3
2. 随机事件：每次试验可能出现也可能不出现的事件
 - 例如：掷一枚骰子可能出现的点数
3. 必然事件：每次试验一定出现的事件，用 Ω 表示
 - 例如：掷一枚骰子出现的点数小于7
4. 不可能事件：每次试验一定不出现的事件，用 Φ 表示
 - 例如：掷一枚骰子出现的点数大于6

事件与样本空间

1. 基本事件

- 一个不可能再分的随机事件
- 例如：掷一枚骰子出现的点数

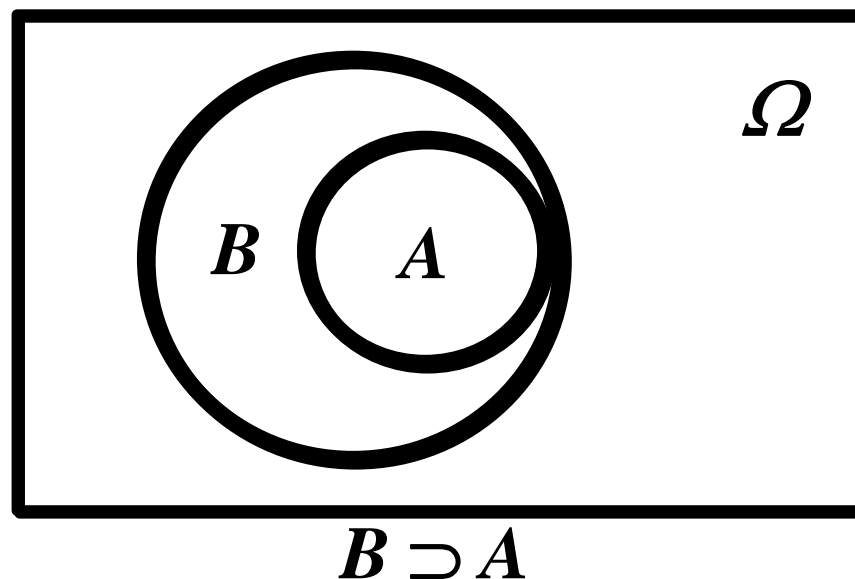
2. 样本空间

- 一个试验中所有基本事件的集合，用 Ω 表示
- 例如：在掷枚骰子的试验中， $\Omega=\{1,2,3,4,5,6\}$
- 在投掷硬币的试验中， $\Omega=\{\text{正面}, \text{反面}\}$

事件的关系和运算

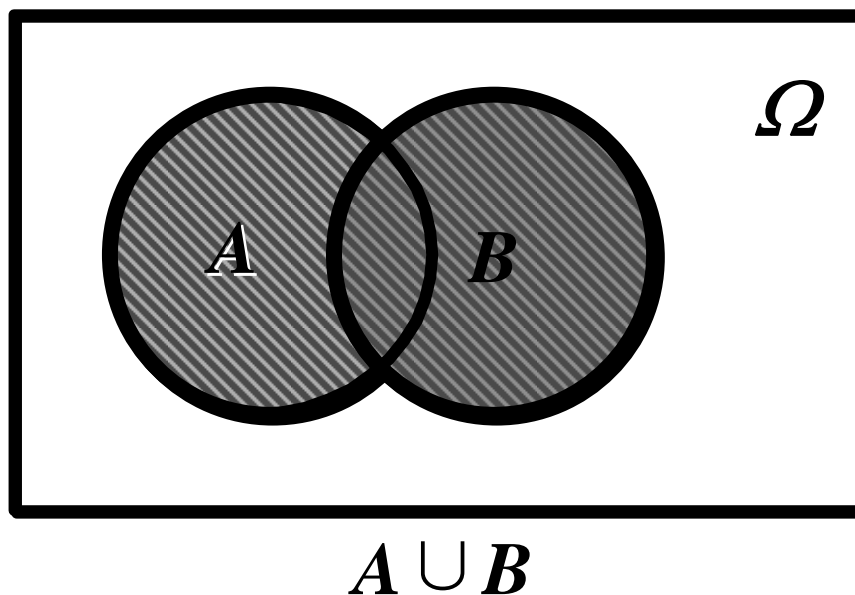
(事件的包含)

➡ 若事件 A 发生必然导致事件 B 发生, 则称事件 B 包含事件 A , 或事件 A 包含于事件 B , 记作或 $A \subset B$ 或 $B \supset A$



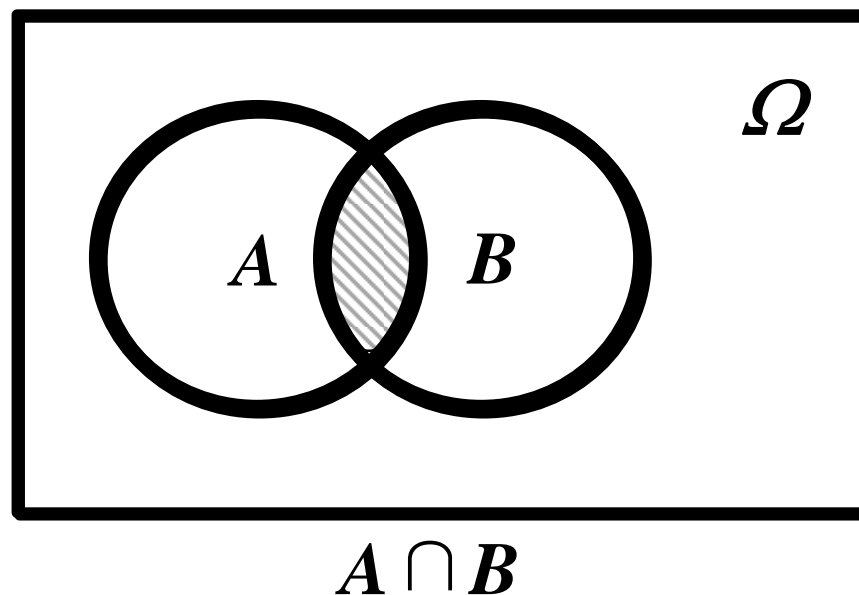
事件的关系和运算 (事件的并或和)

➡ 事件 A 和事件 B 中至少有一个发生的事件称为事件 A 与事件 B 的并。它是由属于事件 A 或事件 B 的所有的样本点组成的集合，记为 $A \cup B$ 或 $A+B$



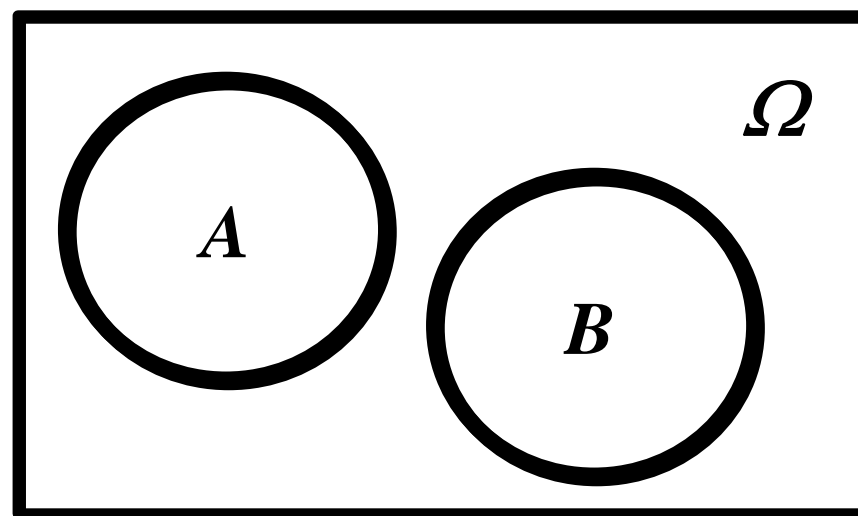
事件的关系和运算 (事件的交或积)

➡ 事件 A 与事件 B 同时发生的事件称为事件 A 与事件 B 的交，它是由属于事件 A 也属于事件 B 的所有公共样本点所组成的集合，记为 $B \cap A$ 或 AB



事件的关系和运算 (互斥事件)

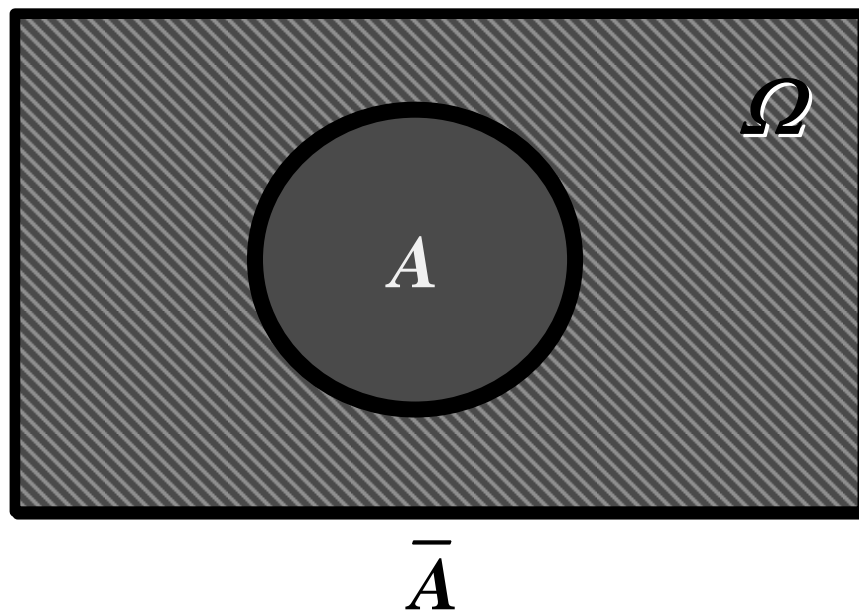
➡ 事件 A 与事件 B 中，若有一个发生，另一个必定不发生，则称事件 A 与事件 B 是互斥的，否则称两个事件是相容的。显然，事件 A 与事件 B 互斥的充分必要条件是事件 A 与事件 B 没有公共的样本点



A 与 B 互不相容

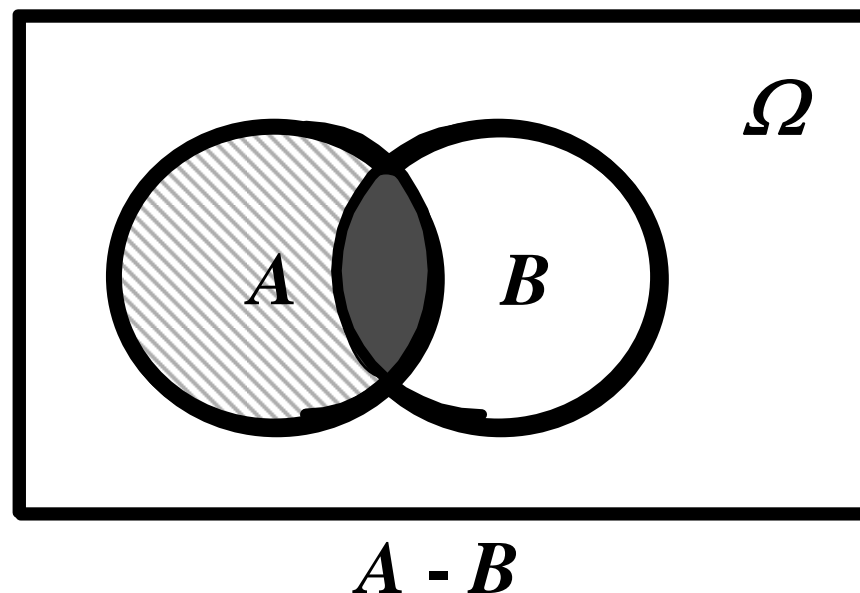
事件的关系和运算 (事件的逆)

➡ 一个事件 B 与事件 A 互斥，且它与事件 A 的并是整个样本空间 Ω ，则称事件 B 是事件 A 的逆事件。它是由样本空间中所有不属于事件 A 的样本点所组成的集合，记为 \bar{A}



事件的关系和运算 (事件的差)

➡ 事件 A 发生但事件 B 不发生的事件称为事件 A 与事件 B 的差，它是由属于事件 A 而不属于事件 B 的那些样本点构成的集合，记为 $A-B$



事件的关系和运算 (事件的性质)

➡ 设 A 、 B 、 C 为三个事件，则有

1. 交换律: $A \cup B = B \cup A$

$$A \cap B = B \cap A$$

2. 结合律: $A \cup (B \cup C) = (A \cup B) \cup C$

$$A(BC) = (AB)C$$

3. 分配律: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

事件的概率

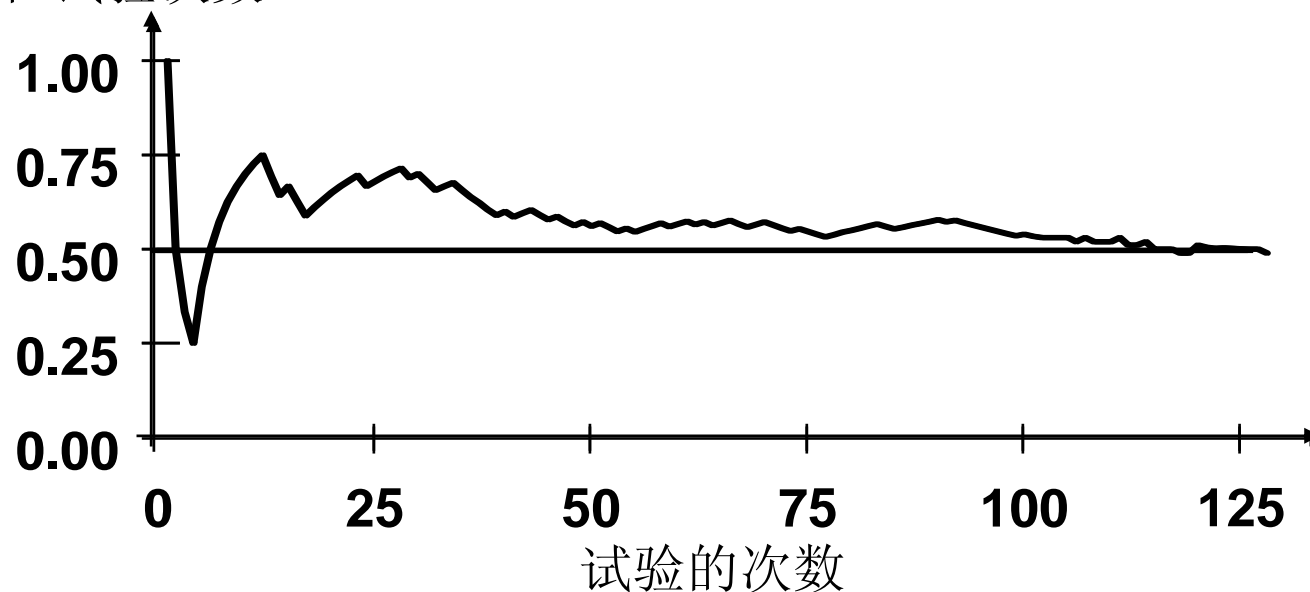
事件的概率

1. 事件A的概率是对事件A在试验中出现的可能性大小的一种度量
2. 表示事件A出现可能性大小的数值
3. 事件A的概率表示为 $P(A)$
4. 概率的定义有：古典定义、统计定义和主观概率定义

事件的概率

➡例如，投掷一枚硬币，出现正面和反面的频率，随着投掷次数 n 的增大，出现正面和反面的频率稳定在 $1/2$ 左右

正面 / 试验次数



概率的古典定义

- ➡ 如果某一随机试验的结果有限，而且各个结果在每次试验中出现的可能性相同，则事件A发生的概率为该事件所包含的基本事件个数 m 与样本空间中所包含的基本事件个数 n 的比值，记为

$$P(A) = \frac{\text{事件A所包含的基本事件个数}}{\text{样本空间所包含的基本事件个数}} = \frac{m}{n}$$

概率的古典定义 (实例)

【例】某钢铁公司所属三个工厂的职工人数如下表。从该公司中随机抽取1人，问：

(1) 该职工为男性的概率

(2) 该职工为炼钢厂职工的概率

某钢铁公司所属企业职工人数			
工厂	男职工	女职工	合计
炼钢厂	4000	1800	6200
炼铁厂	3200	1600	4800
轧钢厂	900	600	1500
合计	8500	4000	12500

概率的古典定义 (计算结果)

解：(1)用 A 表示“抽中的职工为男性”这一事件； A 为全公司男职工的集合；基本空间为全公司职工的集合。则

$$P(A) = \frac{\text{全公司男性职工人数}}{\text{全公司职工总人数}} = \frac{8500}{12500} = 0.68$$

(2) 用 B 表示“抽中的职工为炼钢厂职工”； B 为炼钢厂全体职工的集合；基本空间为全体职工的集合。则

$$P(B) = \frac{\text{炼钢厂职工人数}}{\text{全公司职工总人数}} = \frac{4800}{12500} = 0.384$$

概率的统计定义

- ➡ 在相同条件下进行 n 次随机试验，事件 A 出现 m 次，则比值 m/n 称为事件 A 发生的频率。随着 n 的增大，该频率围绕某一常数 P 上下摆动，且波动的幅度逐渐减小，取向于稳定，这个频率的稳定值即为事件 A 的概率，记为

$$P(A) = \frac{m}{n} = p$$

概率的统计定义 (实例)

【例】：某工厂为节约用电，规定每天的用电量指标为1000度。按照上个月的用电记录，30天中有12天的用电量超过规定指标，若第二个月仍没有具体的节电措施，试问该厂第一天用电量超过指标的概率。

解：上个月30天的记录可以看作是重复进行了30次试验，试验A表示用电超过指标出现了12次。根据概率的统计定义有

$$P(A) = \frac{\text{超过用电指标天数}}{\text{试验的天数}} = \frac{12}{30} = 0.4$$

主观概率定义

1. 对一些无法重复的试验，确定其结果的概率只能根据以往的经验人为确定
2. 概率是一个决策者对某事件是否发生，根据个人掌握的信息对该事件发生可能性的判断
3. 例如，我认为2001年的中国股市是一个盘整年

概率的性质与运算法则

概率的性质

1. 非负性

- 对任意事件 A , 有 $0 \leq P \leq 1$

2. 规范性

- 必然事件的概率为1; 不可能事件的概率为0。
即 $P(\Omega) = 1$; $P(\Phi) = 0$

3. 可加性

- 若 A 与 B 互斥, 则 $P(A \cup B) = P(A) + P(B)$
- 推广到多个两两互斥事件 A_1, A_2, \dots, A_n , 有
 $P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$

概率的加法法则

➡ 法则一

1. 两个互斥事件之和的概率，等于两个事件概率之和。设 A 和 B 为两个互斥事件，则

$$P(A \cup B) = P(A) + P(B)$$

2. 事件 A_1, A_2, \dots, A_n 两两互斥，则有

$$\begin{aligned} &P(A_1 \cup A_2 \cup \dots \cup A_n) \\ &= P(A_1) + P(A_2) + \dots + P(A_n) \end{aligned}$$

概率的加法法则 (实例)

【例】 根据钢铁公司职工的例子，随机抽取一名职工，计算该职工为炼钢厂或轧钢厂职工的概率

解： 用A表示“抽中的为炼钢厂职工”这一事件；B表示“抽中的为轧钢厂职工”这一事件。随机抽取一人为炼钢厂或轧钢厂职工的事件为互斥事件A与B 的和，其发生的概率为

$$P(A \cup B) = P(A) + P(B) = \frac{4800}{12500} + \frac{1500}{12500} = 0.504$$

概率的加法法则

➡ 法则二

对任意两个随机事件 A 和 B ，它们和的概率为两个事件分别概率的和减去两个事件交的概率，即

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

概率的加法法则 (实例)

【例】 设某地有甲、乙两种报纸，该地成年人中有20%读甲报纸，16%读乙报纸，8%两种报纸都读。问成年人中有百分之几至少读一种报纸。

解： 设 $A = \{\text{读甲报纸}\}$ ， $B = \{\text{读乙报纸}\}$ ， $C = \{\text{至少读一种报纸}\}$ 。则

$$\begin{aligned} P(C) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= 0.2 + 0.16 - 0.08 = 0.28 \end{aligned}$$

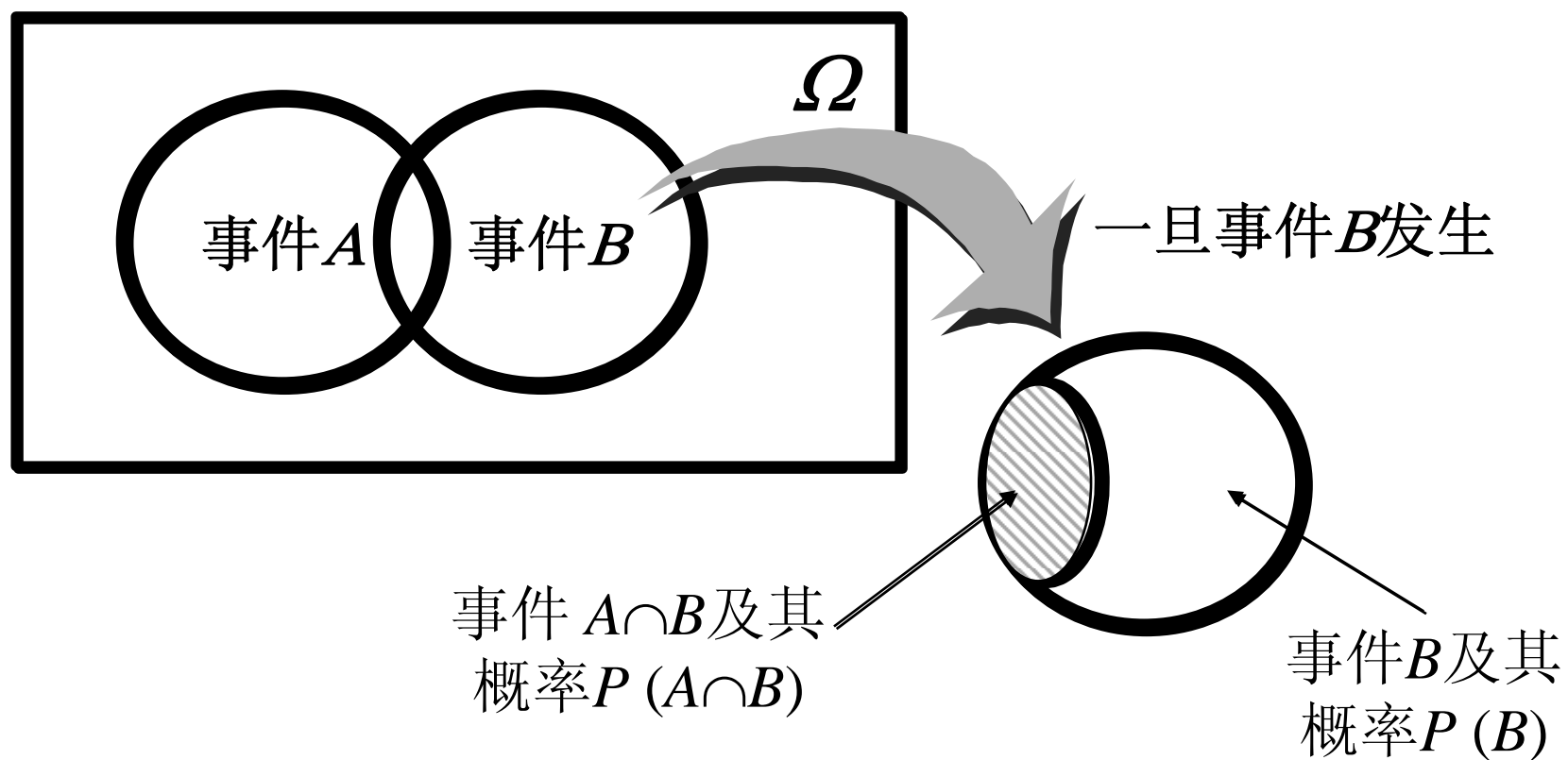
条件概率与独立事件

条件概率

- ➡ 在事件 B 已经发生的条件下，求事件 A 发生的概率，称这种概率为事件 B 发生条件下事件 A 发生的条件概率，记为

$$P(A|B) = \frac{P(AB)}{P(B)}$$

条件概率的图示



概率的乘法公式

1. 用来计算两事件交的概率
2. 以条件概率的定义为基础
3. 设 A 、 B 为两个事件，若 $P(B) > 0$ ，则
$$P(AB) = P(B)P(A|B), \text{ 或 } P(AB) = P(A)P(B|A)$$

概率的乘法公式 (实例)

【例】设有1000中产品，其中850件是正品，150件是次品，从中依次抽取2件，两件都是次品的概率是多少？

解：设 A_i 表示“第 i 次抽到的是次品”($i=1,2$)，所求概率为 $P(A_1A_2)$

$$\begin{aligned} P(A_1A_2) &= P(A_1)P(A_2 | A_1) \\ &= \frac{150}{1000} \cdot \frac{149}{999} = 0.0224 \end{aligned}$$

事件的独立性

1. 一个事件的发生与否并不影响另一个事件发生的概率，则称两个事件独立
2. 若事件 A 与 B 独立，则 $P(B|A)=P(B)$ ， $P(A|B)=P(A)$
3. 此时概率的乘法公式可简化为
4. 推广到 n 个独立事件，有

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n)$$

事件的独立性 (实例)

【例】某工人同时看管三台机床，每单位时间(如30分钟)内机床不需要看管的概率：甲机床为0.9，乙机床为0.8，丙机床为0.85。若机床是自动且独立地工作，求

(1) 在30分钟内三台机床都不需要看管的概率

(2) 在30分钟内甲、乙机床不需要看管，且丙机床需要看管的概率

解：设 A_1 , A_2 , A_3 为甲、乙、丙三台机床不需要看管的事件， \bar{A}_3 为丙机床需要看管的事件，依题意有

$$(1) P(A_1 A_2 A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) = 0.9 \times 0.8 \times 0.85 = 0.612$$

$$(2) P(A_1 A_2 \bar{A}_3) = P(A_1) \cdot P(A_2) \cdot P(\bar{A}_3)$$

$$= 0.9 \times 0.8 \times (1 - 0.85) = 0.108$$

全概公式

- ➡ 设事件 A_1, A_2, \dots, A_n 两两互斥, $A_1+A_2+\dots+A_n=\Omega$ (满足这两个条件的事件组称为一个完备事件组), 且 $P(A_i)>0(i=1,2, \dots,n)$, 则对任意事件 B , 有

$$P(B) = \sum_{i=1}^n p(A_i)P(B | A_i)$$

我们把事件 A_1, A_2, \dots, A_n 看作是引起事件 B 发生的所有可能原因, 事件 B 能且只能在原有 A_1, A_2, \dots, A_n 之一发生的条件下发生, 求事件 B 的概率就是上面的全概公式

全概公式 (实例)

【例】某车间用甲、乙、丙三台机床进行生产，各种机床的次品率分别为5%、4%、2%，它们各自的产品分别占总产量的25%、35%、40%，将它们的产品组合在一起，求任取一个是次品的概率。

解：设 A_1 表示“产品来自甲台机床”， A_2 表示“产品来自乙台机床”， A_3 表示“产品来自丙台机床”， B 表示“取到次品”。根据全概公式有

$$\begin{aligned} P(B) &= \sum_{i=1}^3 p(A_i)P(B | A_i) \\ &= 0.25 \times 0.05 + 0.35 \times 0.04 + 0.40 \times 0.02 \\ &= 0.0345 \end{aligned}$$

贝叶斯公式 (逆概公式)

1. 与全概公式解决的问题相反，贝叶斯公式是建立在条件概率的基础上寻找事件发生的原因
2. 设 n 个事件 A_1, A_2, \dots, A_n 两两互斥， $A_1 + A_2 + \dots + A_n = \Omega$ (满足这两个条件的事件组称为一个完备事件组)，且 $P(A_i) > 0 (i=1, 2, \dots, n)$ ，则

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{\sum_{j=1}^n P(A_j)P(B | A_j)}$$

贝叶斯公式 (实例)

【例】某车间用甲、乙、丙三台机床进行生产，各种机床的次品率分别为5%、4%、2%，它们各自的产品分别占总产量的25%、35%、40%，将它们的产品组合在一起，如果取到的一件产品是次品，分别求这一产品是甲、乙、丙生产的概率

解：设 A_1 表示“产品来自甲台机床”， A_2 表示“产品来自乙台机床”， A_3 表示“产品来自丙台机床”， B 表示“取到次品”。根据贝叶斯公式有：

$$P(A_1 | B) = \frac{0.25 \times 0.05}{0.0345} = 0.3623$$

$$P(A_2 | B) = \frac{0.35 \times 0.04}{0.0345} = 0.406$$

$$P(A_3 | B) = \frac{0.4 \times 0.02}{0.0345} = 0.232$$

第二节 随机变量及其分布

- 一. 随机变量的概念
- 二. 离散型随机变量的概率分布
- 三. 连续型随机变量的概率分布

随机变量的概念

随机变量的概念

1. 一次试验的结果的数值性描述
2. 一般用 X 、 Y 、 Z 来表示
3. 例如: 投掷两枚硬币出现正面的数量
4. 根据取值情况的不同分为离散型随机变量和连续型随机变量

离散型随机变量

1. 随机变量 X 取有限个值或所有取值都可以逐个列举出来 X_1, X_2, \dots
2. 以确定的概率取这些不同的值
3. 离散型随机变量的一些例子

试验	随机变量	可能的取值
抽查 100 个产品	取到次品的个数	0,1,2, ...,100
一家餐馆营业一天	顾客数	0,1,2, ...
电脑公司一个月的销售	销售量	0,1, 2,...
销售一辆汽车	顾客性别	男性为 0 ,女性为 1

连续型随机变量

1. 随机变量 X 取无限个值
2. 所有可能取值不可以逐个列举出来，而是取数轴上某一区间内的任意点
3. 连续型随机变量的一些例子

试验	随机变量	可能的取值
抽查一批电子元件	使用寿命(小时)	$X \geq 0$
新建一座住宅楼	半年后工程完成的百分比	$0 \leq X \leq 100$
测量一个产品的长度	测量误差(cm)	$X \geq 0$

离散型随机变量的概率分布

离散型随机变量的概率分布

1. 列出离散型随机变量 X 的所有可能取值
2. 列出随机变量取这些值的概率
3. 通常用下面的表格来表示

$X = x_i$	x_1, x_2, \dots, x_n
$P(X = x_i) = p_i$	p_1, p_2, \dots, p_n

4. $P(X = x_i) = p_i$ 称为离散型随机变量的概率函数
 - $p_i \geq 0$
 - $\sum_{i=1}^n p_i = 1$

离散型随机变量的概率分布 (实例)

【例】如规定打靶中域 I 得3分，中域 II 得2分，中域 III 得1分，中域外得0分。今某射手每100次射击，平均有30次中域 I，55次中域 II，10次中 III，5次中域外。则考察每次射击得分为0,1,2,3这一离散型随机变量，其概率分布为

$X = x_i$	0	1	2	3
$P(X=x_i)=p_i$	0.05	0.10	0.55	0.30

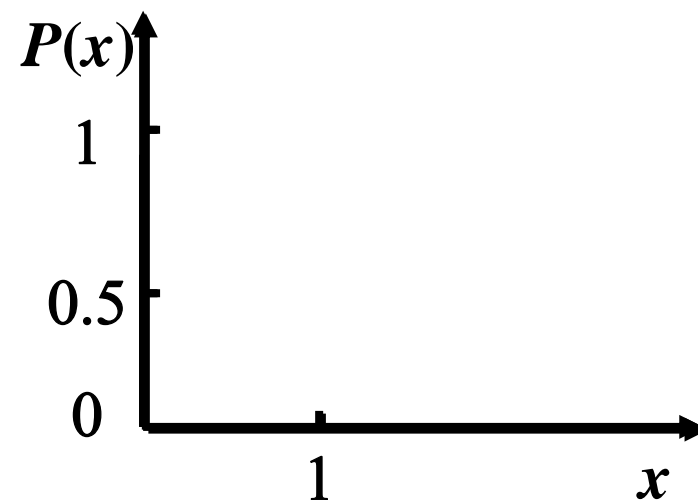
离散型随机变量的概率分布 (0-1分布)

1. 一个离散型随机变量 X 只取两个可能的值
 - 例如，男性用 1 表示，女性用 0 表示；
合格品用 1 表示，不合格品用 0 表示
2. 列出随机变量取这两个值的概率

离散型随机变量的概率分布 (0-1分布实例)

【例】已知一批产品的次品率为 $p=0.05$ ，合格率为 $q=1-p=1-0.05=0.95$ 。并指定废品用1表示，合格品用0表示。则任取一件为废品或合格品这一离散型随机变量，其概率分布为

$X = x_i$	0	1
$P(X=x_i)=p_i$	0.05	0.95



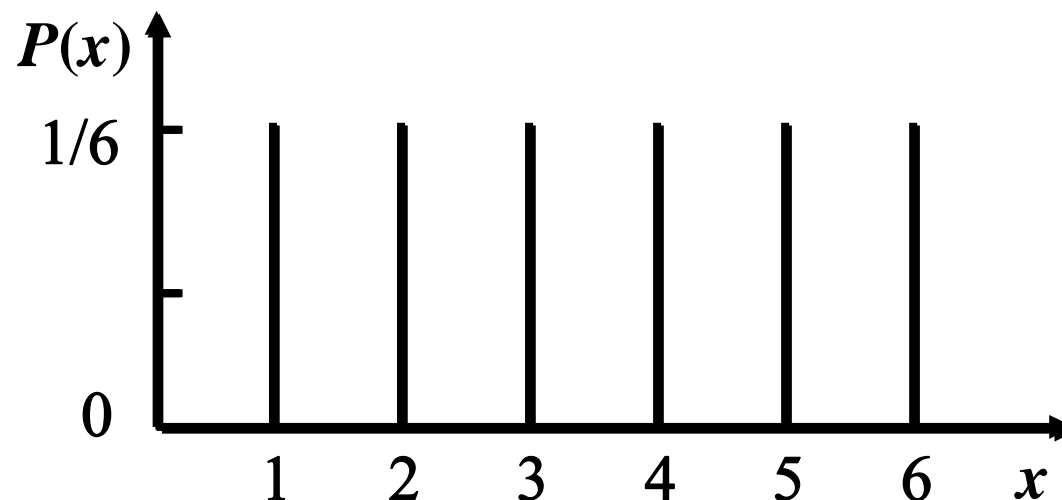
离散型随机变量的概率分布 (均匀分布)

1. 一个离散型随机变量取各个值的概率相同
2. 列出随机变量取值及其取值的概率
3. 例如，投掷一枚骰子，出现的点数及其出现各点的概率

离散型随机变量的概率分布 (均匀分布实例)

【例】投掷一枚骰子，出现的点数是个离散型随机变量，其概率分布为

$X = x_i$	1	2	3	4	5	6
$P(X=x_i)=p_i$	1/6	1/6	1/6	1/6	1/6	1/6



离散型随机变量的数学期望和方差

离散型随机变量的数学期望

1. 在离散型随机变量 X 的一切可能取值的完备组中，各可能取值 x_i 与其取相对应的概率 p_i 乘积之和
2. 描述离散型随机变量取值的集中程度
3. 计算公式为

$$E(X) = \sum_{i=1}^n x_i p_i \quad (X \text{取有限个值})$$

$$E(X) = \sum_{i=1}^{\infty} x_i p_i \quad (X \text{取无穷个值})$$

离散型随机变量的方差

1. 随机变量 X 的每一个取值与期望值的离差平方和的数学期望，记为 $D(X)$
2. 描述离散型随机变量取值的分散程度
3. 计算公式为

$$D(X) = E[X - E(X)]^2$$

若 X 是离散型随机变量，则

$$D(X) = \sum_{i=1}^{\infty} [x_i - E(X)]^2 \cdot p_i$$

离散型随机变量的方差 (实例)

【例】投掷一枚骰子，出现的点数是个离散型随机变量，其概率分布为如下。计算数学期望和方差

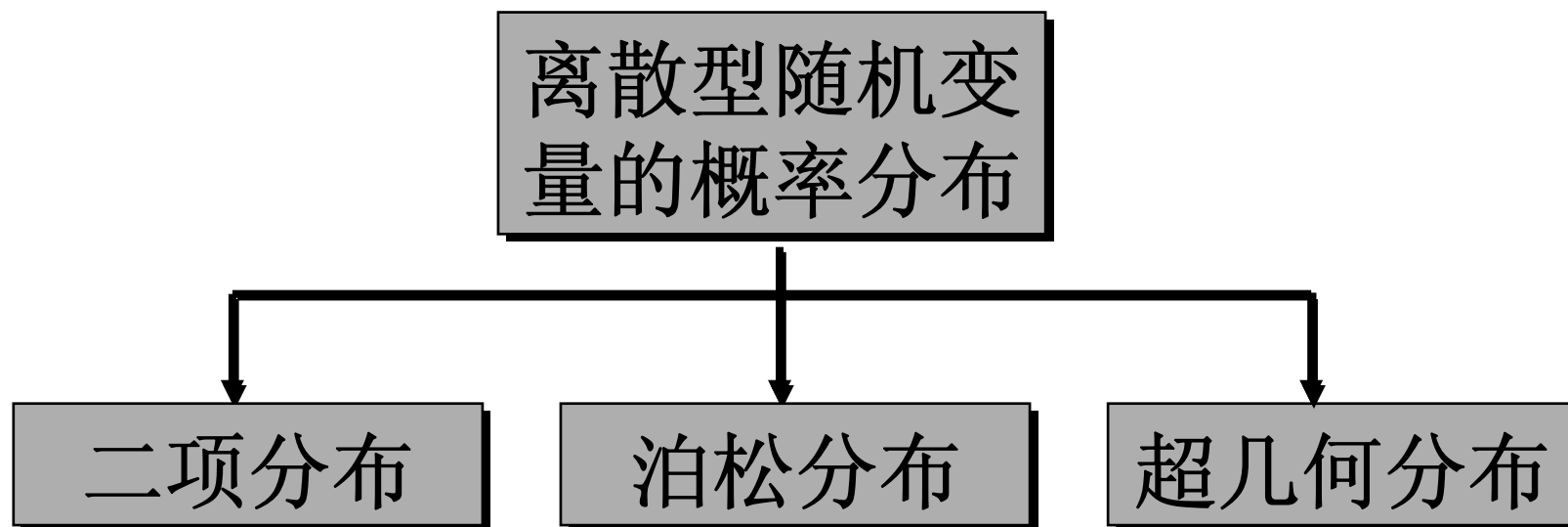
$X = x_i$	1	2	3	4	5	6
$P(X = x_i) = p_i$	1/6	1/6	1/6	1/6	1/6	1/6

解：数学期望为：
$$E(X) = \sum_{i=1}^6 x_i p_i = 1 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$$

方差为：
$$D(X) = \sum_{i=1}^6 [x_i - E(X)]^2 \cdot p_i$$
$$= (1 - 3.5)^2 \times \frac{1}{6} + \cdots + (6 - 3.5)^2 \times \frac{1}{6} = 2.9167$$

几种常见的离散型概率分布

常见的离散型概率分布



二项试验 (贝努里试验)

1. 二项分布与贝努里试验有关
2. 贝努里试验具有如下属性
 - 试验包含了 n 个相同的试验
 - 每次试验只有两个可能的结果，即“成功”和“失败”
 - 出现“成功”的概率 p 对每次试验结果是相同的；“失败”的概率 q 也相同，且 $p + q = 1$
 - 试验是相互独立的
 - 试验“成功”或“失败”可以计数

二项分布

1. 进行 n 次重复试验，出现“成功”的次数的概率分布称为二项分布
2. 设 X 为 n 次重复试验中事件 A 出现的次数， X 取 x 的概率为

$$P\{X = x\} = C_n^x p^x q^{n-x} \quad (x = 0, 1, 2, \dots, n)$$

$$\text{式中: } C_n^x = \frac{n!}{x!(n-x)!}$$

二项分布

1. 显然, 对于 $P\{X=x\} \geq 0, \quad x=1,2,\dots,n$, 有

$$\sum_{x=0}^n C_n^x p^x q^{n-x} = (p+q)^n = 1$$

2. 同样有

$$P\{0 \leq X \leq m\} = \sum_{x=0}^m C_n^x p^x q^{n-x}$$

$$P\{m \leq X \leq n\} = \sum_{x=m}^n C_n^x p^x q^{n-x}$$

3. 当 $n=1$ 时, 二项分布化简为

$$P\{X=x\} = p^x q^{1-x} = 1 \quad x=0,1$$

二项分布的数学期望和方差

1. 二项分布的数学期望为

$$E(X) = np$$

2. 方差为

$$D(X) = npq$$

二项分布 (实例)

【例】 已知100件产品中有5件次品，现从中任取一件，有放回地抽取3次。求在所抽取的3件产品中恰好有2件次品的概率

解：设 X 为所抽取的3件产品中的次品数，则 $X \sim B(3, 0.05)$ ，根据二项分布公式有

$$P\{X = 2\} = C_3^2 (0.05)^2 (0.95)^{3-2} = 0.007125$$

泊松分布

1. 用于描述在一指定时间范围内或在一定的长度、面积、体积之内每一事件出现次数的分布
2. 泊松分布的例子
 - 一个城市在一个月内发生的交通事故次数
 - 消费者协会一个星期内收到的消费者投诉次数
 - 人寿保险公司每天收到的死亡声明的人数

泊松概率分布函数

$$P\{X = x\} = \frac{\lambda e^{-\lambda}}{x!} \quad (x = 0, 1, 2, \dots, n)$$

λ — 给定的时间间隔、长度、面积、体积内“成功”的平均数

$e = 2.71828$

x — 给定的时间间隔、长度、面积、体积内“成功”的次数

泊松概率分布的期望和方差

1. 泊松分布的数学期望为

$$E(X) = \lambda$$

2. 方差为

$$D(X) = \lambda$$

泊松分布 (实例)

【例】假定某企业的职工中在周一请假的人数 X 服从泊松分布，且设周一请事假的平均人数为2.5人。求

(1) X 的均值及标准差

(2) 在给定的某周一正好请事假是5人的概率

解：(1) $E(X)=\lambda=2.5$ ； $D(X) = \lambda = \sqrt{2.5}=1.581$

$$(2) P\{X = 5\} = \frac{(2.5)^5 e^{-2.5}}{5!} = 0.067$$

泊松分布

(作为二项分布的近似)

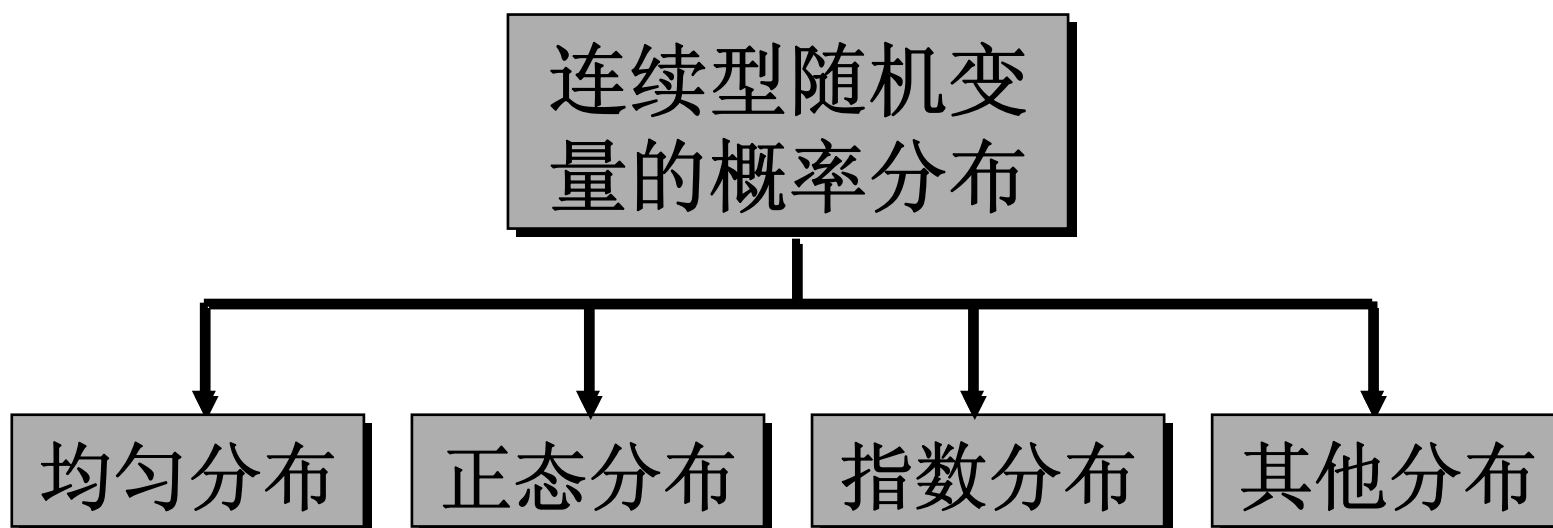
1. 当试验的次数 n 很大，成功的概率 p 很小时，可用泊松分布来近似地计算二项分布的概率，即

$$C_n^x p^x q^{n-x} \approx \frac{\lambda e^{-\lambda}}{x!}$$

2. 实际应用中，当 $P \leq 0.25$ ， $n > 20$ ， $np \leq 5$ 时，近似效果良好

连续型随机变量的概率分布

连续型随机变量的概率分布



连续型随机变量的概率分布

1. 连续型随机变量可以取某一区间或整个实数轴上的任意一个值
2. 它取任何一个特定的值的概率都等于0
3. 不能列出每一个值及其相应的概率
4. 通常研究它取某一区间值的概率
5. 用数学函数的形式和分布函数的形式来描述

概率密度函数

1. 设 X 为一连续型随机变量， x 为任意实数， X 的概率密度函数记为 $f(x)$ ，它满足条件

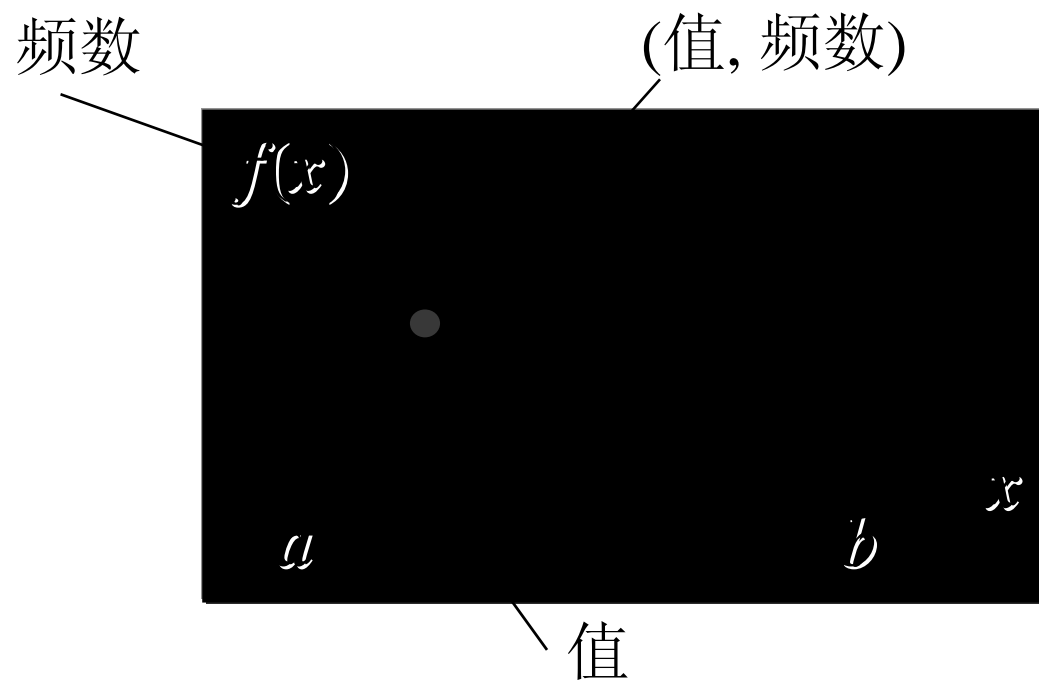
$$(1) f(x) \geq 0$$

$$(2) \int_{-\infty}^{+\infty} f(x) dx = 1$$

2. $f(x)$ 不是概率

概率密度函数

➔ 密度函数 $f(x)$ 表示 X 的所有取值 x 及其频数 $f(x)$



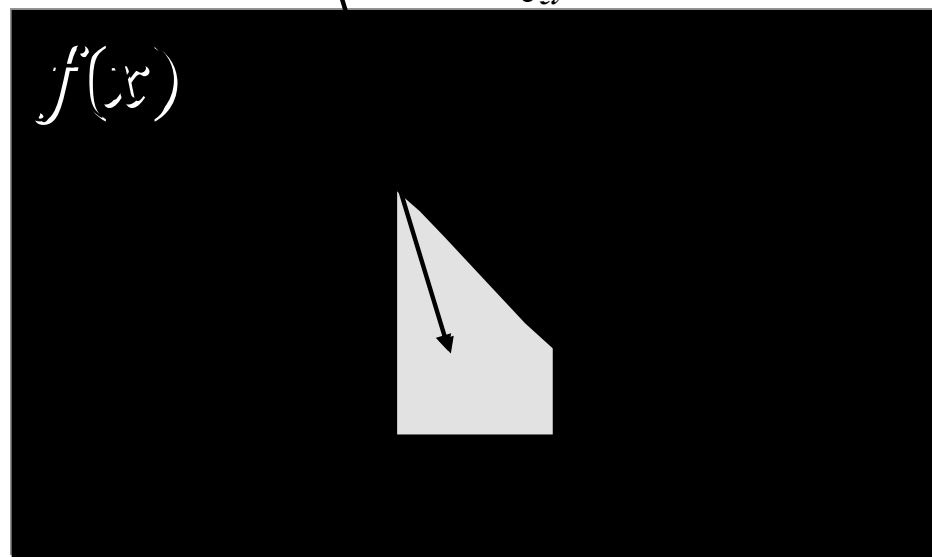
概率密度函数

- ➔ 在平面直角坐标系中画出 $f(x)$ 的图形，则对于任何实数 $x_1 < x_2$ ， $P(x_1 < X \leq x_2)$ 是该曲线下从 x_1 到 x_2 的面积

概率是曲线下的面积



$$P(a < X \leq b) = \int_a^b f(x) dx$$



分布函数

1. 连续型随机变量的概率也可以用分布函数 $F(x)$ 来表示
2. 分布函数定义为

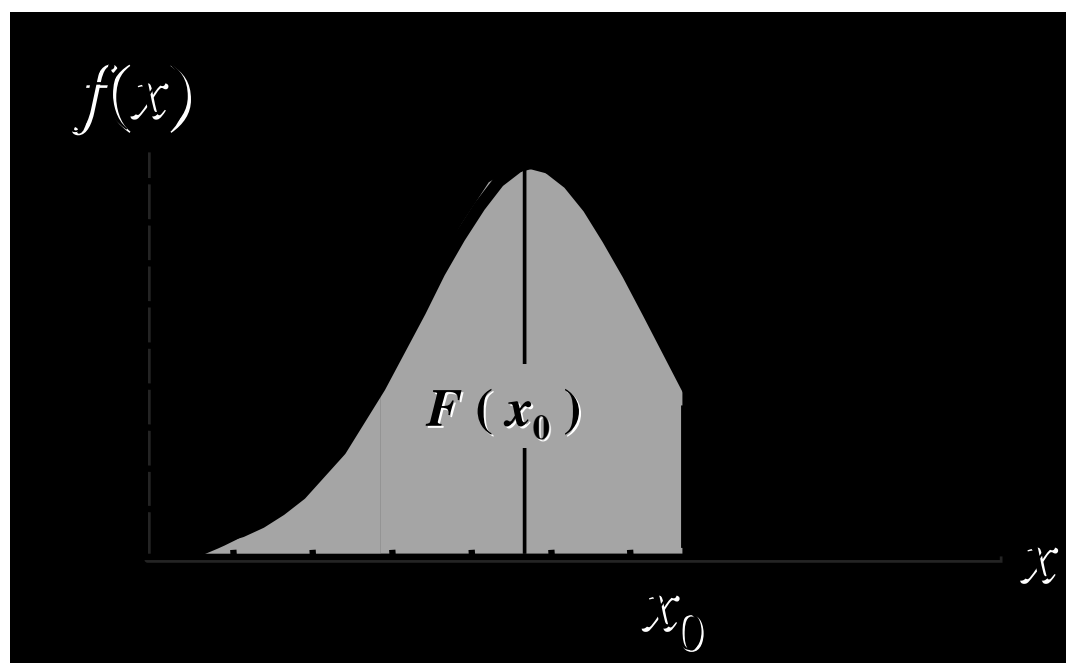
$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt \quad (-\infty < x < +\infty)$$

3. 根据分布函数, $P(a < X < b)$ 可以写为

$$P(a < X < b) = \int_a^b f(x) dx = F(b) - F(a)$$

分布函数与密度函数的图示

1. 密度函数曲线下的面积等于1
2. 分布函数是曲线下小于 x_0 的面积



连续型随机变量的期望和方差

1. 连续型随机变量的数学期望为

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \mu$$

2. 方差为

$$D(X) = \int_{-\infty}^{+\infty} [x - E(X)]^2 f(x)dx = \sigma^2$$

均匀分布

均匀分布

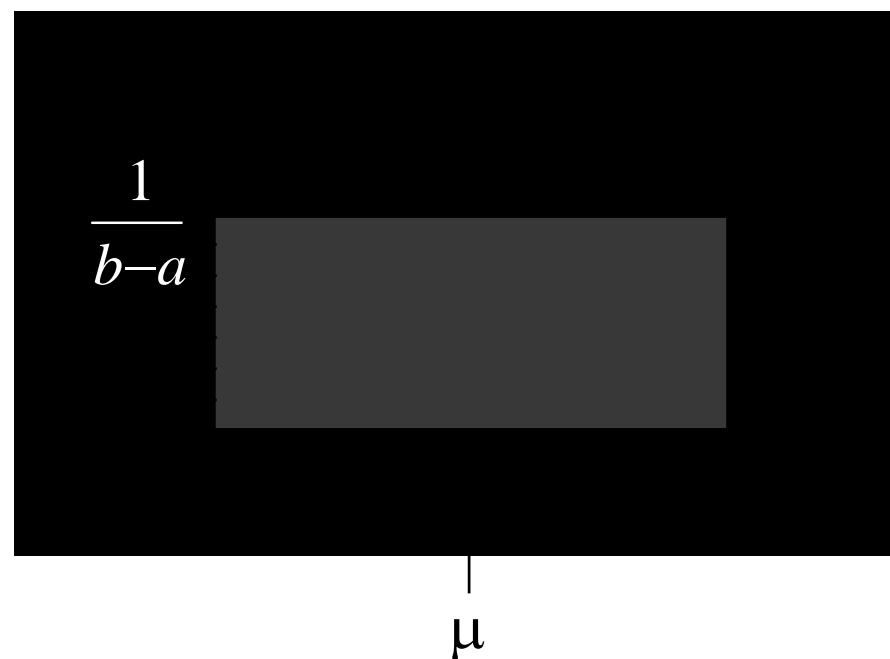
1. 若随机变量 X 的概率密度函数为

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq X \leq b \\ 0 & \text{其他} \end{cases}$$

称 X 在区间 $[a, b]$ 上均匀分布

2. 数学期望和方差分别为

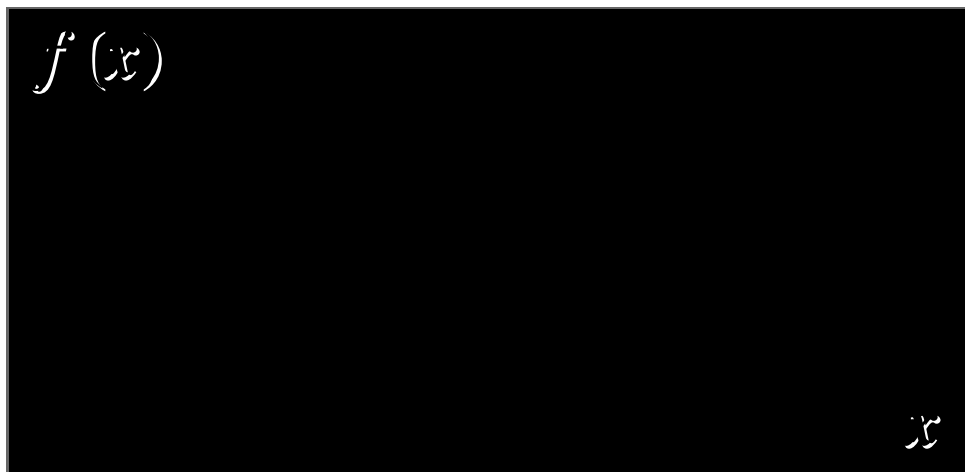
$$E(X) = \frac{a+b}{2} \quad ; \quad D(X) = \frac{(b-a)^2}{12}$$



正态分布

正态分布的重要性

1. 描述连续型随机变量的最重要的分布
2. 可用于近似离散型随机变量的分布
 - 例如: 二项分布
3. 经典统计推断的基础



概率密度函数

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty$$

$f(x)$ = 随机变量 X 的频数

σ^2 = 总体方差

$\pi = 3.14159$; $e = 2.71828$

x = 随机变量的取值 $(-\infty < x < \infty)$

μ = 总体均值

正态分布函数的性质

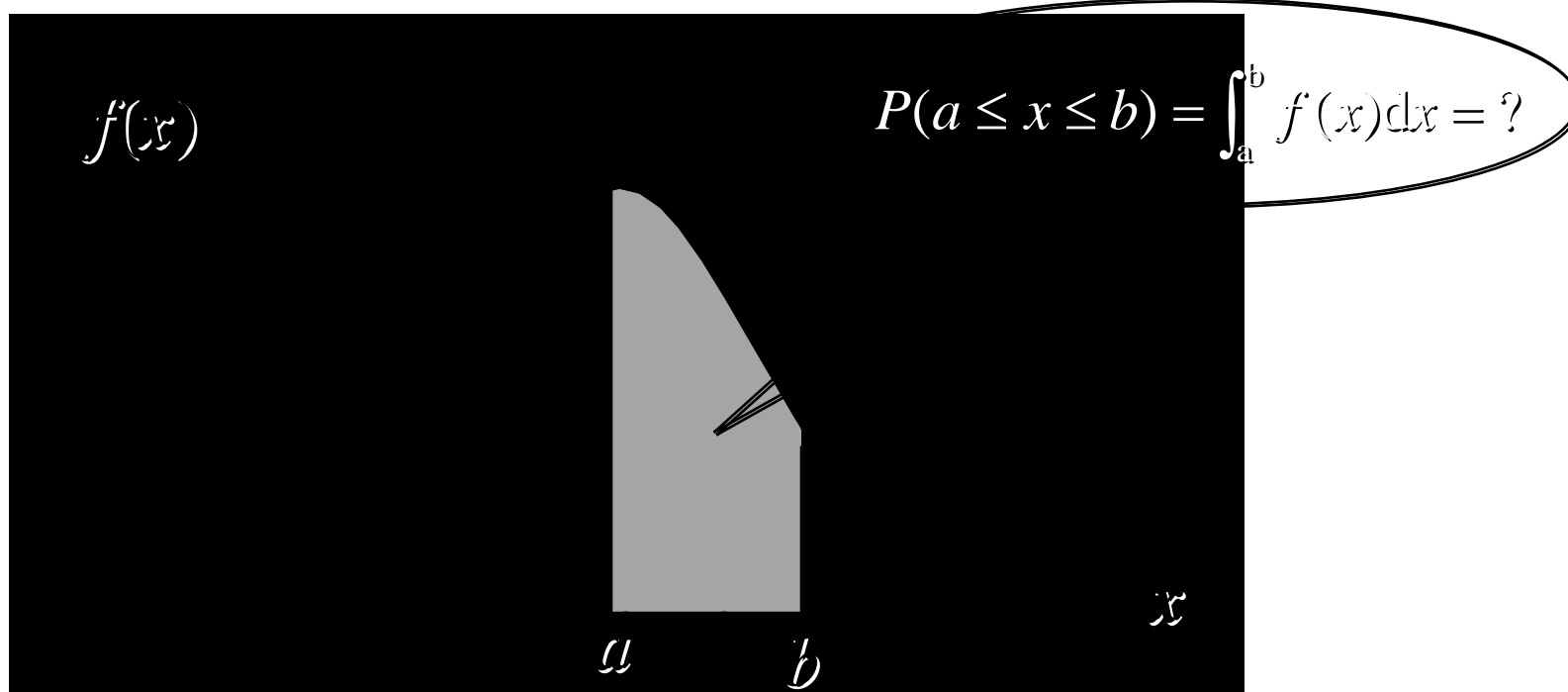
1. 概率密度函数在 x 的上方，即 $f(x)>0$
2. 正态曲线的最高点在均值 μ ，它也是分布的中位数和众数
3. 正态分布是一个分布族，每一特定正态分布通过均值 μ 的标准差 σ 来区分。 μ 决定曲线的高度， σ 决定曲线的平缓程度，即宽度
4. 曲线 $f(x)$ 相对于均值 μ 对称，尾端向两个方向无限延伸，且理论上永远不会与横轴相交
5. 正态曲线下的总面积等于1
6. 随机变量的概率由曲线下的面积给出

μ 和 σ 对正态曲线的影响

$f(x)$

正态分布的概率

概率是曲线下的面积!



标准正态分布的重要性

1. 一般的正态分布取决于均值 μ 和标准差 σ
2. 计算概率时，每一个正态分布都需要有自己的正态概率分布表，这种表格是无穷多的
3. 若能将一般的正态分布转化为标准正态分布，计算概率时只需要查一张表

标准正态分布函数

1. 任何一个一般的正态分布，可通过下面的线性变换转化为标准正态分布

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

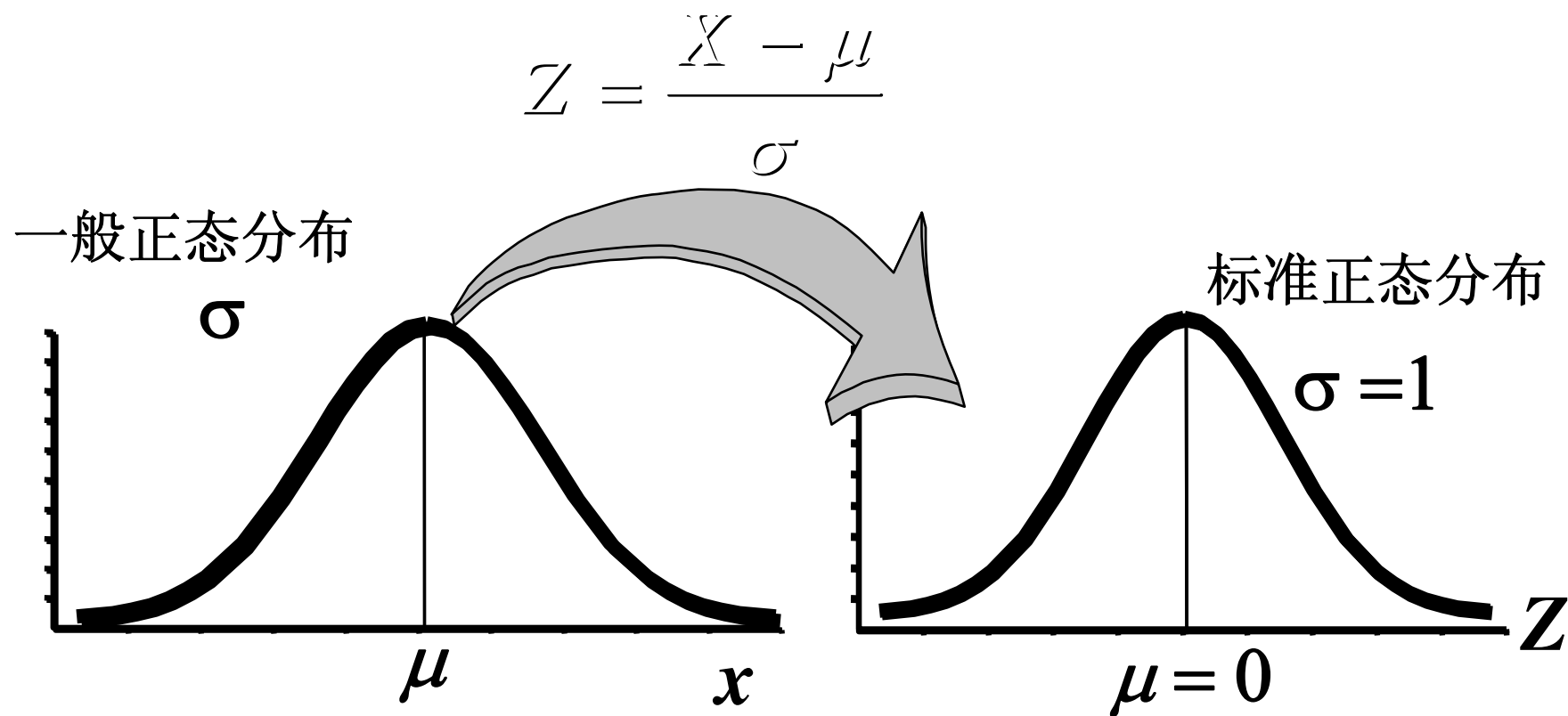
2. 标准正态分布的概率密度函数

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty$$

3. 标准正态分布的分布函数

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

标准正态分布



标准正态分布表的使用

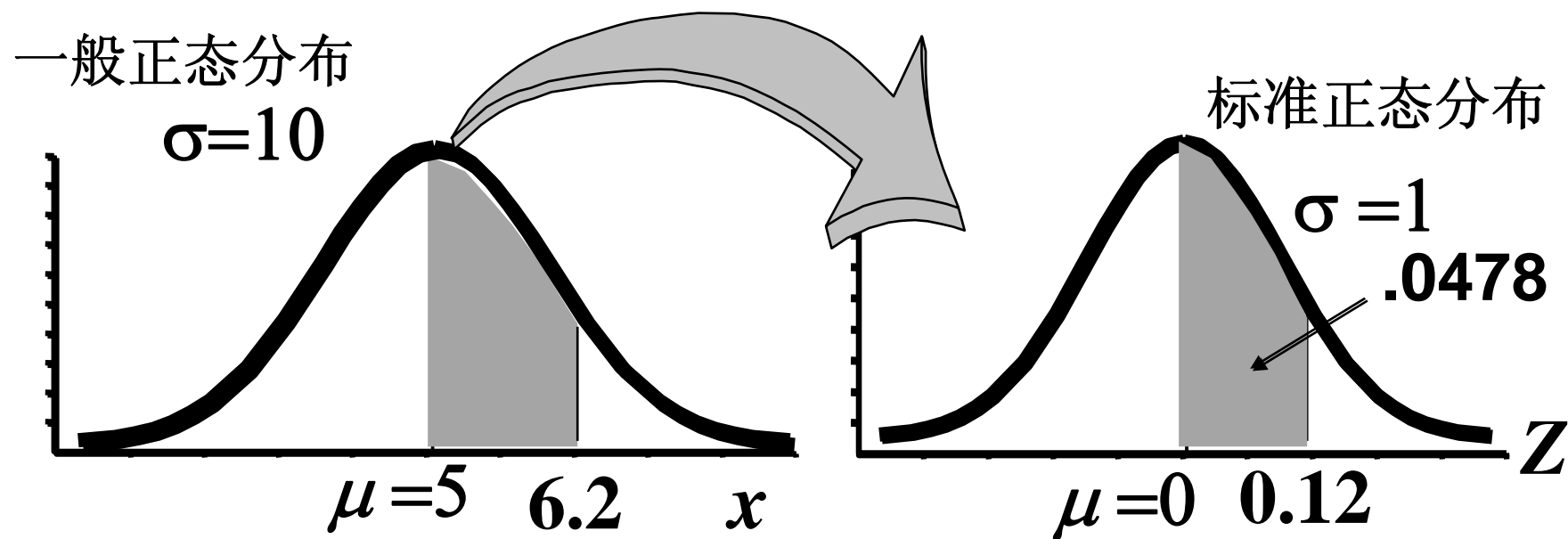
1. 将一个一般的转换为标准正态分布
2. 计算概率时，查标准正态概率分布表
3. 对于负的 x ，可由 $\Phi(-x)=1-\Phi(x)$ 得到
4. 对于标准正态分布，即 $X \sim N(0,1)$ ，有
 - $P(a \leq X \leq b) = \Phi(b) - \Phi(a)$
 - $P(|X| \leq a) = 2\Phi(a) - 1$
5. 对于一般正态分布，即 $X \sim N(\mu, \sigma)$ ，有

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)$$

标准化的例子

$P(5 \leq X \leq 6.2)$

$$Z = \frac{X - \mu}{\sigma} = \frac{6.2 - 5}{10} = 0.12$$



标准化的例子

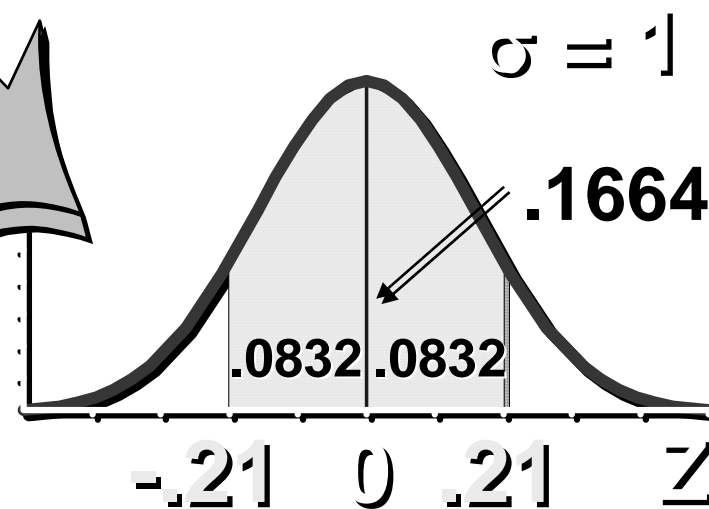
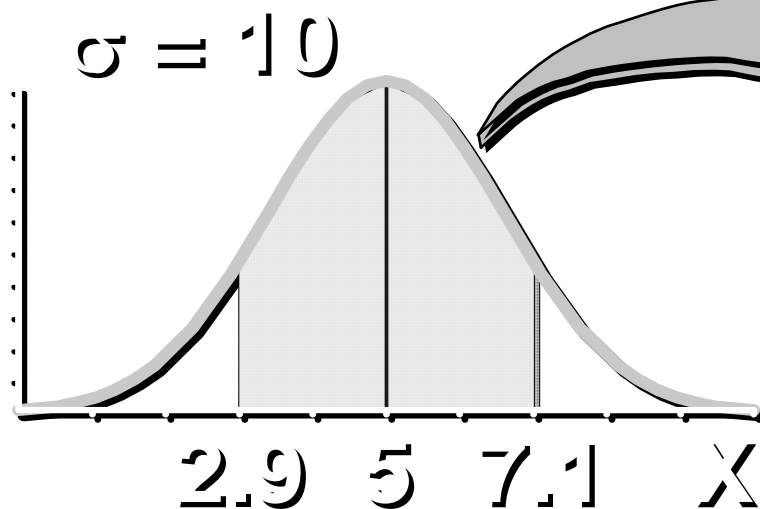
$P(2.9 \leq X \leq 7.1)$

$$Z = \frac{X - \mu}{\sigma} = \frac{2.9 - 5}{10} = -.21$$

一般正态分布

$$Z = \frac{X - \mu}{\sigma} = \frac{7.1 - 5}{10} = .21$$

标准正态分布



正态分布 (实例)

【例】设 $X \sim N(0, 1)$ ，求以下概率：

(1) $P(X < 1.5)$; (2) $P(X > 2)$; (3) $P(-1 < X \leq 3)$; (4) $P(|X| \leq 2)$

解：(1) $P(X < 1.5) = \Phi(1.5) = 0.9332$

(2) $P(X > 2) = 1 - P(2 \leq X) = 1 - 0.9973 = 0.0027$

(3) $P(-1 < X \leq 3) = P(X \leq 3) - P(X \leq -1)$
 $= \Phi(3) - \Phi(-1) = \Phi(3) - [1 - \Phi(1)]$
 $= 0.9987 - (1 - 0.8413) = 0.8390$

(4) $P(|X| \leq 2) = P(-2 \leq X \leq 2) = \Phi(2) - \Phi(-2)$
 $= \Phi(2) - [1 - \Phi(2)] = 2\Phi(2) - 1 = 0.9545$

正态分布 (实例)

【例】 设 $X \sim N(5, 3^2)$, 求以下概率

(1) $P(X \leq 10)$; (2) $P(2 < X < 10)$

解: (1)
$$P(X \leq 10) = P\left(\frac{X - 5}{3} \leq \frac{10 - 5}{3}\right)$$
$$= P\left(\frac{X - 5}{3} \leq 1.67\right) = \Phi(1.67) = 0.9525$$

(2)
$$P(2 < X < 10) = P\left(\frac{2 - 5}{3} < \frac{X - 5}{3} < \frac{10 - 5}{3}\right)$$
$$= P\left(-1 < \frac{X - 5}{3} < 1.67\right)$$
$$= \Phi(1.67) - \Phi(-1) = 0.7938$$

二项分布的正态近似

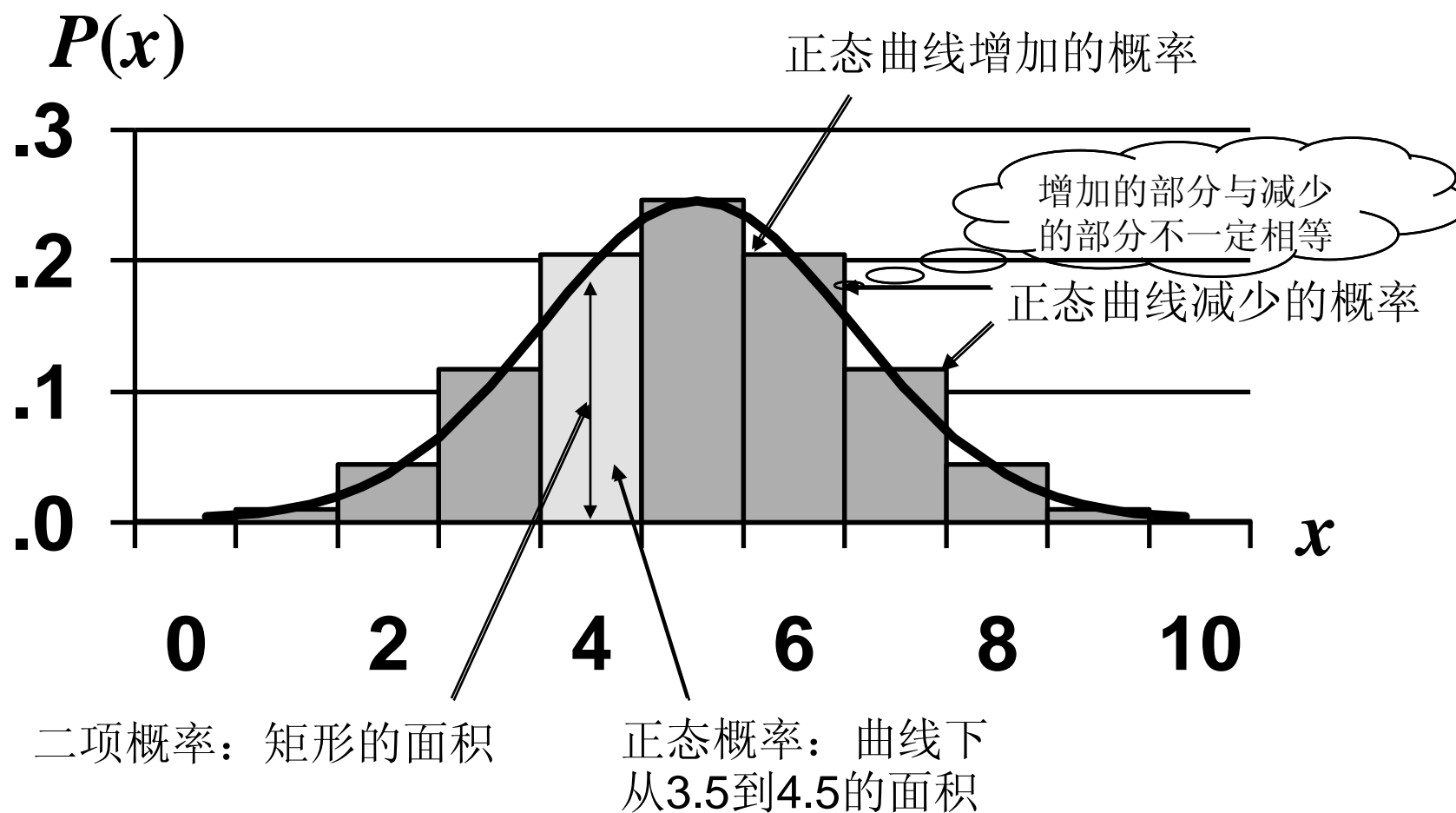
二项分布的正态近似

1. 当 n 很大时, 二项随机变量 X 近似服从正态分布 $N\{np, np(1-p)\}$
2. 对于一个二项随机变量 X , 当 n 很大时, 求 $P(x_1 \leq X \leq x_2)$ 时可用正态分布近似为

$$\begin{aligned} P\{x_1 \leq X \leq x_2\} &= \sum_{x=x_1}^{x_2} C_n^x p^x q^{n-x} = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \\ &= \Phi(b) - \Phi(a) \end{aligned}$$

$$\text{式中: } a = \frac{x_1 - np}{\sqrt{npq}}, \quad b = \frac{x_2 - np}{\sqrt{npq}}, \quad q = 1 - p$$

为什么概率是近似的



二项分布的正态近似 (实例)

【例】100台机床彼此独立地工作，每台机床的实际工作时间占全部工作时间的8%。求

(1)任一时刻有70~80台机床在工作的概率

(2)任一时刻有80台以上机床在工作的概率

解：设 X 表示100机床中工作着的机床数，则 $X \sim B(100, 0.8)$ 。
现用正态分布近似计算， $np=80$ ， $npq=16$

$$\begin{aligned}(1) \quad P(70 \leq X \leq 80) &= P\left(\frac{70 - 80}{4} \leq \frac{X - 80}{4} \leq \frac{80 - 80}{4}\right) \\ &= \Phi(1.5) - \Phi(0) = 0.927\end{aligned}$$

$$(2) \quad P(X \geq 80) = P\left(\frac{X - 80}{4} \geq 0\right) = 1 - \Phi(0) = 0.5$$

本章小结

1. 定义试验、结果、事件、样本空间、概率
2. 描述和使用概率的运算法则
3. 定义和解释随机变量及其分布
4. 计算随机变量的数学期望和方差
5. 计算离散型随机变量的概率和概率分布
6. 计算连续型随机变量的概率
7. 用正态分布近似二项分布
8. 用Excel计算分布的概率

结 束



第七章 假设检验

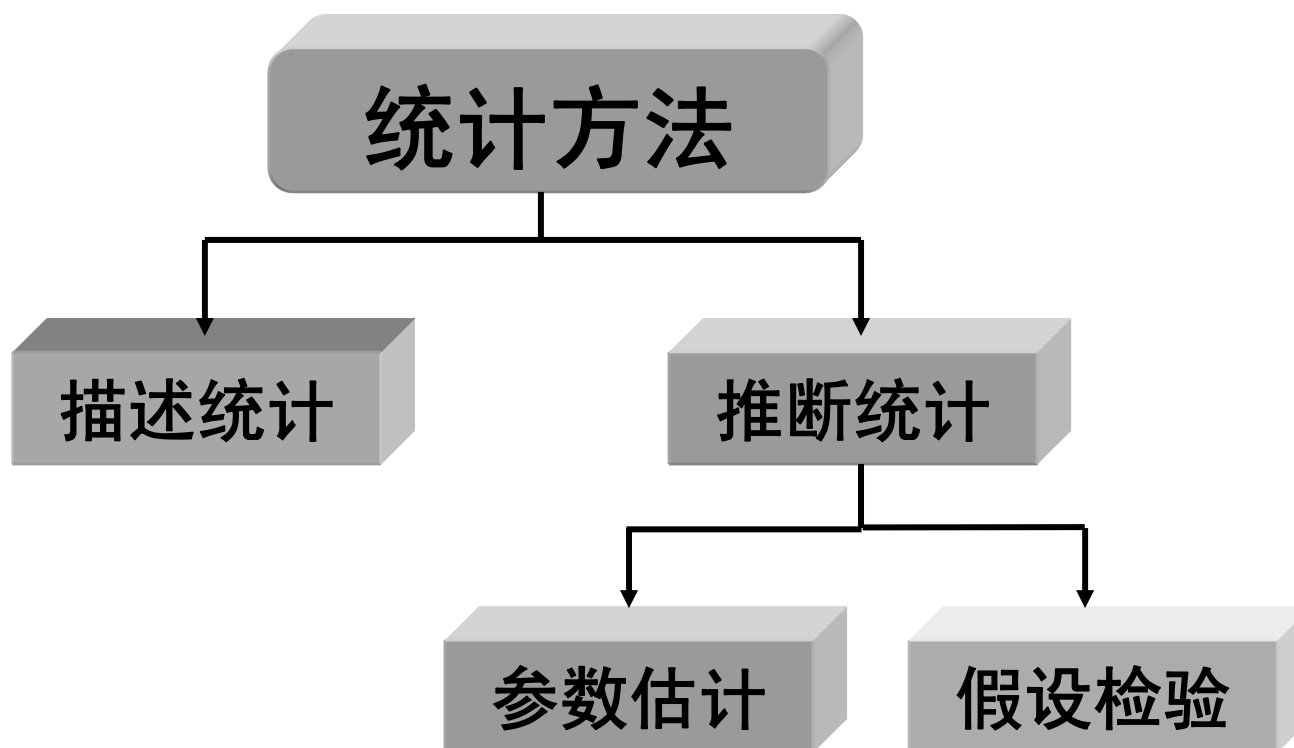
PowerPoint



第七章 假设检验

- 第一节 假设检验的一般问题
- 第二节 一个正态总体的参数检验
- 第三节 两个正态总体的参数检验
- 第四节 假设检验中的其他问题

假设检验在统计方法中的地位



学习目标

1. 了解假设检验的基本思想
2. 掌握假设检验的步骤
3. 能对实际问题作假设检验
4. 利用置信区间进行假设检验
5. 利用 P -值进行假设检验

第一节 假设检验的一般问题

- 一. 假设检验的概念
- 二. 假设检验的步骤
- 三. 假设检验中的小概率原理
- 四. 假设检验中的两类错误
- 五. 双侧检验和单侧检验

经济、管理类
基础课程

统计学

假设检验的概念与思想

什么是假设?

- ➔ 对总体参数的一种看法
 - 总体参数包括总体均值、比例、方差等
 - 分析之前必需陈述

我认为该企业生产的零件的平均长度为4厘米!



什么是假设检验？

1. 概念

- 事先对总体参数或分布形式作出某种假设
- 然后利用样本信息来判断原假设是否成立

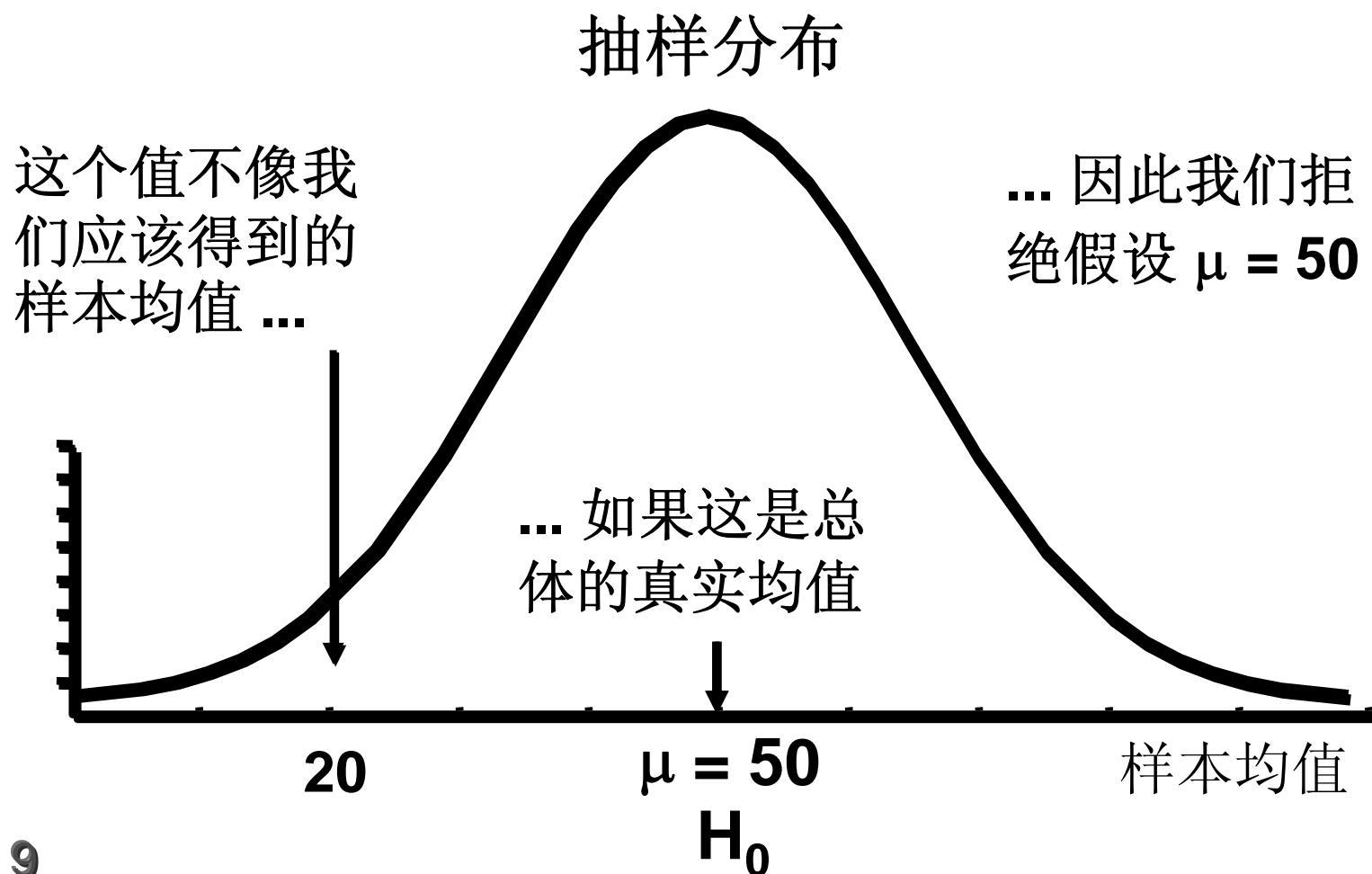
2. 类型

- 参数假设检验
- 非参数假设检验

3. 特点

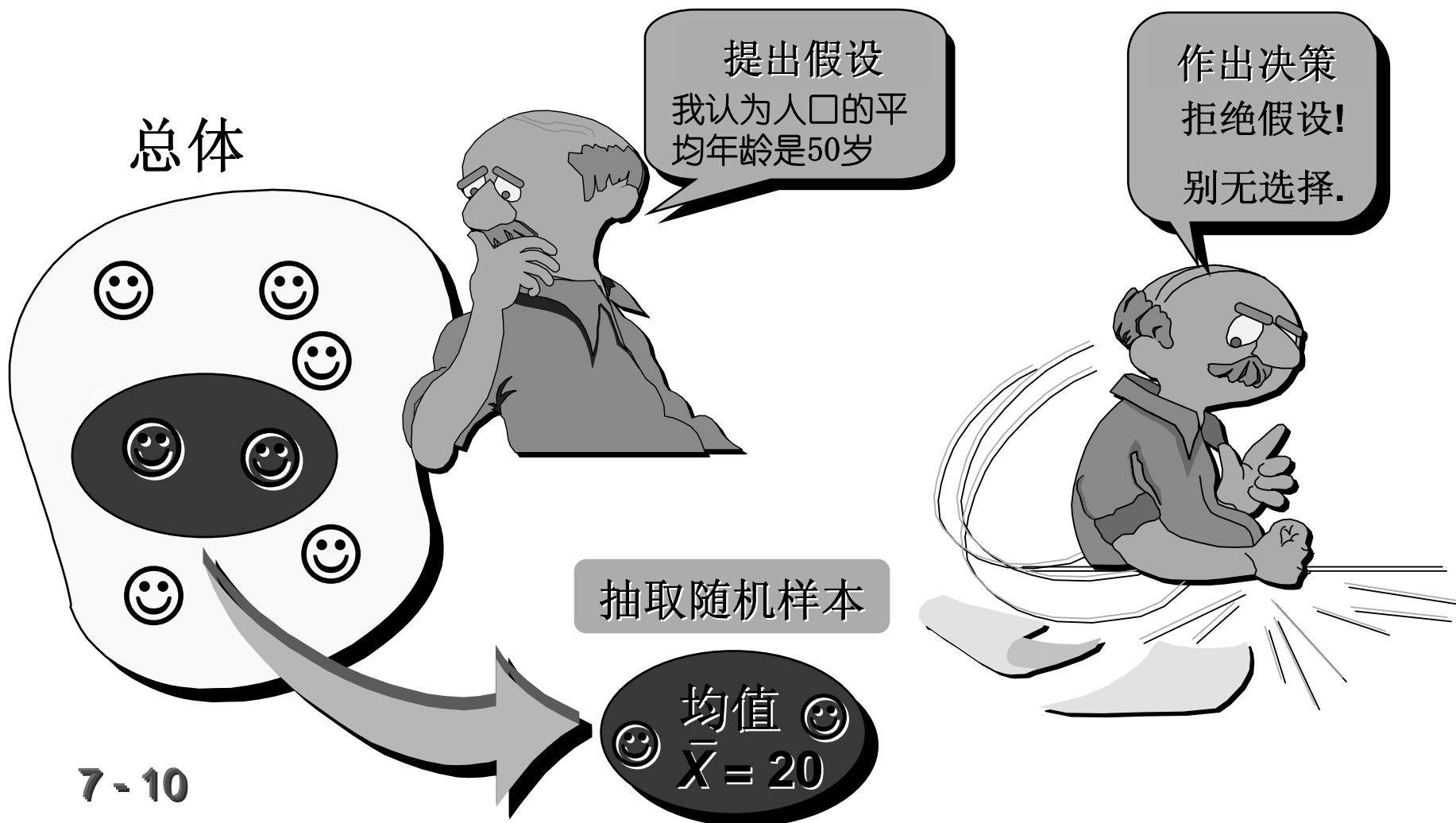
- 采用逻辑上的反证法
- 依据统计上的小概率原理

假设检验的基本思想



假设检验的过程

(提出假设→抽取样本→作出决策)



假设检验的步骤

- 提出原假设和备择假设
- 确定适当的检验统计量
- 规定显著性水平 α
- 计算检验统计量的值
- 作出统计决策

提出原假设和备择假设

➡ 什么是原假设? (Null Hypothesis)

为什么
叫0假设

1. 待检验的假设, 又称“0假设”
2. 如果错误地作出决策会导致一系列后果
3. 总是有等号 $=$, \leq 或 \geq
4. 表示为 H_0
 - $H_0: \mu =$ 某一数值
 - 指定为 $=$ 号, 即 \leq 或 \geq
 - 例如, $H_0: \mu = 3190$ (克)

提出原假设和备择假设

➡ 什么是备择假设? (**Alternative Hypothesis**)

1. 与原假设对立的假设
2. 总是有不等号: \neq , $<$ 或 $>$
3. 表示为 H_1
 - $H_1: \mu < \text{某一数值}$, 或 $\mu > \text{某一数值}$
 - 例如, $H_1: \mu < 3910(\text{克})$, 或 $\mu > 3910(\text{克})$

确定适当的检验统计量

➡ 什么检验统计量？

1. 用于假设检验问题的统计量
2. 选择统计量的方法与参数估计相同，需考虑
 - 是大样本还是小样本
 - 总体方差已知还是未知
3. 检验统计量的基本形式为

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

规定显著性水平 α

➡ 什么显著性水平？

1. 是一个概率值
2. 原假设为真时，拒绝原假设的概率
 - 被称为抽样分布的拒绝域
3. 表示为 α (alpha)
 - 常用的 α 值有0.01, 0.05, 0.10
4. 由研究者事先确定

作出统计决策

1. 计算检验的统计量
2. 根据给定的显著性水平 α ，查表得出相应的临界值 Z_α 或 $Z_{\alpha/2}$
3. 将检验统计量的值与 α 水平的临界值进行比较
4. 得出接受或拒绝原假设的结论

经济、管理类
基础课程

统计学

假设检验中的小概率原理

假设检验中的小概率原理

➡ 什么小概率？

1. 在一次试验中，一个几乎不可能发生的事件发生的概率
2. 在一次试验中小概率事件一旦发生，我们就有理由拒绝原假设
3. 小概率由研究者事先确定

什么是小概率

假设检验中的两类错误 (决策风险)

假设检验中的两类错误

1. 第一类错误（弃真错误）

- 原假设为真时拒绝原假设
- 会产生一系列后果
- 第一类错误的概率为 α
 - 被称为显著性水平

2. 第二类错误（取伪错误）

- 原假设为假时接受原假设
- 第二类错误的概率为 β (Beta)

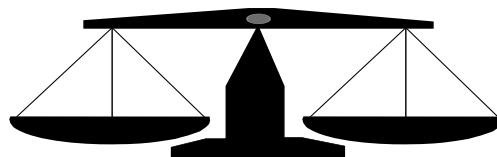
假设检验中的两类错误 (决策结果)

H_0 : 无罪

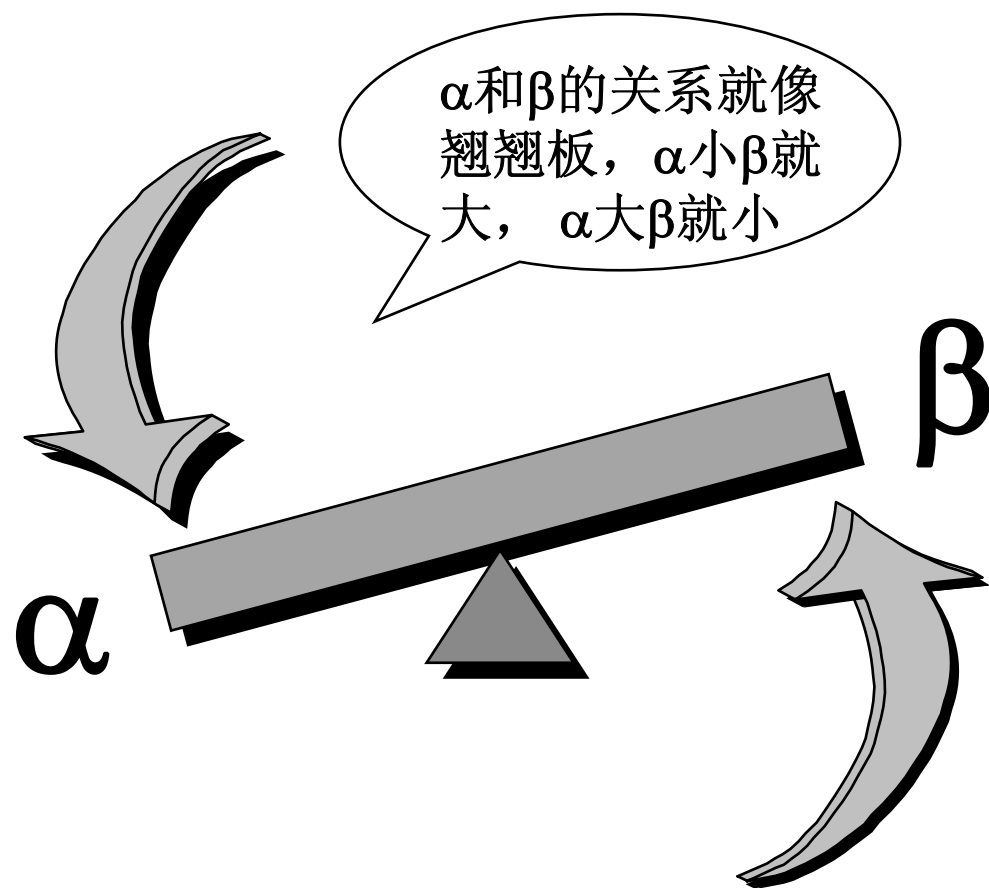
假设检验就好像一场审判过程

统计检验过程

陪审团审判			H_0 检验		
裁决	实际情况		决策	实际情况	
	无罪	有罪		H_0 为真	H_0 为假
无罪	正确	错误	接受 H_0	$1 - \alpha$	第二类错误(β)
有罪	错误	正确	拒绝 H_0	第一类错误(α)	功效($1 - \beta$)



α 错误和 β 错误的关系



影响 β 错误的因素

1. 总体参数的真值
 - 随着假设的总体参数的减少而增大
2. 显著性水平 α
 - 当 α 减少时增大
3. 总体标准差 σ
 - 当 σ 增大时增大
4. 样本容量 n
 - 当 n 减少时增大

双侧检验和单侧检验

双侧检验与单侧检验 (假设的形式)

假设	研究的问题		
	双侧检验	左侧检验	右侧检验
H_0	$\mu = \mu_0$	$\mu \geq \mu_0$	$\mu \leq \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$

双侧检验

（原假设与备择假设的确定）

1. 双侧检验属于 *决策中的假设检验*。也就是说，不论是拒绝 H_0 还是接受 H_0 ，我们都必需采取相应的行动措施
2. 例如，某种零件的尺寸，要求其平均长度为10厘米，大于或小于10厘米均属于不合格
3. 建立的原假设与备择假设应为

$$H_0: \mu = 10 \quad H_1: \mu \neq 10$$

双侧检验 (确定假设的步骤)

1. 例如问题为：检验该企业生产的零件平均长度为4厘米
2. 步骤
 - 从统计角度陈述问题 ($\mu = 4$)
 - 从统计角度提出相反的问题 ($\mu \neq 4$)
 - 必需互斥和穷尽
 - 提出原假设 ($\mu = 4$)
 - 提出备择假设 ($\mu \neq 4$)
 - 有 \neq 符号

双侧检验 (例子)

该企业生产的零件平均长度是**4**厘米吗？

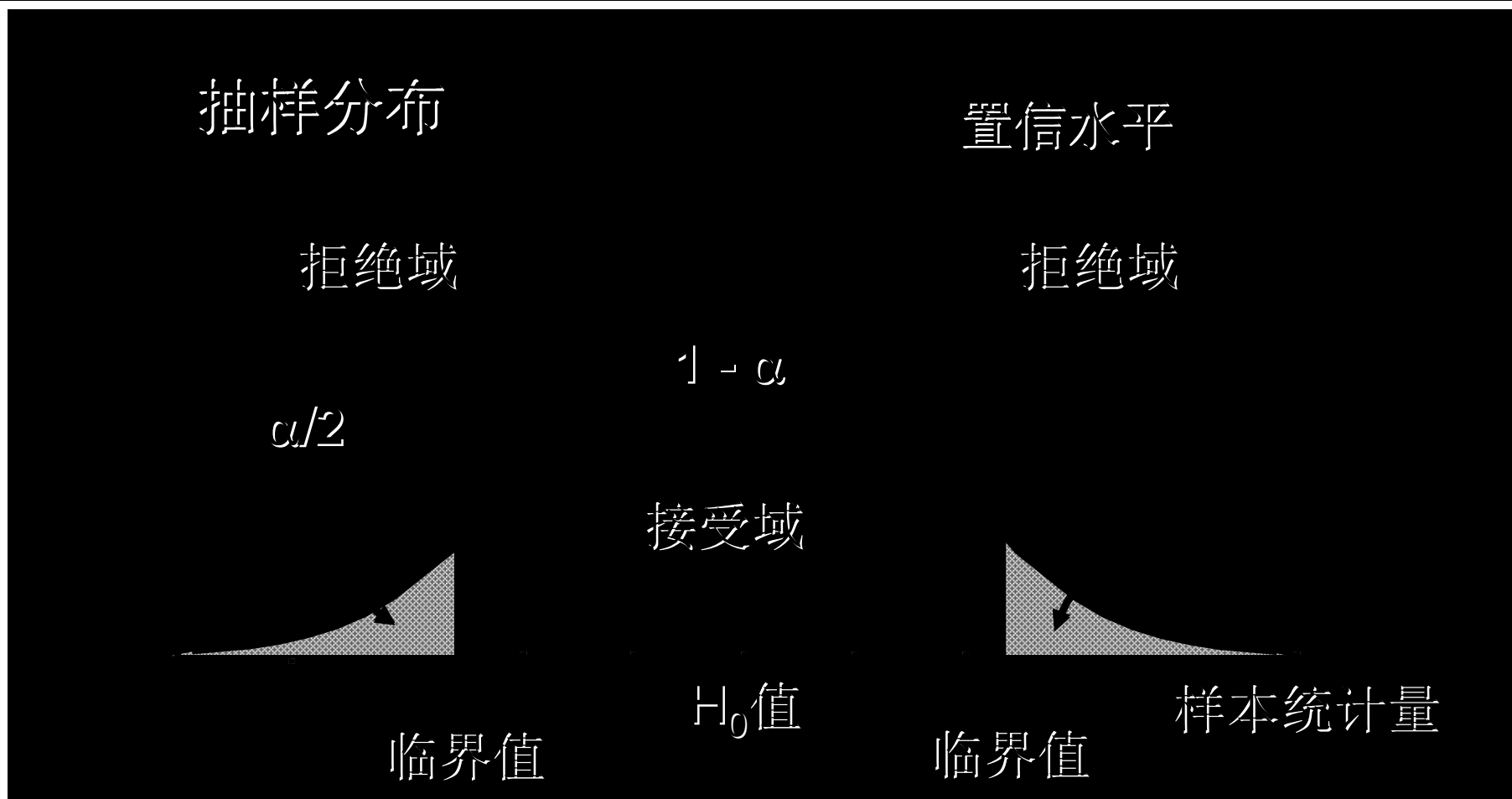
(属于决策中的假设)

提出原假设: $H_0: \mu = 4$

提出备择假设: $H_1: \mu \neq 4$

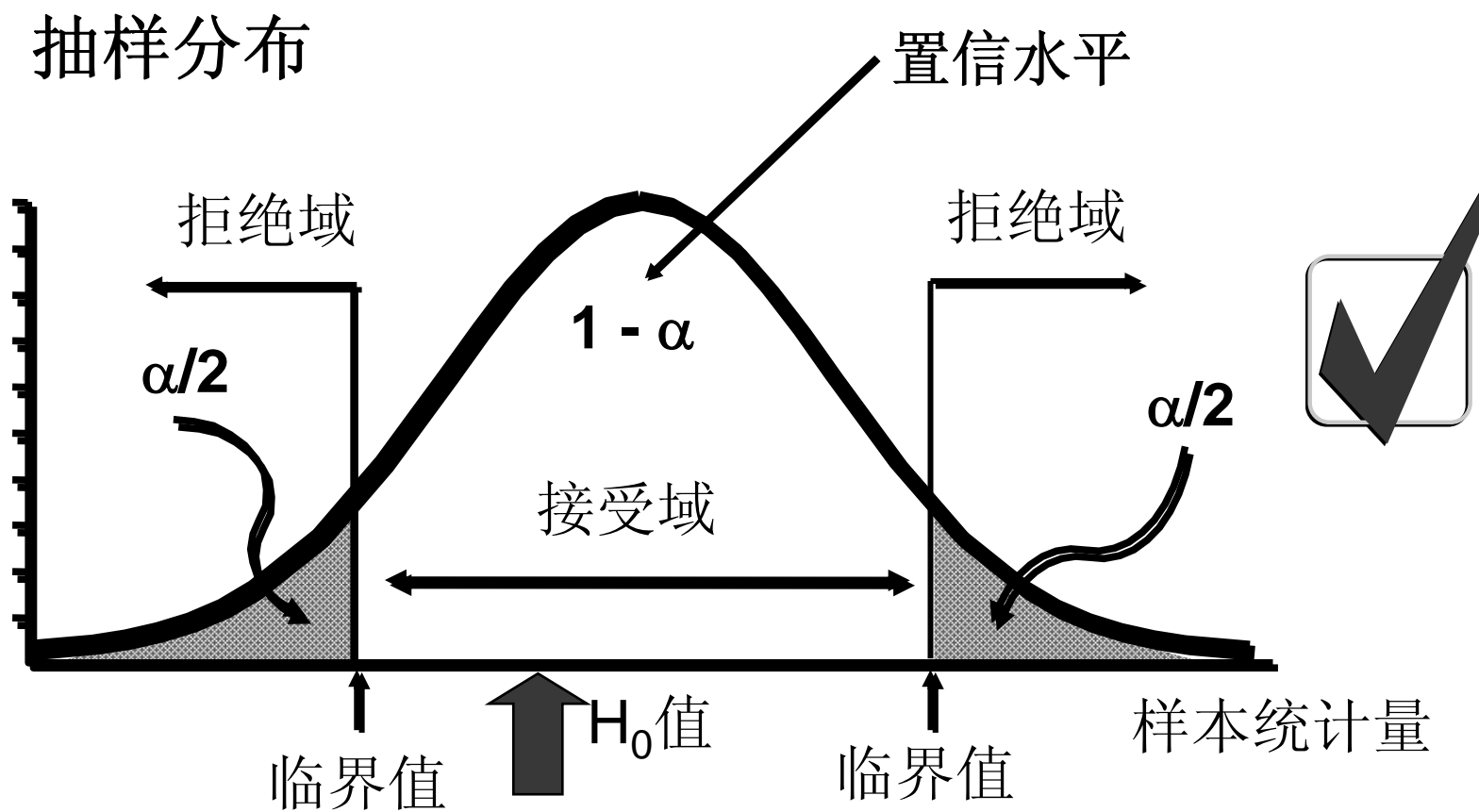
双侧检验

(显著性水平与拒绝域)



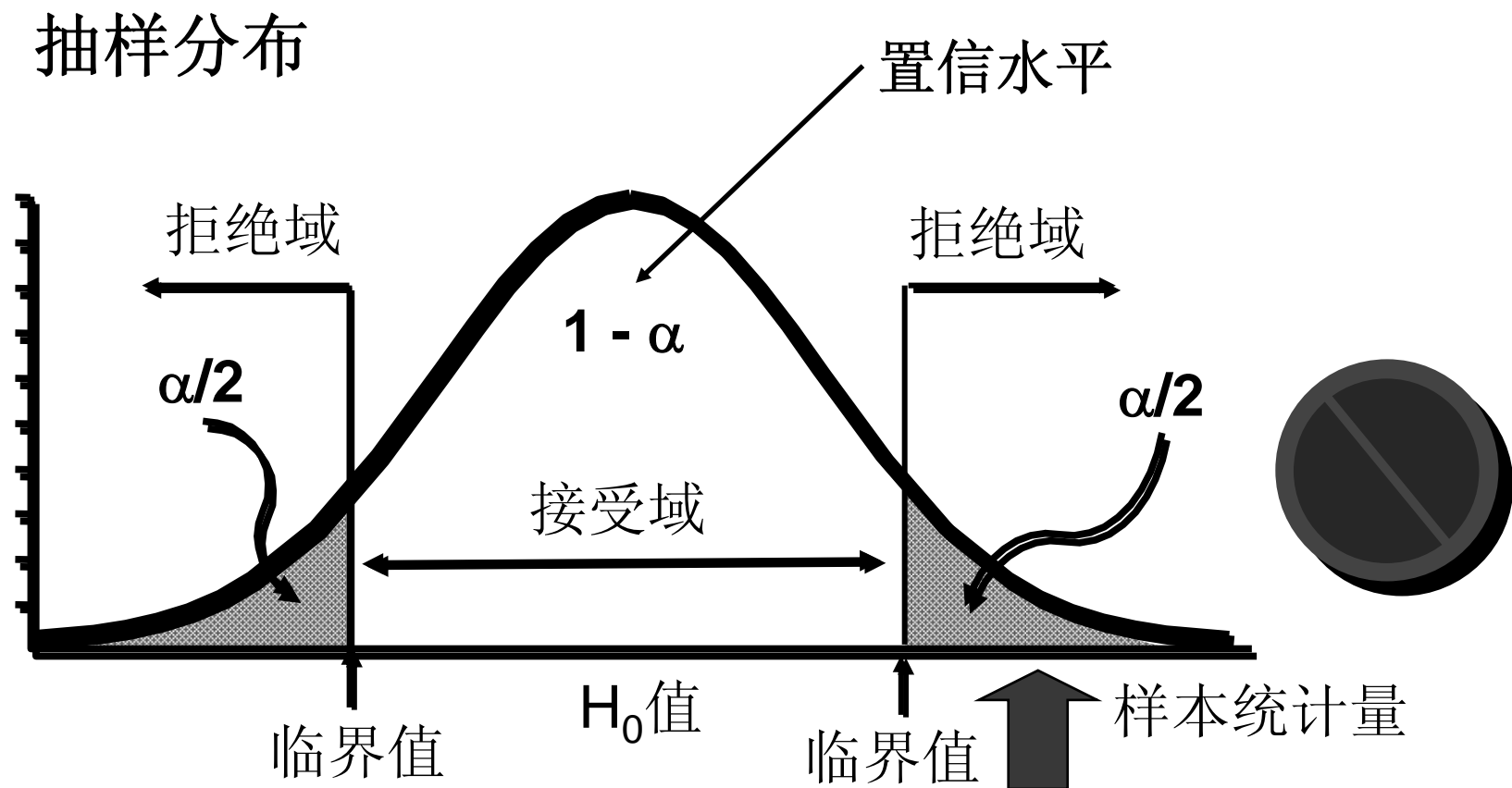
双侧检验

(显著性水平与拒绝域)



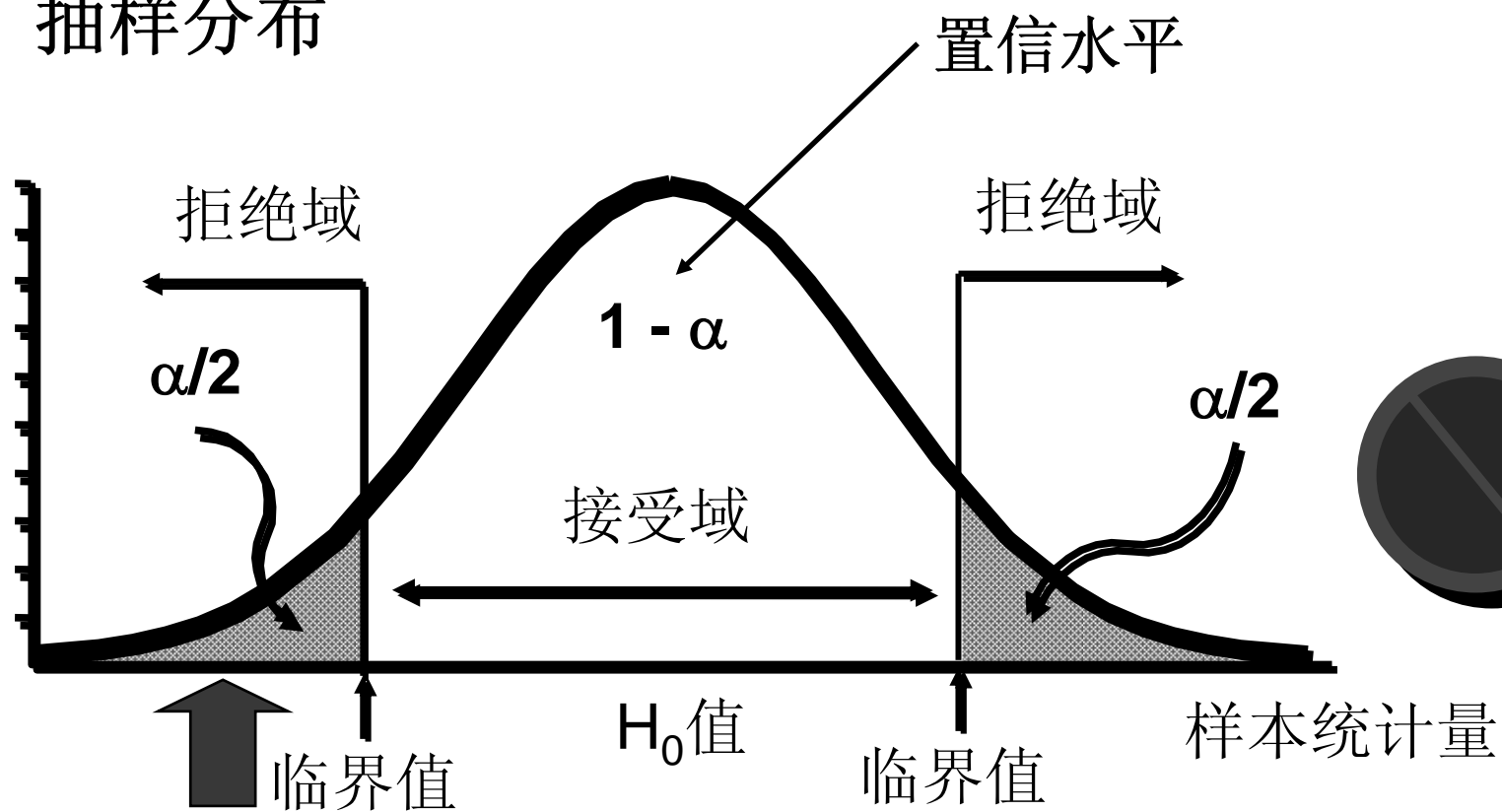
双侧检验

(显著性水平与拒绝域)



双侧检验 (显著性水平与拒绝域)

抽样分布



单侧检验

（原假设与备择假设的确定）

➡ 检验研究中的假设

1. 将所研究的假设作为备择假设 H_1
2. 将认为研究结果是无效的说法或理论作为原假设 H_0 。或者说，把希望(想要)证明的假设作为备择假设
3. 先确立备择假设 H_1

单侧检验

（原假设与备择假设的确定）

□ 例如，采用新技术生产后，将会使产品的使用寿命明显延长到1500小时以上

- 属于研究中的假设
- 建立的原假设与备择假设应为

$$H_0: \mu \leq 1500 \quad H_1: \mu > 1500$$

□ 例如，改进生产工艺后，会使产品的废品率降低到2%以下

- 属于研究中的假设
- 建立的原假设与备择假设应为

$$H_0: \mu \geq 2\% \quad H_1: \mu < 2\%$$

单侧检验

（原假设与备择假设的确定）

➡ 检验某项声明的有效性

1. 将所作出的说明(声明)作为原假设
2. 对该说明的质疑作为备择假设
3. 先确立原假设 H_0
 - 除非我们有证据表明“声明”无效，否则就应认为该“声明”是有效的

单侧检验

（原假设与备择假设的确定）

- 例如，某灯泡制造商声称，该企业所生产的灯泡的平均使用寿命在1000小时以上
 - 除非样本能提供证据表明使用寿命在1000小时以下，否则就应认为厂商的声称是正确的
 - 建立的原假设与备择假设应为
$$H_0: \mu \geq 1000 \quad H_1: \mu < 1000$$

单侧检验 (例子)

□该批产品的平均使用寿命超过**1000**小时吗？

(属于检验声明的有效性，先提出原假设)

提出原假设: $H_0: \mu \geq 1000$

选择备择假设: $H_1: \mu < 1000$

单侧检验 (例子)

□ 学生中经常上网的人数超过**25%**吗？

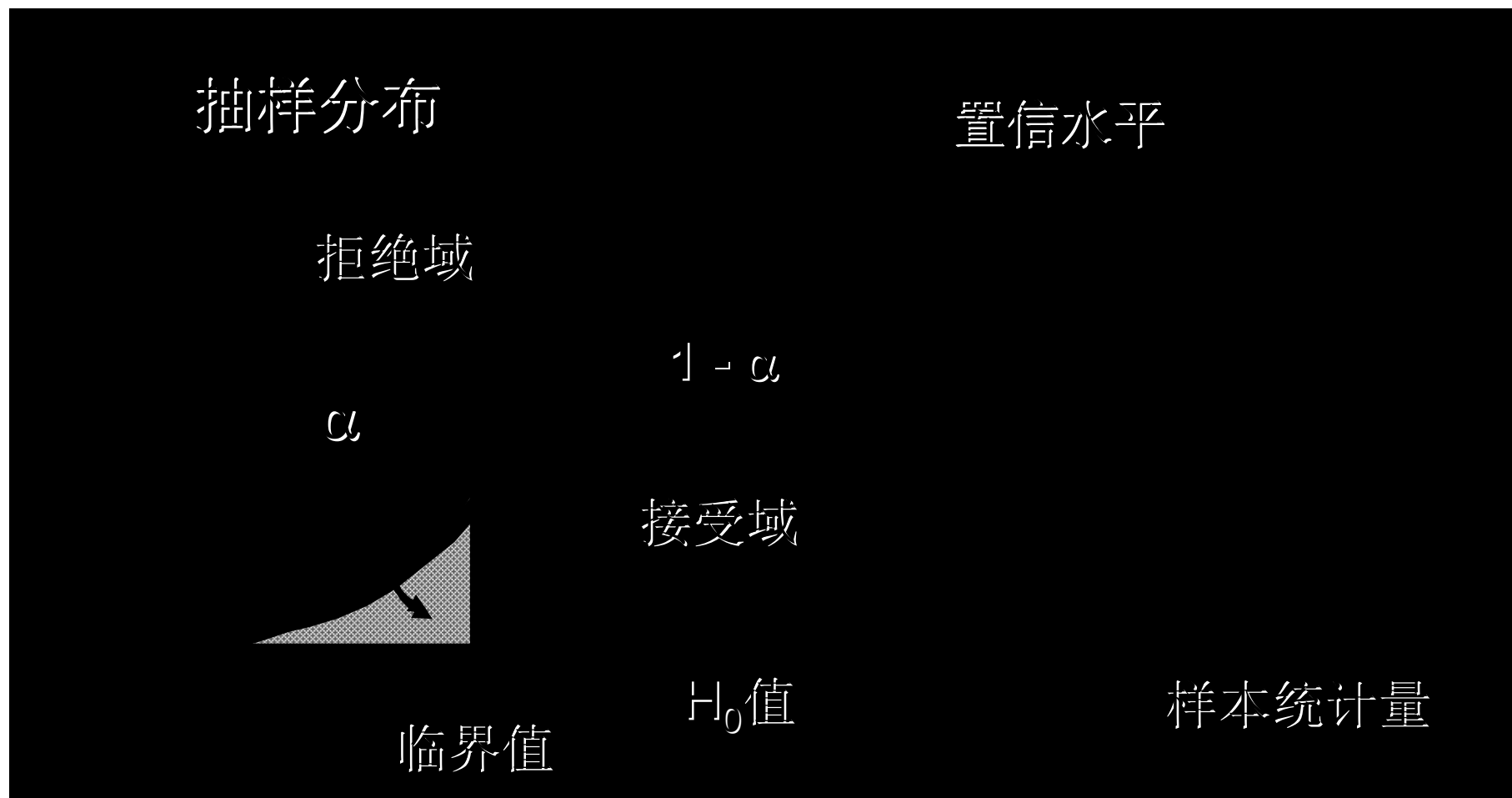
（属于研究中的假设，先提出备择假设）

提出原假设: $H_0: \mu \leq 25$

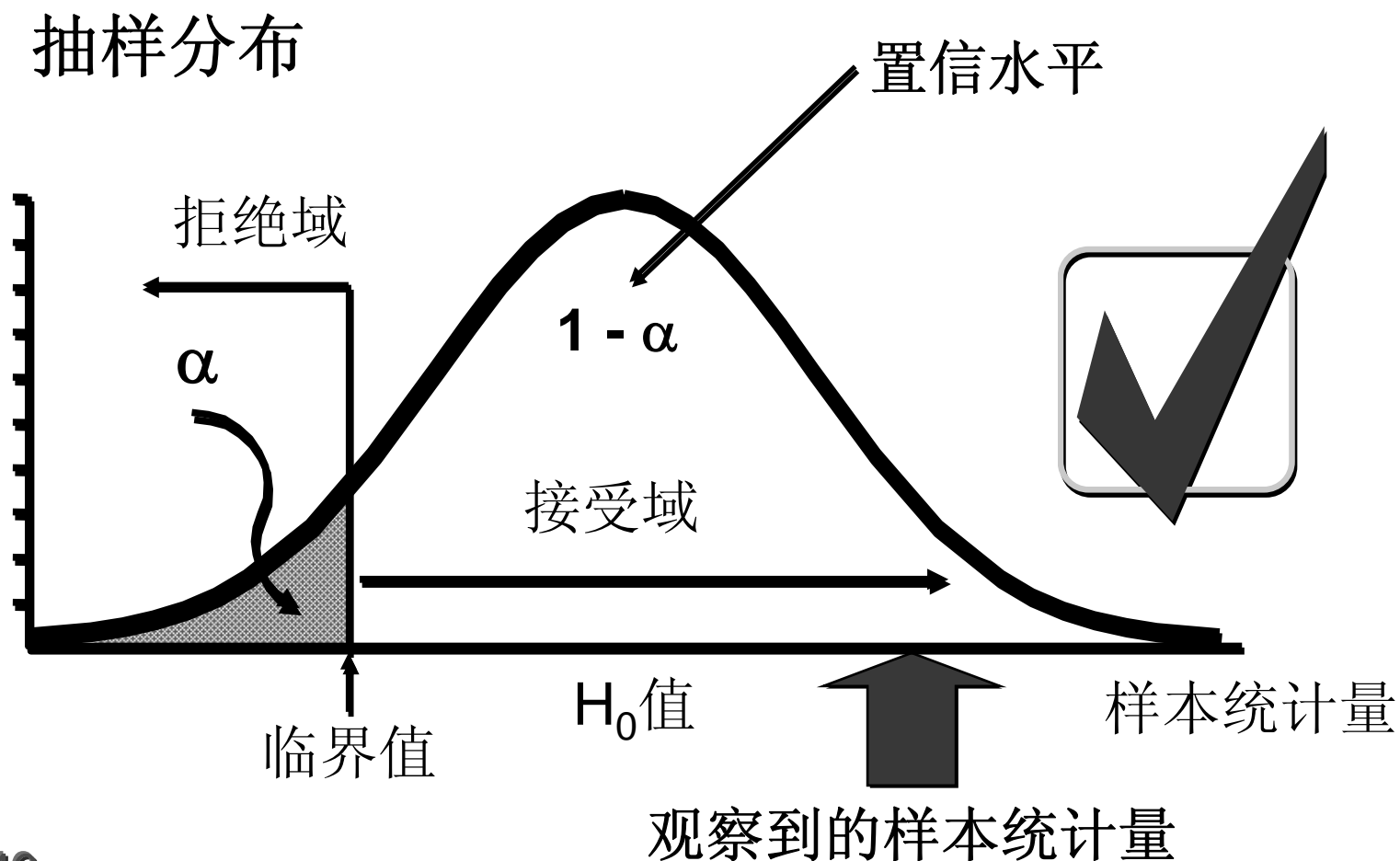
选择备择假设: $H_1: \mu > 25$

单侧检验

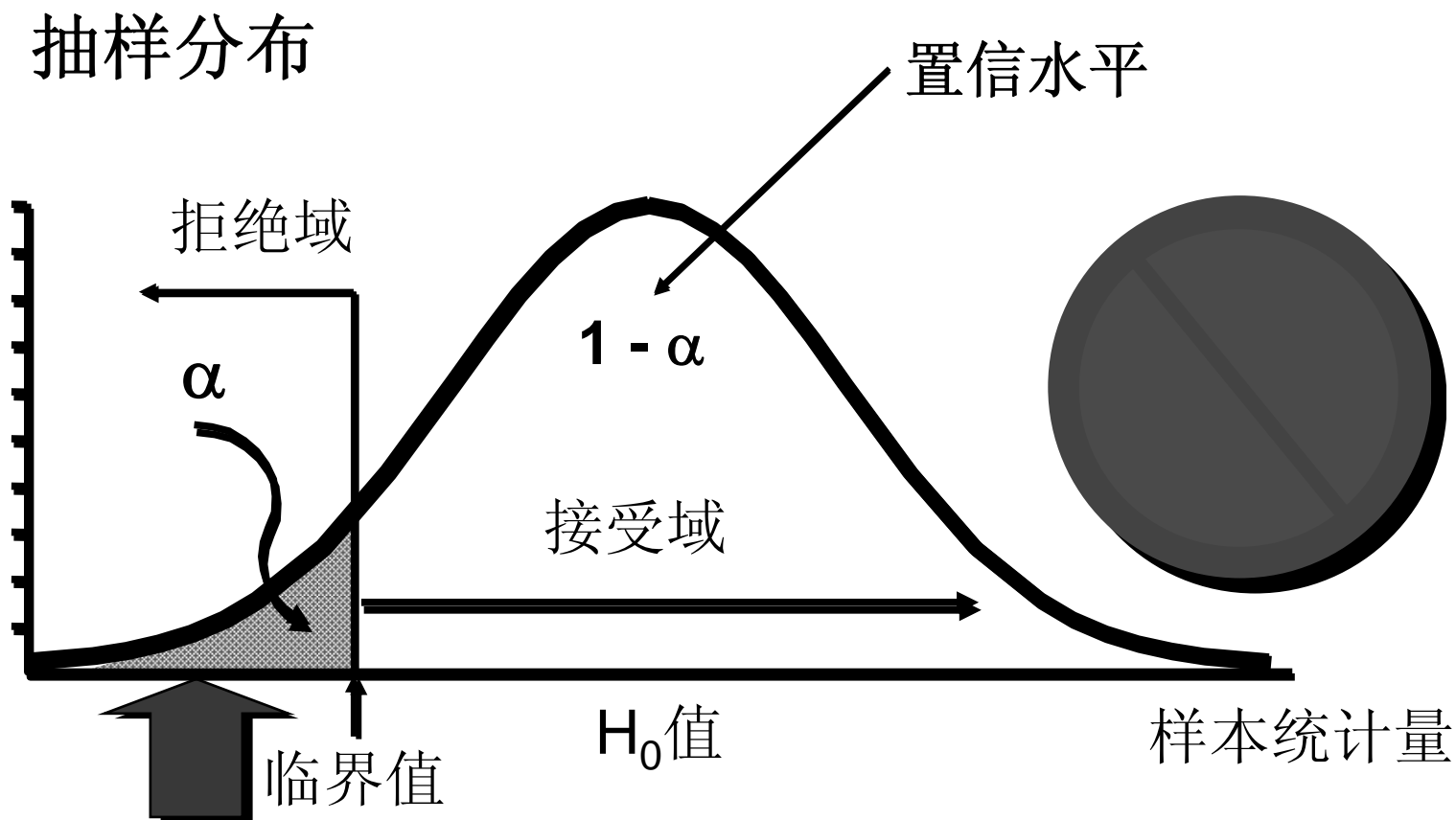
(显著性水平与拒绝域)



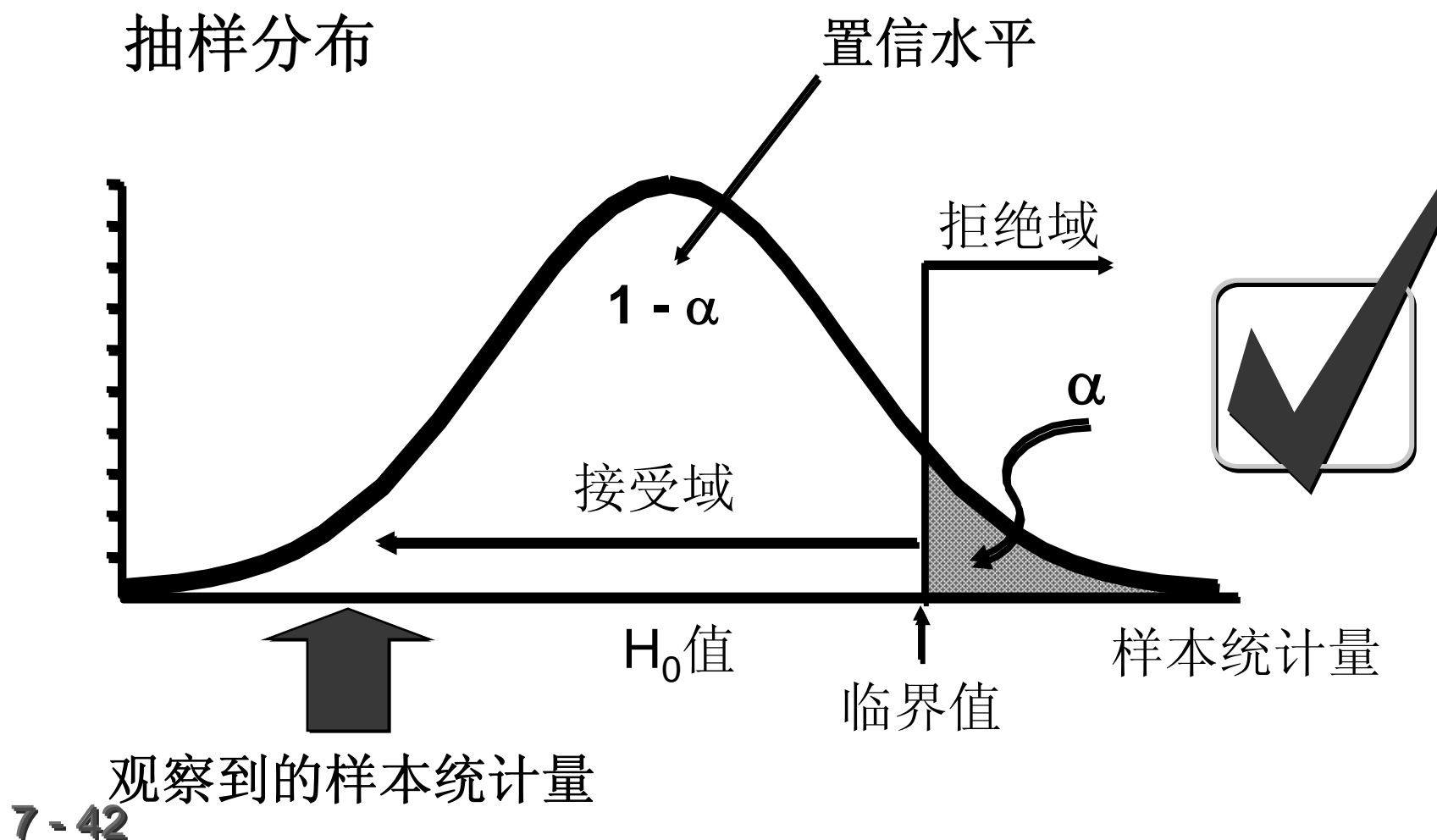
左侧检验 (显著性水平与拒绝域)



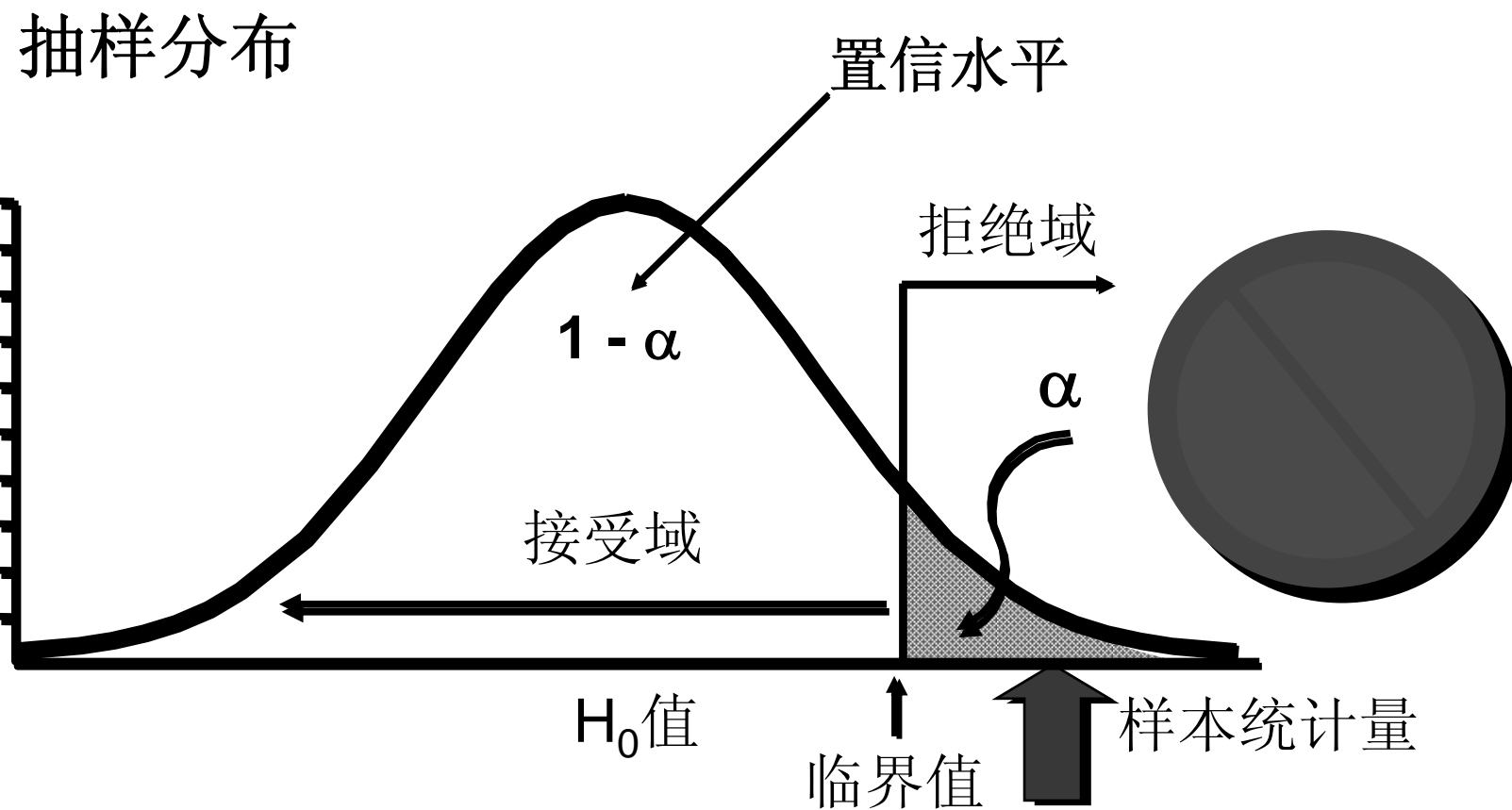
左侧检验 (显著性水平与拒绝域)



右侧检验 (显著性水平与拒绝域)



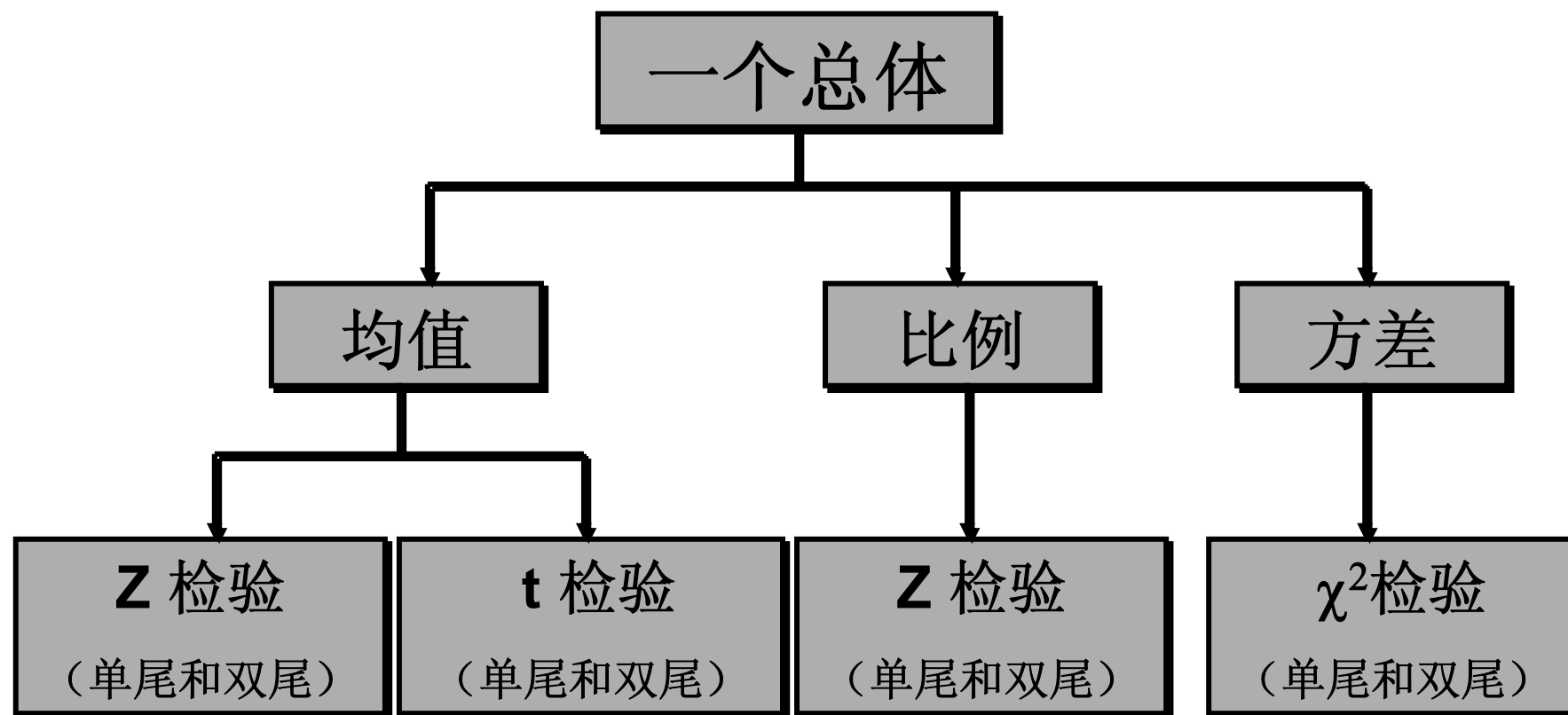
右侧检验 (显著性水平与拒绝域)



第二节 一个正态总体的参数检验

- 一. 总体方差已知时的均值检验
- 二. 总体方差未知时的均值检验
- 三. 总体比例的假设检验

一个总体的检验

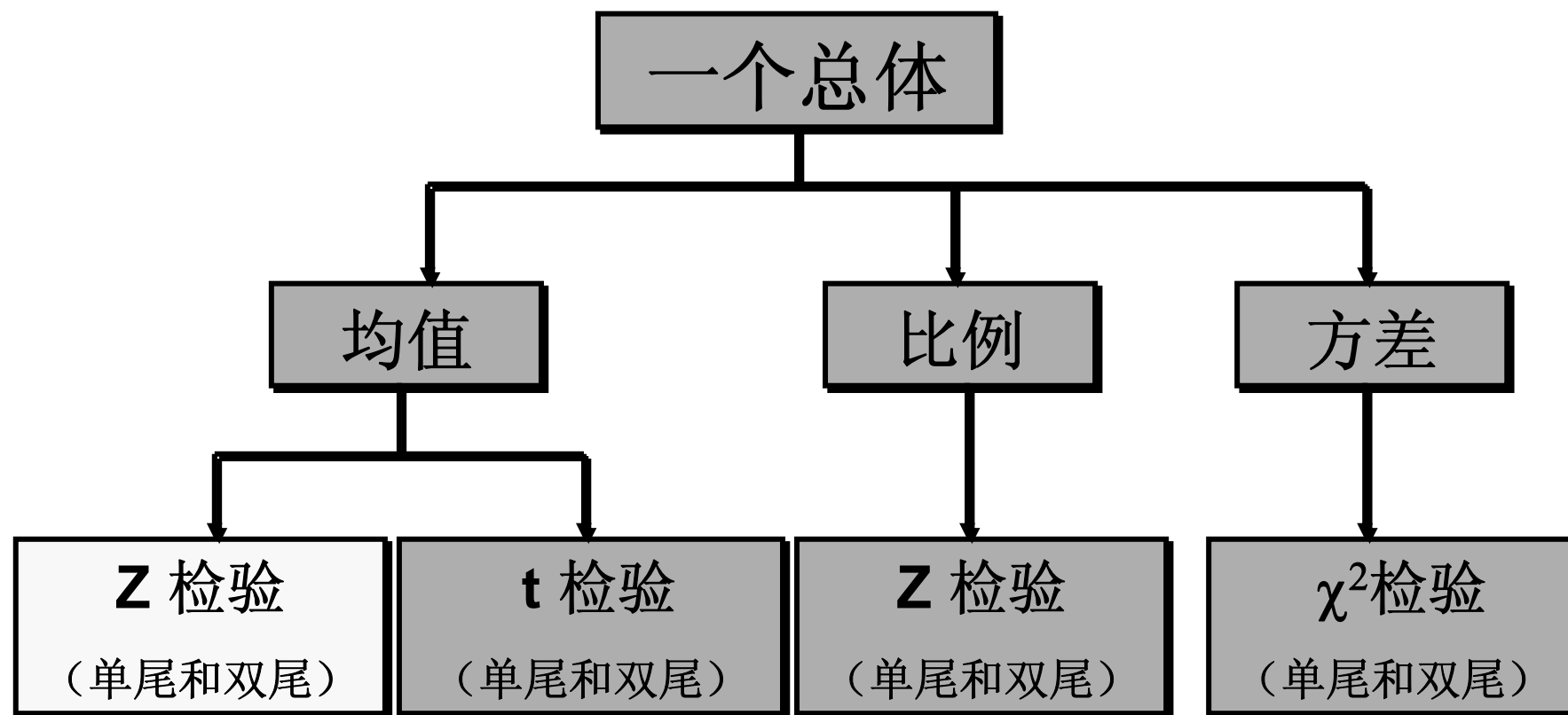


检验的步骤

- 陈述原假设 H_0
- 陈述备择假设 H_1
- 选择显著性水平 α
- 选择检验统计量
- 选择 n
- 给出临界值
- 搜集数据
- 计算检验统计量
- 进行统计决策
- 表述决策结果

总体方差已知时的均值检验 (双尾 Z 检验)

一个总体的检验



均值的双尾 Z 检验 (σ^2 已知)

1. 假定条件

- 总体服从正态分布
- 若不服从正态分布，可用正态分布来近似
($n \geq 30$)

2. 原假设为： $H_0: \mu = \mu_0$ ；备择假设为： $H_1: \mu \neq \mu_0$

3. 使用z-统计量

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

均值的双尾 Z 检验 (实例)

【例】某机床厂加工一种零件，根据经验知道，该厂加工零件的椭圆度近似服从正态分布，其总体均值为 $\mu_0=0.081\text{mm}$ ，总体标准差为 $\sigma=0.025$ 。今换一种新机床进行加工，抽取 $n=200$ 个零件进行检验，得到的椭圆度为 0.076mm 。试问新机床加工零件的椭圆度的均值与以前有无显著差异？（ $\alpha=0.05$ ）

属于决策中的
假设！

均值的双尾 Z 检验 (计算结果)

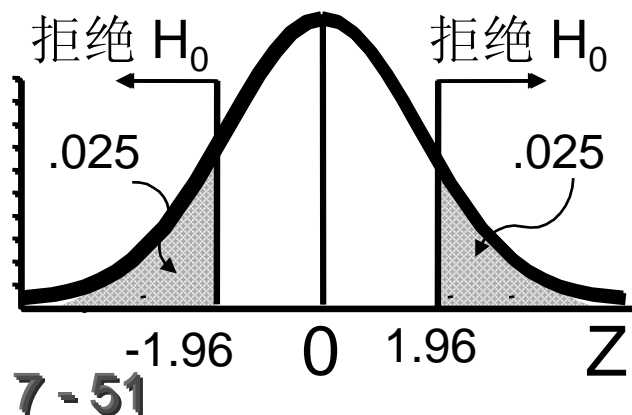
$H_0: \mu = 0.081$

$H_1: \mu \neq 0.081$

$\alpha = 0.05$

$n = 200$

临界值(s):



检验统计量:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{0.076 - 0.081}{0.025 / \sqrt{200}} = -2.83$$

决策:

拒绝 H_0

结论:

有证据表明新机床加工的零件的椭圆度与以前有显著差异

总体方差已知时的均值检验 (单尾 Z 检验)

均值的单尾 Z 检验 (σ^2 已知)

1. 假定条件

- 总体服从正态分布
- 若不服从正态分布，可以用正态分布来近似 ($n \geq 30$)

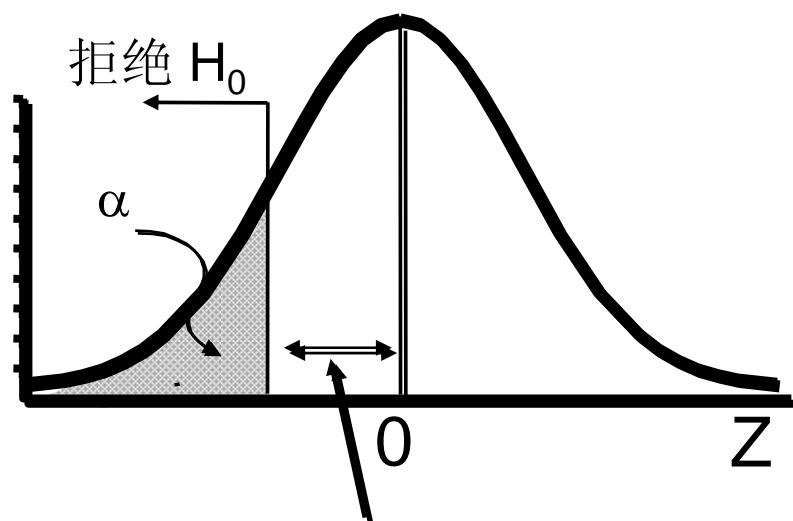
2. 备择假设有<或>符号

3. 使用z-统计量

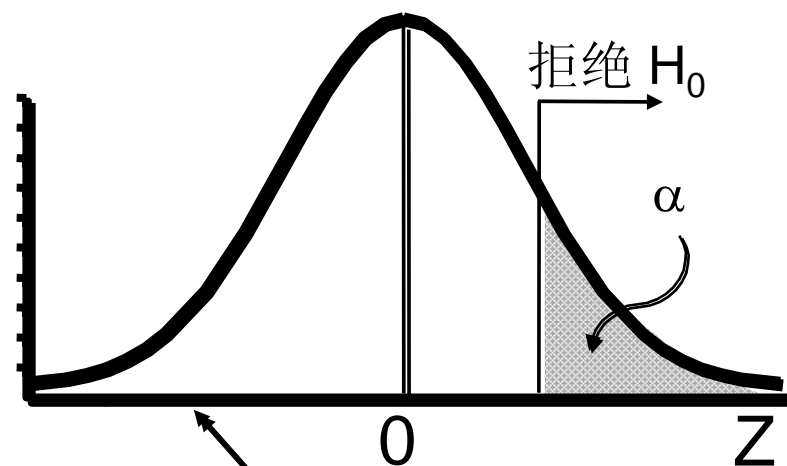
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

均值的单尾 Z 检验 (提出假设)

左侧: $H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$ 右侧: $H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$



必须是显著地低于 μ_0 , 大的值满足 H_0 , 不能拒绝



必须显著地大于 μ_0 , 小的值满足 H_0 , 不能拒绝

均值的单尾Z检验 (实例)

【例】某批发商欲从生产厂家购进一批灯泡，根据合同规定，灯泡的使用寿命平均不能低于1000小时。已知灯泡使用寿命服从正态分布，标准差为20小时。在总体中随机抽取100只灯泡，测得样本均值为960小时。批发商是否应该购买这批灯泡？ ($\alpha=0.05$)

属于检验声明
的有效性！



均值的单尾Z检验 (计算结果)

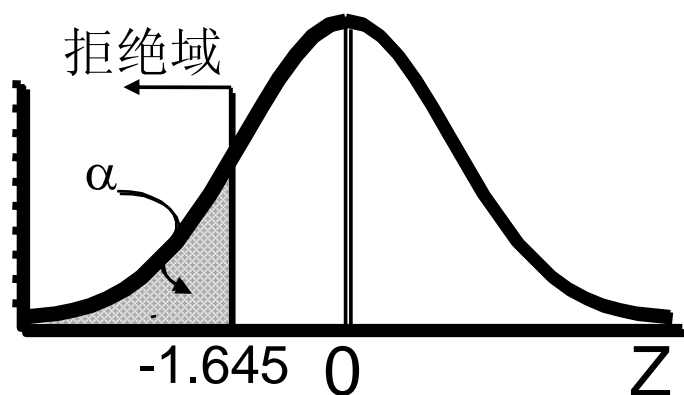
$$H_0: \mu \geq 1000$$

$$H_1: \mu < 1000$$

$$\alpha = 0.05$$

$$n = 100$$

临界值(s):



7-56

检验统计量:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{960 - 1000}{20 / \sqrt{100}} = -2$$

决策:

在 $\alpha = 0.05$ 的水平上拒绝 H_0

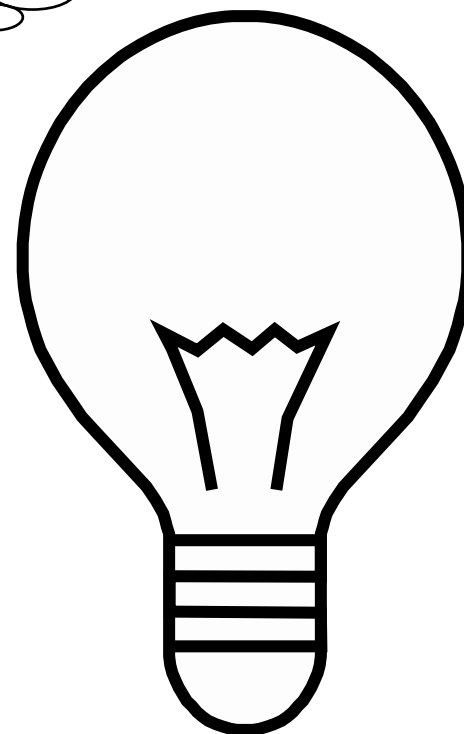
结论:

有证据表明这批灯泡的使用寿命低于1000小时

均值的单尾Z检验 (实例)

【例】 根据过去大量资料，某厂生产的灯泡的使用寿命服从正态分布 $N(1020, 100^2)$ 。现从最近生产的一批产品中随机抽取16只，测得样本平均寿命为1080小时。试在0.05的显著性水平下判断这批产品的使用寿命是否有显著提高？($\alpha=0.05$)

属于研究中的假设！



均值的单尾Z检验 (计算结果)

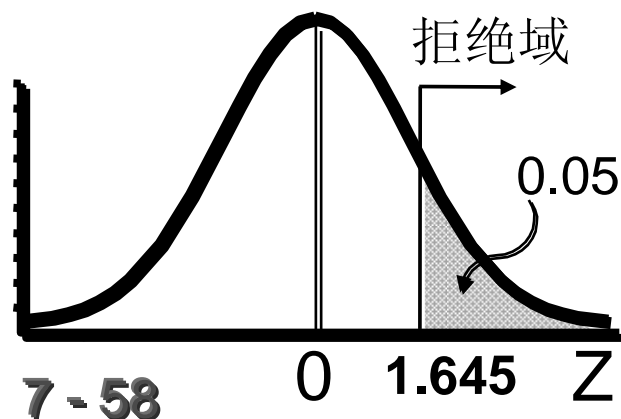
$$H_0: \mu \leq 1020$$

$$H_1: \mu > 1020$$

$$\alpha = 0.05$$

$$n = 16$$

临界值(s):



检验统计量:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{1080 - 1020}{100 / \sqrt{14}} = 2.4$$

决策:

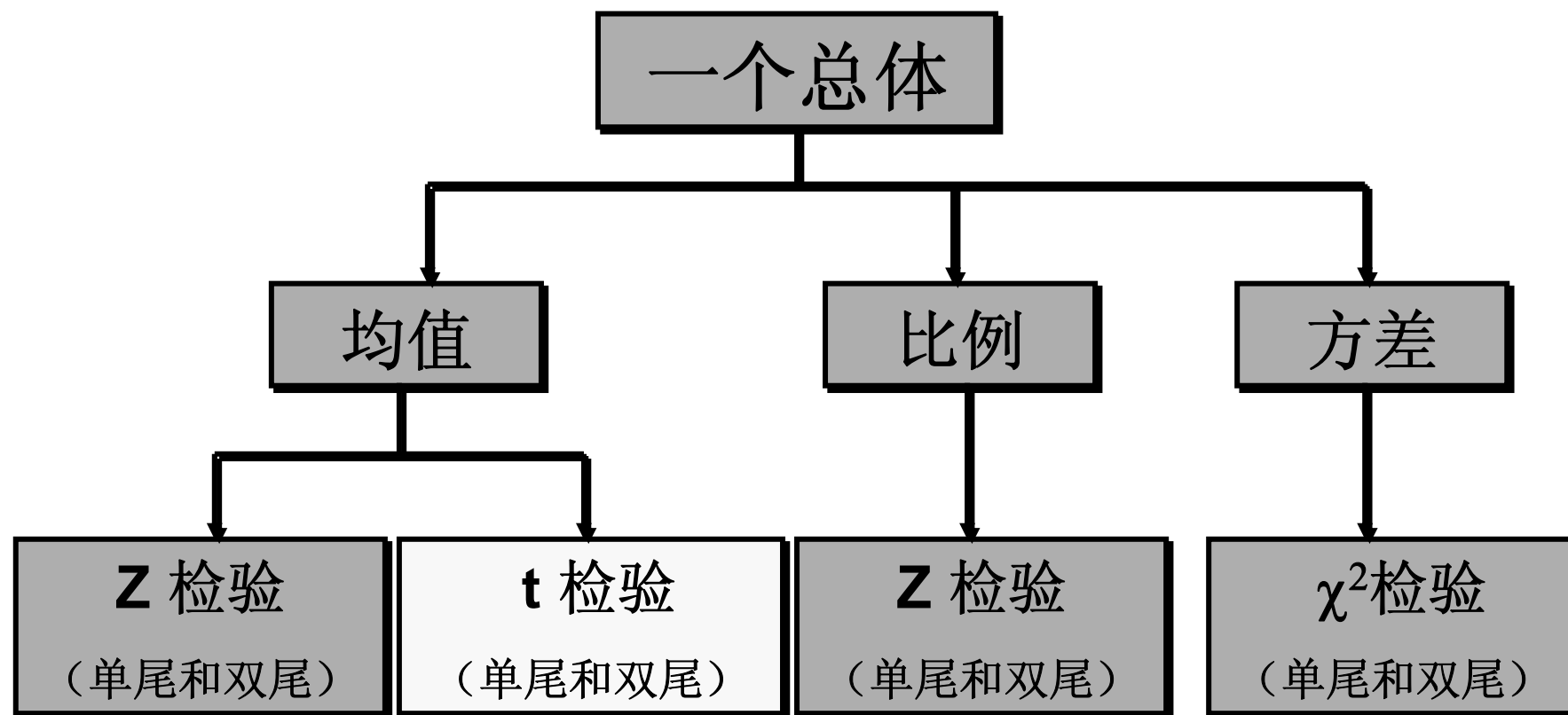
在 $\alpha = 0.05$ 的水平上拒绝 H_0

结论:

有证据表明这批灯泡的使用寿命有显著提高

总体方差未知时的均值检验 (双尾 t 检验)

一个总体的检验



均值的双尾 t 检验 (σ^2 未知)

1. 假定条件

- 总体为正态分布
- 如果不是正态分布, 只有轻微偏斜和大样本 ($n \geq 30$) 条件下

2. 使用 t 统计量

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \sim t(n-1)$$

均值的双尾 t 检验 (实例)

【例】某厂采用自动包装机分装产品，假定每包产品的重量服从正态分布，每包标准重量为1000克。某日随机抽查9包，测得样本平均重量为986克，样本标准差为24克。试问在0.05的显著性水平上，能否认为这天自动包装机工作正常？

属于决策中的假设！



均值的双尾 t 检验 (计算结果)

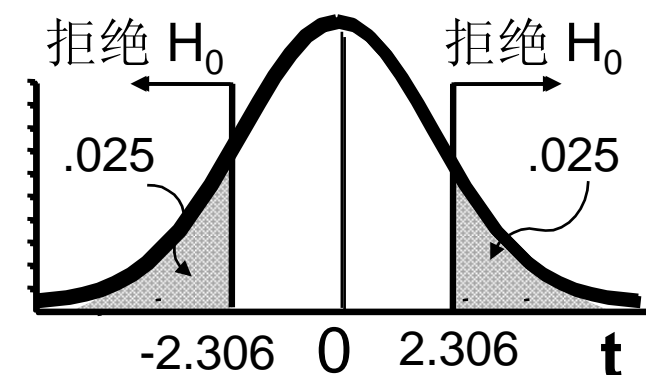
$$H_0: \mu = 1000$$

$$H_1: \mu \neq 1000$$

$$\alpha = 0.05$$

$$df = 9 - 1 = 8$$

临界值(s):



7-63

检验统计量:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{986 - 1000}{24/\sqrt{9}} = -1.75$$

决策:

在 $\alpha = 0.05$ 的水平上接受 H_0

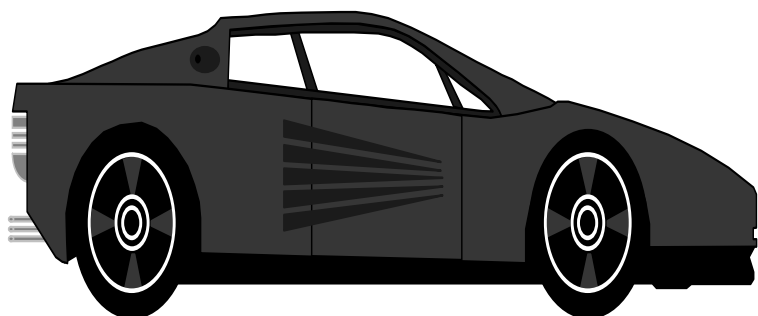
结论:

有证据表明这天自动包装机
工作正常

总体方差未知时的均值检验 (单尾 t 检验)

均值的单尾 t 检验 (实例)

属于检验声明有
效性的假设!



7-65

【例】一个汽车轮胎制造商声称，某一等级的轮胎的平均寿命在一定的汽车重量和正常行驶条件下大于**40000**公里，对一个由**20**个轮胎组成的随机样本作了试验，测得平均值为**41000**公里，标准差为**5000**公里。已知轮胎寿命的公里数服从正态分布，我们能否根据这些数据作出结论，该制造商的产品同他所说的标准相符？($\alpha = 0.05$)

均值的单尾 t 检验 (计算结果)

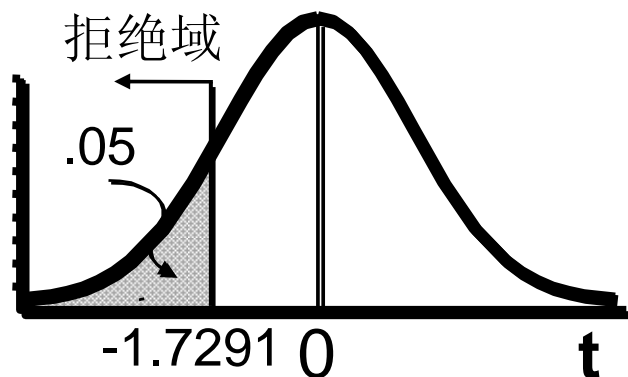
$$H_0: \mu \geq 40000$$

$$H_1: \mu < 40000$$

$$\alpha = 0.05$$

$$df = 20 - 1 = 19$$

临界值(s):



7 - 66

检验统计量:

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \\ &= \frac{41000 - 40000}{5000 / \sqrt{20}} = 0.894 \end{aligned}$$

决策:

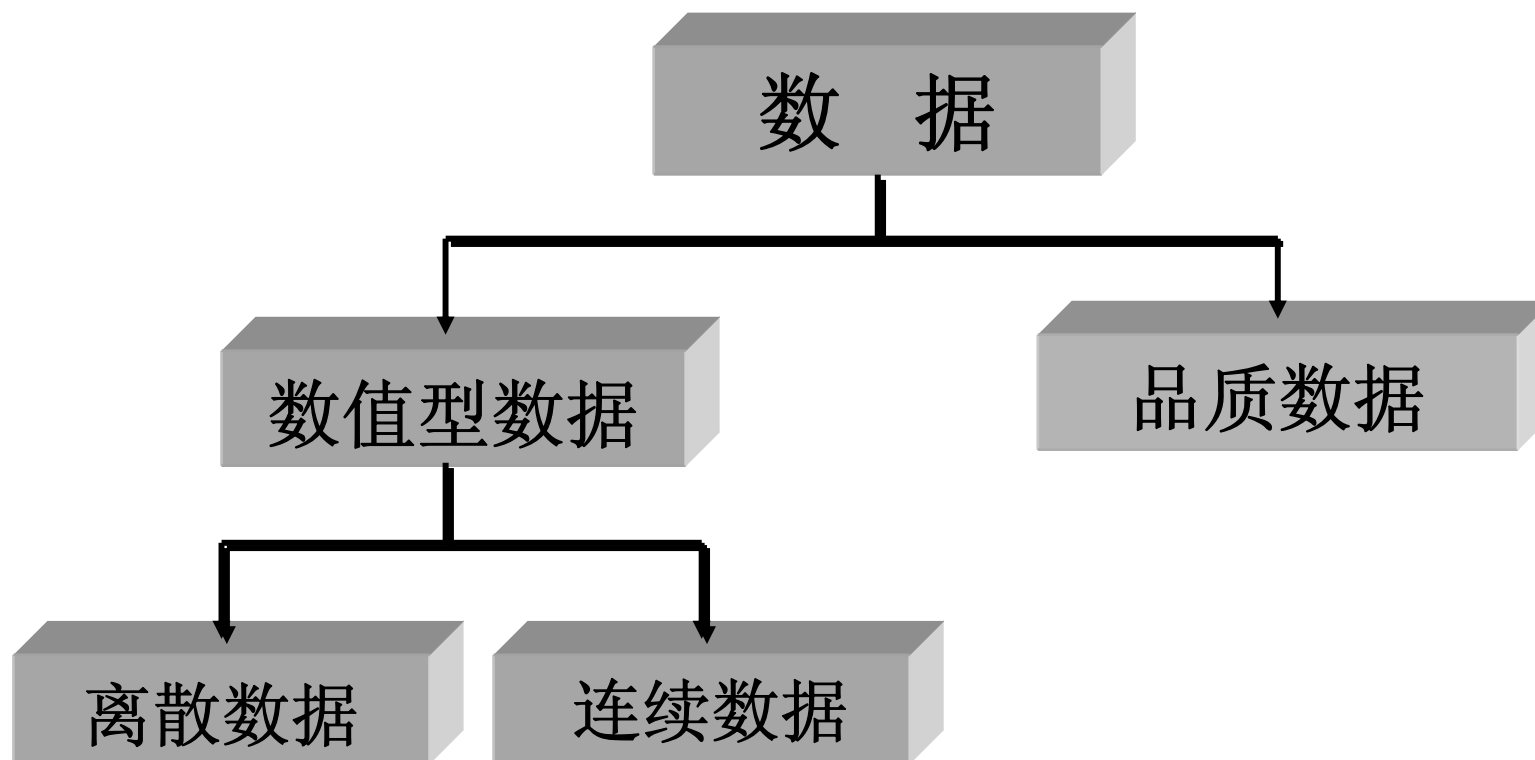
在 $\alpha = 0.05$ 的水平上接受 H_0

结论:

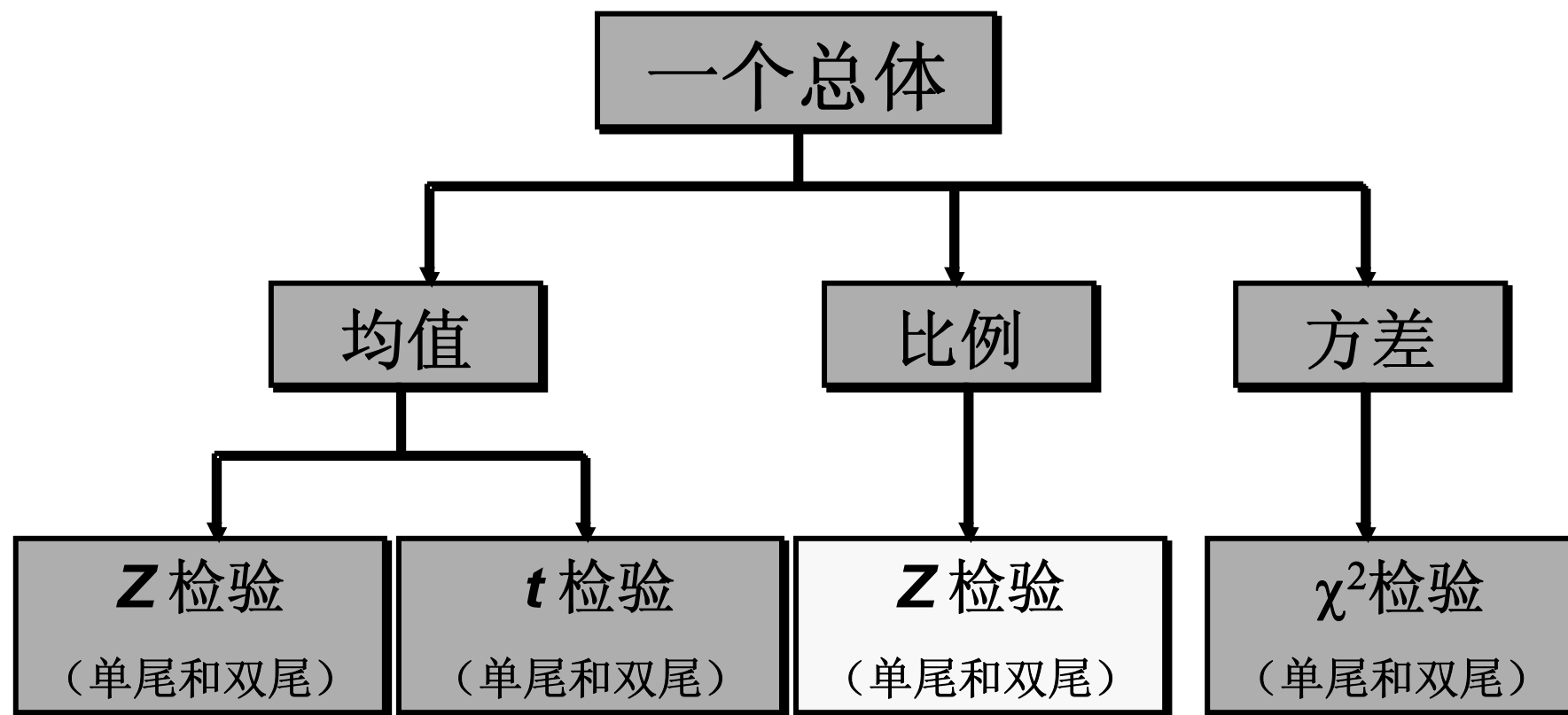
有证据表明轮胎使用寿命显著地大于40000公里

总体比例的假设检验 (Z 检验)

适用的数据类型



一个总体的检验



一个总体比例的 Z 检验

1. 假定条件

- 有两类结果
- 总体服从二项分布
- 可用正态分布来近似

2. 比例检验的 z 统计量

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

P_0 为假设的总体比例

一个总体比例的 Z 检验 (实例)

【例】某研究者估计本市居民家庭的电脑拥有率为**30%**。现随机抽查了**200**的家庭，其中**68**个家庭拥有电脑。试问研究者的估计是否可信？
($\alpha = 0.05$)



一个样本比例的 Z 检验 (结果)

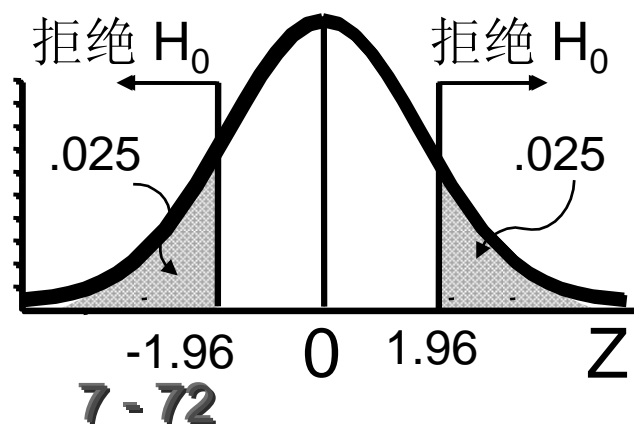
$$H_0: p = 0.3$$

$$H_1: p \neq 0.3$$

$$\alpha = 0.05$$

$$n = 200$$

临界值(s):



检验统计量:

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\ &= \frac{0.34 - 0.3}{\sqrt{\frac{0.3 \times 0.7}{200}}} = 1.234 \end{aligned}$$

决策:

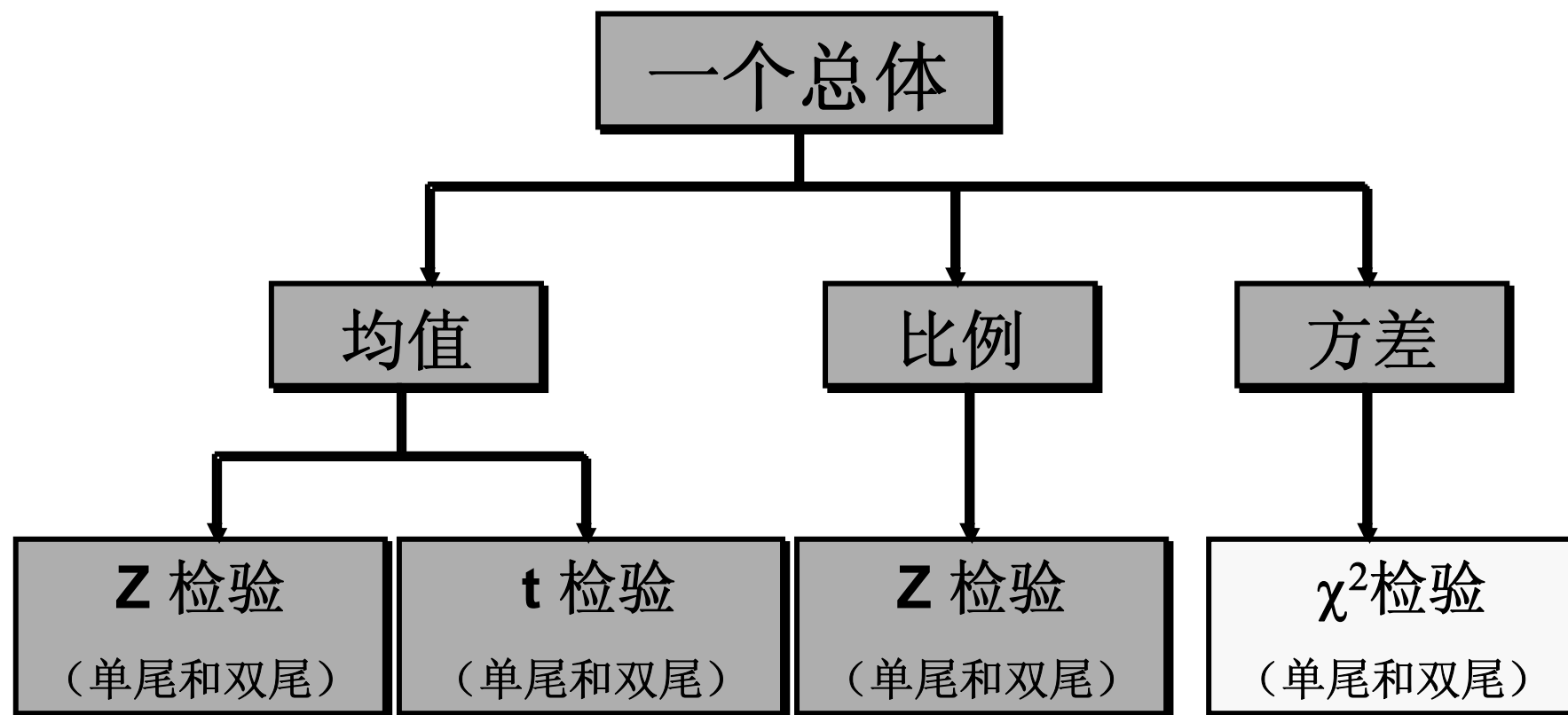
在 $\alpha = 0.05$ 的水平上接受 H_0

结论:

有证据表明研究者的估计可信

总体方差的检验 (χ^2 检验)

一个总体的检验



方差的卡方 (χ^2) 检验

1. 检验一个总体的方差或标准差
2. 假设总体近似服从正态分布
3. 原假设为 $H_0: \sigma^2 = \sigma_0^2$
4. 检验统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$$

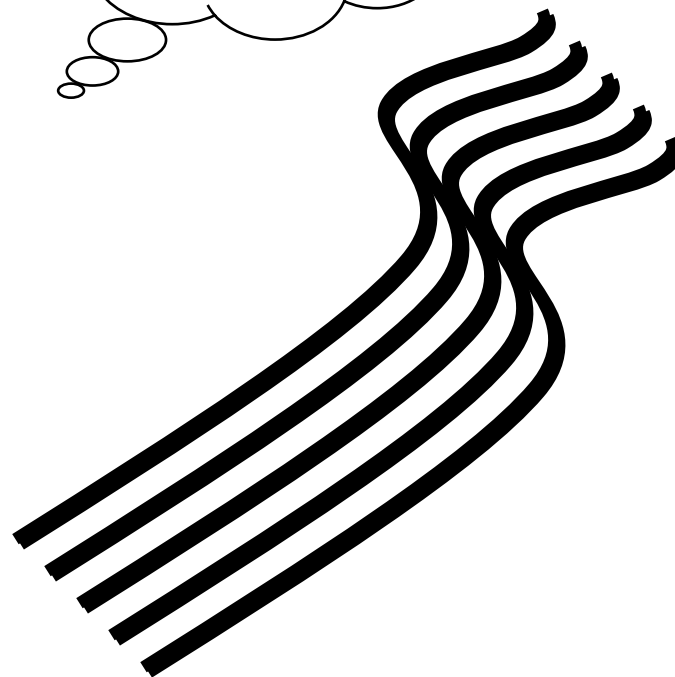
样本方差

假设的总体方差

卡方 (χ^2) 检验 实例

【例】根据长期正常生产的资料可知，某厂所产维尼纶的纤度服从正态分布，其方差为**0.0025**。现从某日产品中随机抽取**20**根，测得样本方差为**0.0042**。试判断该日纤度的波动与平日有无显著差异？($\alpha=0.05$)

属于决策中的
假设！



卡方 (χ^2) 检验 计算结果

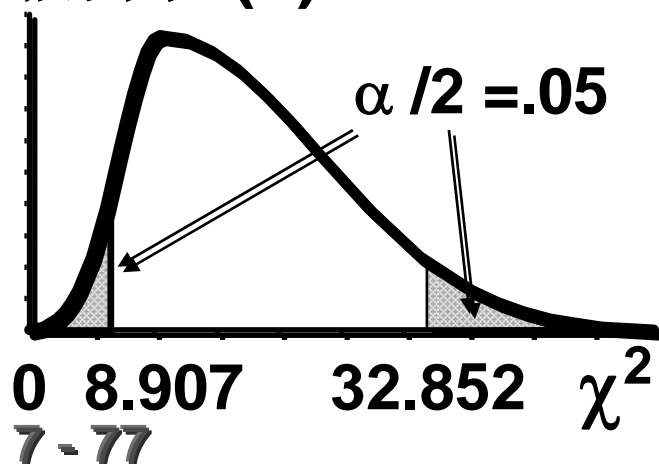
$$H_0: \sigma^2 = 0.0025$$

$$H_1: \sigma^2 \neq 0.0025$$

$$\alpha = 0.05$$

$$df = 20 - 1 = 19$$

临界值(s):



统计量:

$$\begin{aligned}\chi^2 &= \frac{(n-1)s^2}{\sigma_0^2} \\ &= \frac{(20-1)0.0042}{0.0025} = 31.92\end{aligned}$$

决策:

在 $\alpha = 0.05$ 的水平上接受 H_0

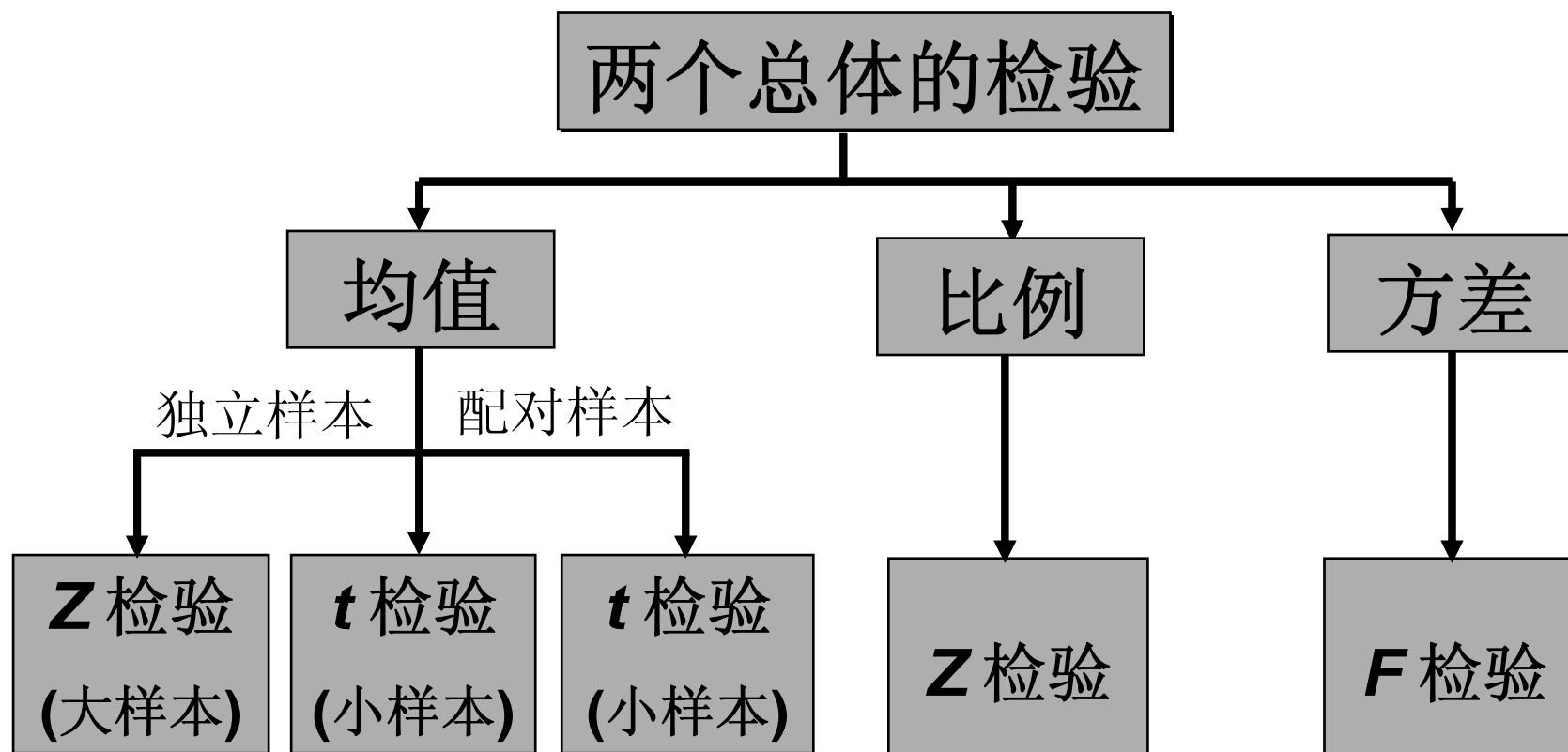
结论:

有证据表明该日纤度的波动比平时没有显著差异

第三节 两个正态总体的参数检验

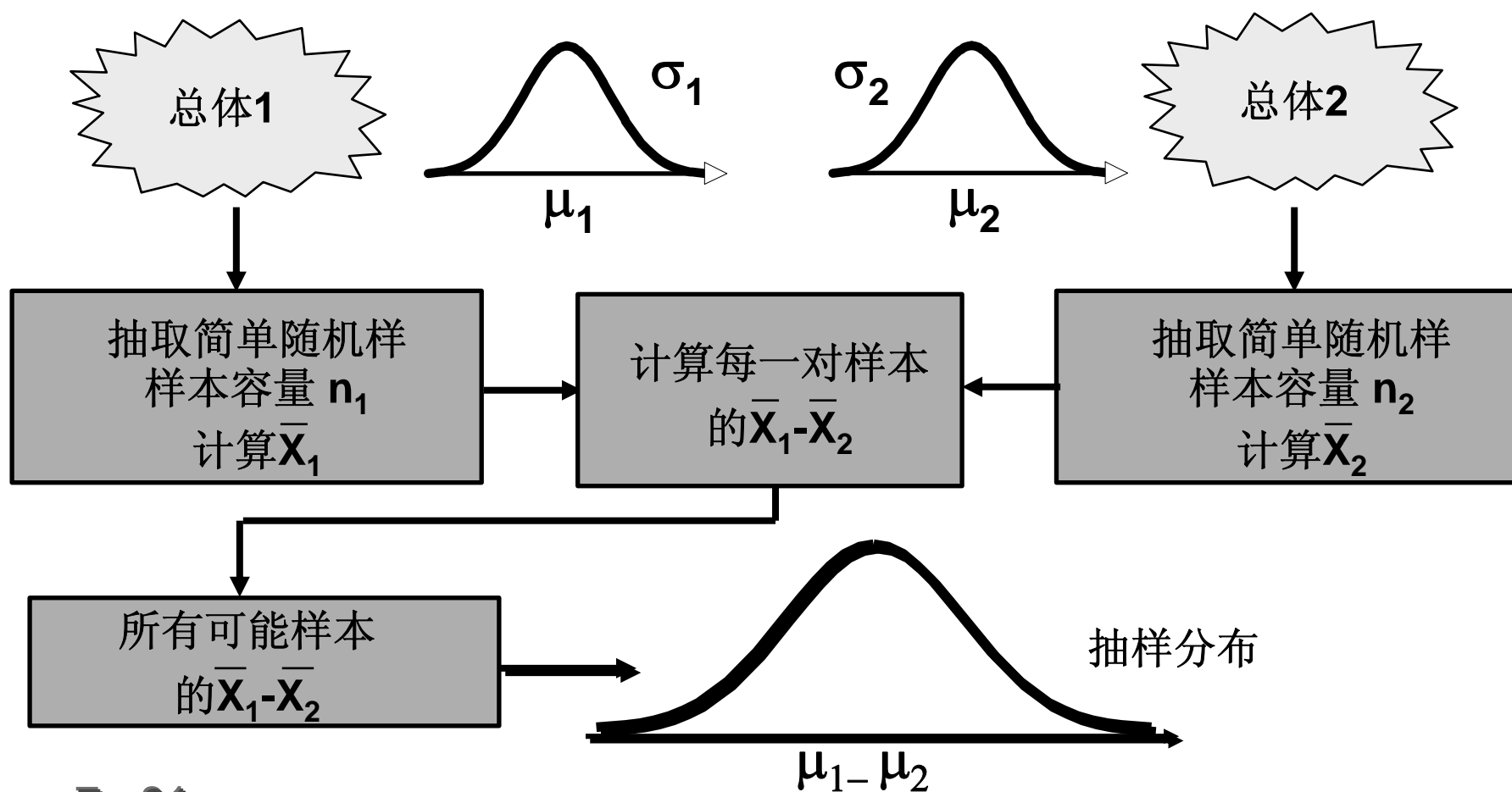
- 一. 两个总体参数之差的抽样分布
- 二. 两个总体均值之差的检验
- 三. 假设检验中相关样本的利用
- 四. 两个总体比例之差的检验

两个正态总体的参数检验



两个独立样本的均值检验

两个独立样本之差的抽样分布



两个总体均值之差的Z检验 (σ_1^2 、 σ_2^2 已知)

1. 假定条件

- 两个样本是独立的随机样本
- 两个总体都是正态分布
- 若不是正态分布，可以用正态分布来近似 ($n_1 \geq 30$ 和 $n_2 \geq 30$)

2. 原假设： $H_0: \mu_1 - \mu_2 = 0$ ，备择假设： $H_1: \mu_1 - \mu_2 \neq 0$

3. 检验统计量为

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

两个总体均值之差的Z检验 (假设的形式)

假设	研究的问题		
	没有差异 有差异	均值 ₁ ≥ 均值 ₂ 均值 ₁ < 均值 ₂	均值 ₁ ≤ 均值 ₂ 均值 ₁ > 均值 ₂
H₀	$\mu - \mu = 0$	$\mu - \mu \geq 0$	$\mu - \mu \leq 0$
H₁	$\mu - \mu \neq 0$	$\mu - \mu < 0$	$\mu - \mu > 0$

两个总体均值之差的Z检验 (例子)

【例】有两种方法可用于制造某种以抗拉强度为重要特征的产品。根据以往的资料得知，第一种方法生产出的产品其抗拉强度的标准差为**8**公斤，第二种方法的标准差为**10**公斤。从两种方法生产的产品中各抽取一个随机样本，样本容量分别为 **$n_1=32$** ， **$n_2=40$** ，测得 $\bar{x}_2=50$ 公斤， $\bar{x}_1=44$ 公斤。问这两种方法生产的产品平均抗拉强度是否有显著差别？ ($\alpha = 0.05$)

属于决策中的假设！

两个总体均值之差的Z检验 (计算结果)

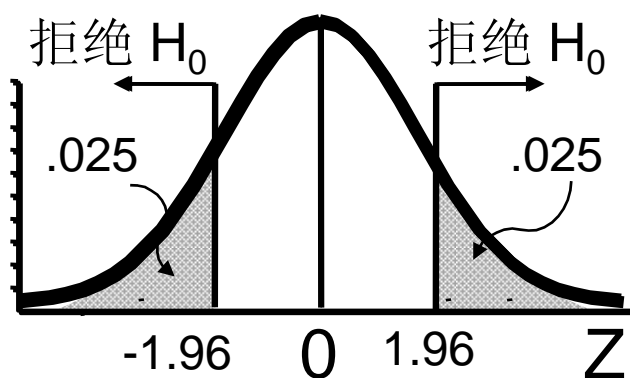
$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

$$\alpha = 0.05$$

$$n_1 = 32, n_2 = 40$$

临界值(s):



7-85

检验统计量:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{50 - 40 - 0}{\sqrt{\frac{64}{32} + \frac{100}{40}}} = 2.83$$

决策:

拒绝 H_0

结论:

有证据表明两种方法生产的产品其抗拉强度有显著差异

两个总体均值之差的 t 检验 (σ_1^2 、 σ_2^2 未知)

1. 检验具有等方差的两个总体的均值

2. 假定条件

- 两个样本是独立的随机样本
- 两个总体都是正态分布
- 两个总体方差未知但相等 $\sigma_1^2 = \sigma_2^2$

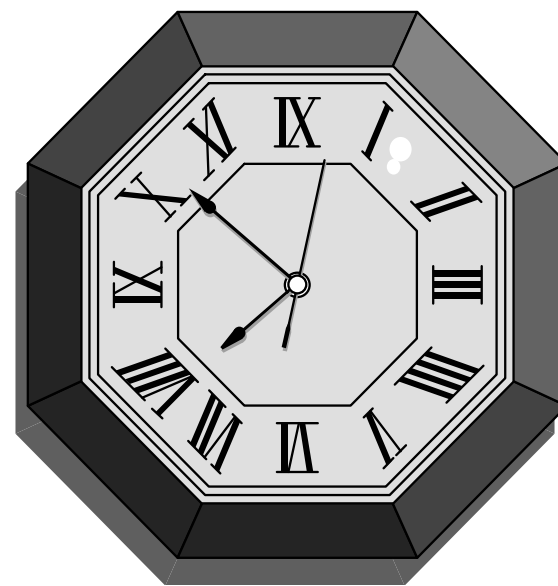
3. 检验统计量

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{其中: } S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

两个总体均值之差的 t 检验 (例子)

【例】一个车间研究用两种不同的工艺组装某种产品所用的时间是否相同。让一个组的**10**名工人用第一种工艺组装该产品，平均所需时间为**26.1**分钟，样本标准差为**12**分钟；另一组**8**名工人用第二种工艺组装，平均所需时间为**17.6**分钟，样本标准差为**10.5**分钟。已知用两种工艺组装产品所用时间服从正态分布，且 $\sigma_1^2 = \sigma_2^2$ 。试问能否认为用第二种方法组装比用第一中方法组装更好？($\alpha = 0.05$)

属于研究中的假设！



两个总体均值之差的 t 检验 (计算结果)

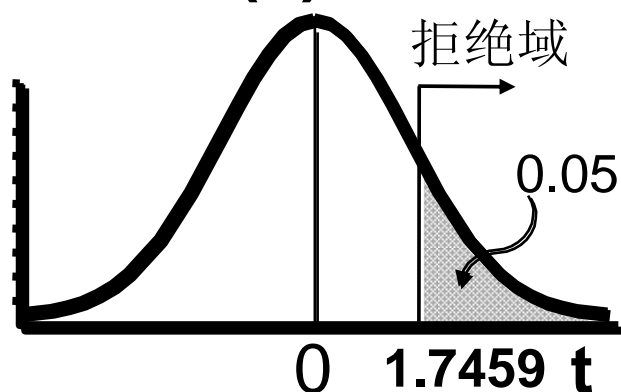
$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

$$\alpha = 0.05$$

$$n_1 = 10, n_2 = 8$$

临界值(s):



7-88

检验统计量:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{261 - 176 - 0}{1137 \sqrt{\frac{1}{10} + \frac{1}{8}}} = 1.576$$

决策:

接受 H_0

结论:

没有证据表明用第二种方法组装更好

假设检验中相关样本的利用

两个相关（配对或匹配）样本 的均值检验

两个总体均值之差的检验 (配对样本的 t 检验)

1. 检验两个相关总体的均值
 - 配对或匹配
 - 重复测量 (前/后)
2. 利用相关样本可消除项目间的方差
3. 假定条件
 - 两个总体都服从正态分布
 - 如果不服从正态分布, 可用正态分布来近似 ($n_1 \geq 30, n_2 \geq 30$)

配对样本的 t 检验 (假设的形式)

假设	研究的问题		
	没有差异 有差异	总体 ₁ ≥ 总体 ₂ 总体 ₁ < 总体 ₂	总体 ₁ ≥ 总体 ₂ 总体 ₁ > 总体 ₂
H₀	$\mu_D = 0$	$\mu_D \geq 0$	$\mu_D \leq 0$
H₁	$\mu_D \neq 0$	$\mu_D < 0$	$\mu_D > 0$

注： $D_i = X_{1i} - X_{2i}$ ， 对第 i 对观察值

配对样本的 t 检验 (数据形式)

观察序号	样本1	样本2	差值
1	x_{11}	x_{21}	$D_1 = x_{11} - x_{21}$
2	x_{12}	x_{22}	$D_1 = x_{12} - x_{22}$
\vdots	\vdots	\vdots	\vdots
i	x_{1i}	x_{2i}	$D_1 = x_{1i} - x_{2i}$
\vdots	\vdots	\vdots	\vdots
n	x_{1n}	x_{2n}	$D_1 = x_{1n} - x_{2n}$

配对样本的 t 检验 (检验统计量)

统计量

$$t = \frac{\bar{x}_D - D_0}{s_D / \sqrt{n_D}}$$

自由度 $df = n_D - 1$

样本均值

$$\bar{x}_D = \frac{\sum_{i=1}^n D_i}{n_D}$$

样本标准差

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{x}_D)^2}{n_D - 1}}$$

配对样本的 t 检验 (例子)

【例】一个以减肥为主要目标的健美俱乐部声称，参加其训练班至少可以使减肥者平均体重减重**8.5**公斤以上。为了验证该宣称是否可信，调查人员随机抽取了**10**名参加者，得到他们的体重记录如下表：

训练前	94.5	101	110	103.5	97	88.5	96.5	101	104	116.5
训练后	85	89.5	101.5	96	86	80.5	87	93.5	93	102

在 $\alpha = 0.05$ 的显著性水平下，调查结果是否支持该俱乐部的声称？

属于检验某项
声明的假设！

配对样本的 t 检验 (计算表)

样本差值计算表		
训练前	训练后	差值 D_i
94.5	85	9.5
101	89.5	11.5
110	101.5	8.5
103.5	96	7.5
97	86	11
88.5	80.5	8
96.5	87	9.5
101	93.5	7.5
104	93	11
116.5	102	14.5
合计	—	98.5

配对样本的 t 检验 (计算结果)

样本均值

$$\bar{x}_D = \frac{\sum_{i=1}^n D_i}{n_D} = \frac{98.5}{10} = 9.85$$

样本标准差

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{x}_D)^2}{n_D - 1}} = \sqrt{\frac{43.525}{10-1}} = 2.199$$

配对样本的 t 检验 (计算结果)

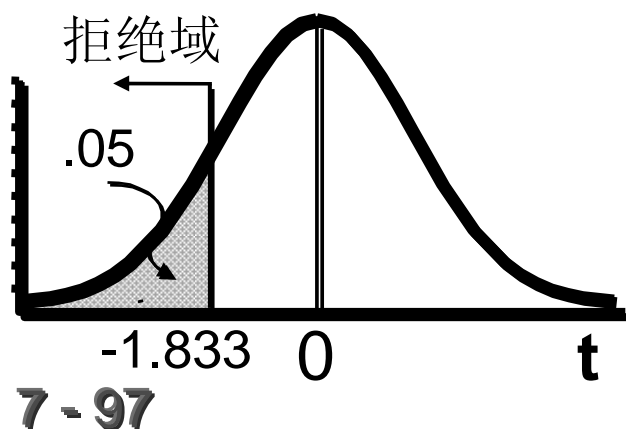
$$H_0: \mu_1 - \mu_2 \geq 8.5$$

$$H_1: \mu_1 - \mu_2 < 8.5$$

$$\alpha = 0.05$$

$$df = 10 - 1 = 9$$

临界值(s):



检验统计量:

$$t = \frac{\bar{x}_D - D_0}{s_D / \sqrt{n_D}} = \frac{9.85 - 0}{2.199 / \sqrt{10}} = 14.165$$

决策:

接受 H_0

结论:

有证据表明该俱乐部的宣称是可信的

两个总体比例之差的检验 (Z 检验)

两个总体比例之差的Z检验

1. 假定条件

- 两个总体是独立的
- 两个总体都服从二项分布
- 可以用正态分布来近似

2. 检验统计量

$$z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{\hat{P}_1(1 - \hat{P}_1)}{n_1} + \frac{\hat{P}_2(1 - \hat{P}_2)}{n_2}}} \sim N(0,1)$$

两个总体比例之差的检验 (假设的形式)

假设	研究的问题		
	没有差异 有差异	比例 ₁ ≥ 比例 ₂ 比例 ₁ < 比例 ₂	总体 ₁ ≤ 比例 ₂ 总体 ₁ > 比例 ₂
H ₀	$P_1 - P_2 = 0$	$P_1 - P_2 \geq 0$	$P_1 - P_2 \leq 0$
H ₁	$P_1 - P_2 \neq 0$	$P_1 - P_2 < 0$	$P_1 - P_2 > 0$

两个总体比例之差的Z检验 (例子)

【例】对两个大型企业青年工人参加技术培训的情况进行调查，调查结果如下：甲厂：调查**60**人，**18**人参加技术培训。乙厂调查**40**人，**14**人参加技术培训。能否根据以上调查结果认为乙厂工人参加技术培训的人数比例高于甲厂？($\alpha = 0.05$)



两个总体比例之差的Z检验 (计算结果)

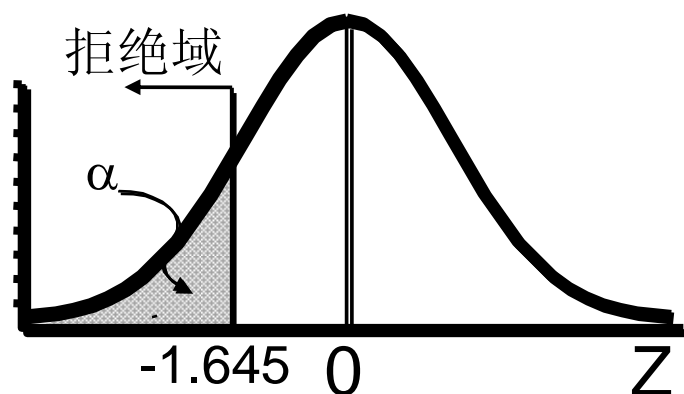
$$H_0: P_1 - P_2 \geq 0$$

$$H_1: P_1 - P_2 < 0$$

$$\alpha = 0.05$$

$$n_1 = 60, n_2 = 40$$

临界值(s):



7-102

检验统计量:

$$z = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} = \frac{0.30 - 0.35 - 0}{\sqrt{\frac{0.30(1-0.30)}{60} + \frac{0.35(1-0.35)}{40}}} = -0.52$$

决策:

接受 H_0

结论:

没有证据表明乙厂工人参加技术培训的人数比例高于甲厂

第四节 假设检验中的其他问题

- 一. 用置信区间进行检验
- 二. 利用 **P** - 值进行检验

利用置信区间进行假设检验

利用置信区间进行假设检验 (双侧检验)

1. 求出双侧检验均值的置信区间

$$\sigma^2 \text{已知时: } \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$$\sigma^2 \text{未知时: } \left(\bar{x} - t_{\alpha/2} \frac{S_{n-1}}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{S_{n-1}}{\sqrt{n}} \right)$$

2. 若总体的假设值 μ_0 在置信区间外, 拒绝 H_0

利用置信区间进行假设检验 (左侧检验)

1. 求出单边置信下限

$$\bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \text{ 或 } \bar{x} - t_{\alpha} \frac{S_{n-1}}{\sqrt{n}}$$

2. 若总体的假设值 μ_0 小于单边置信下限, 拒绝 H_0

利用置信区间进行假设检验 (右侧检验)

1. 求出单边置信上限

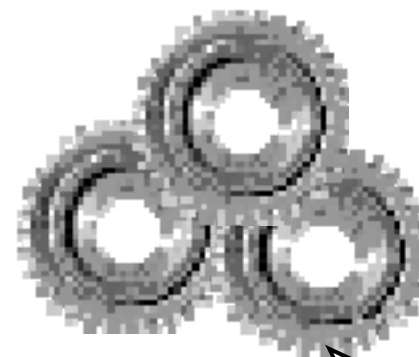
$$\bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \text{ 或 } \bar{x} + t_{\alpha} \frac{s_{n-1}}{\sqrt{n}}$$

2. 若总体的假设值 μ_0 大于单边置信上限，拒绝 H_0

利用置信区间进行假设检验 (例子)

【例】一种袋装食品每包的标准重量应为**1000**克。现从生产的一批产品中随机抽取**16**袋，测得其平均重量为**991**克。已知这种产品重量服从标准差为**50**克的正态分布。试确定这批产品的包装重量是否合格？($\alpha = 0.05$)

属于决策的
假设！



香脆
蛋卷

利用置信区间进行假设检验 (计算结果)

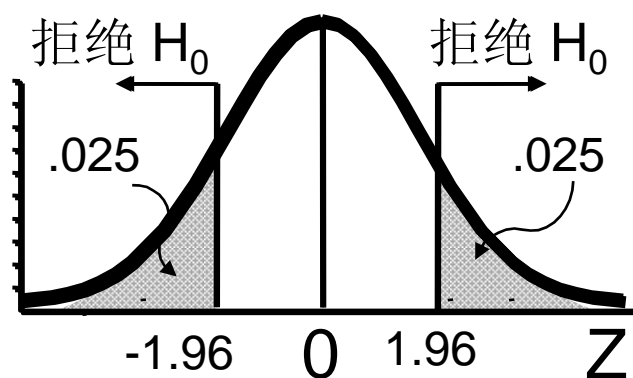
$$H_0: \mu = 1000$$

$$H_1: \mu \neq 1000$$

$$\alpha = 0.05$$

$$n = 49$$

临界值(s):



7-109

置信区间为

$$\begin{aligned} & \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \\ &= \left(991 - 1.96 \frac{50}{\sqrt{16}}, 991 + 1.96 \frac{50}{\sqrt{16}} \right) \\ &= (966.5, 1015.5) \end{aligned}$$

决策:

假设的 $\mu_0 = 1000$ 在置信区间内, 接受 H_0

结论:

表明这批产品的包装重量合格

观察到的显著性水平 P -值

利用 P -值进行假设检验

什么是 P 值? (P -Value)

1. 是一个概率值
2. 如果我们假设原假设为真, P -值是观测到的样本均值不同于(<或 >) 实测值的概率
 - 左侧检验时, P -值为曲线上方小于等于检验统计量部分的面积
 - 右侧检验时, P -值为曲线上方大于等于检验统计量部分的面积
3. 被称为观察到的(或实测的)显著性水平
 - H_0 能被拒绝的 α 的最小值

利用 P 值进行决策

1. 单侧检验

- 若 p -值 $\geq \alpha$, 不能拒绝 H_0
- 若 p -值 $< \alpha$, 拒绝 H_0

2. 双侧检验

- 若 p -值 $\geq \alpha/2$, 不能拒绝 H_0
- 若 p -值 $< \alpha/2$, 拒绝 H_0

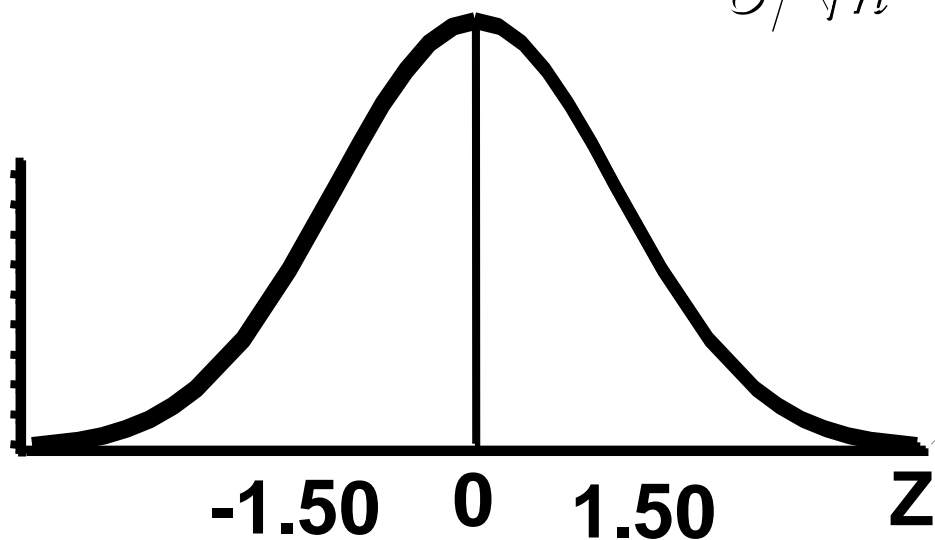
双尾 Z 检验 (P -值计算实例)

【例】欣欣儿童食品厂生产的盒装儿童食品每盒的标准重量为368克。现从某天生产的一批食品中随机抽取25盒进行检查，测得每盒的平均重量为 $\bar{x} = 372.5$ 克。企业规定每盒重量的标准差 σ 为15克。确定 P -值。



双尾 Z 检验 (*P*-值计算结果)

计算的检验统计量为：
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{372.5 - 368}{15 / \sqrt{25}} = 1.5$$

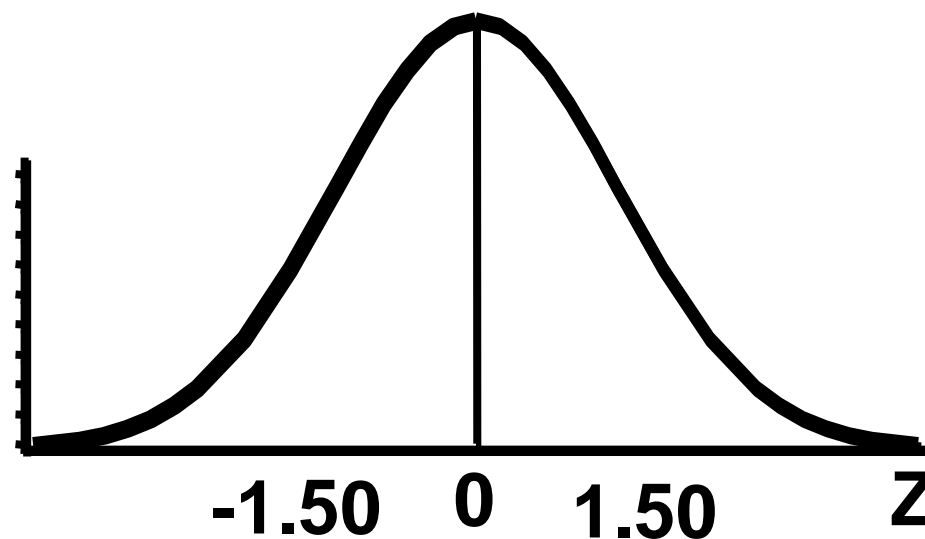


① 样本统计量的Z值

(观察到的)

双尾 Z 检验 (*P*-值计算结果)

p-值为 $P(Z \leq -1.50 \text{ 或 } Z \geq 1.50)$

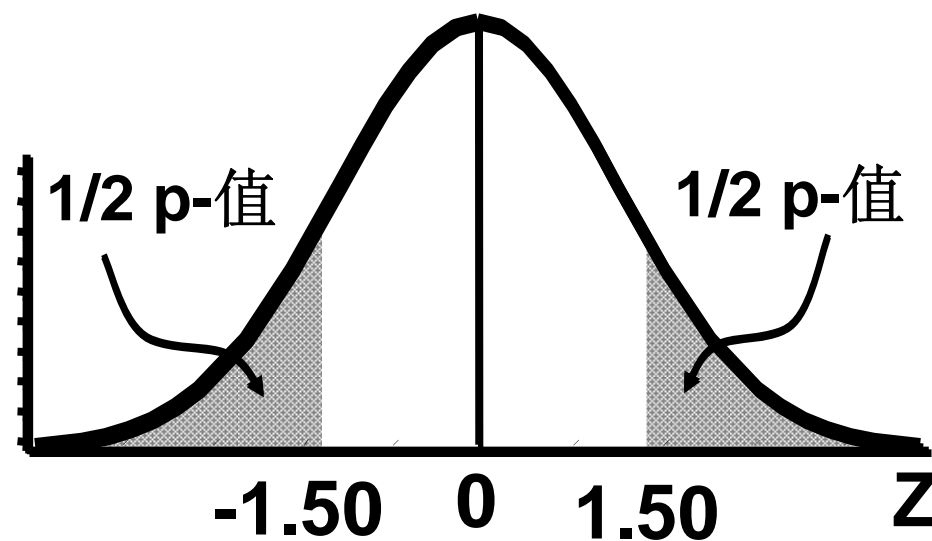


① 样本统计量的Z值

(观察到的)

双尾 Z 检验 (*P*-值计算结果)

p-值为 $P(Z \leq -1.50 \text{ 或 } Z \geq 1.50)$

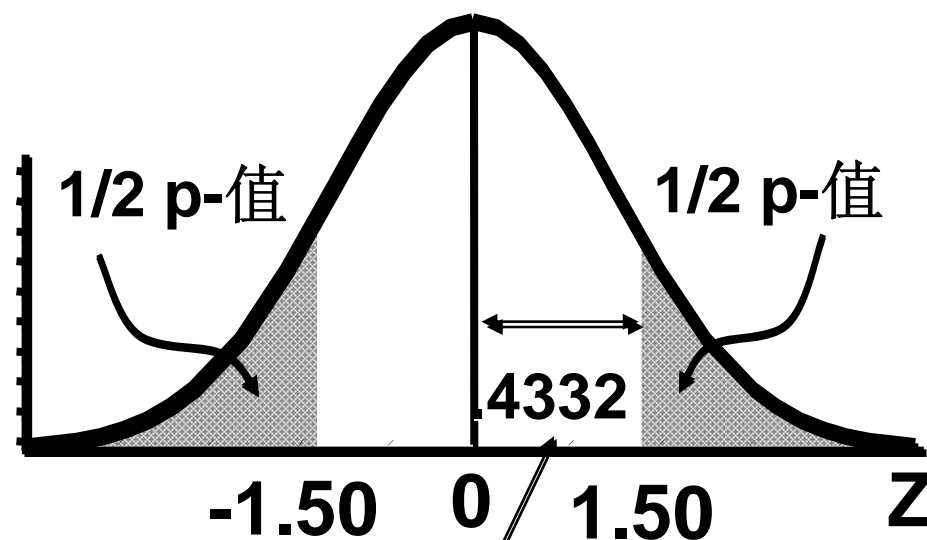


① 样本统计量的 Z 值

(观察到的)

双尾 Z 检验 (*P*-值计算结果)

***p*-值为 $P(Z \leq -1.50 \text{ 或 } Z \geq 1.50)$**



注: $0.9332 - 0.5$
 $= 0.4332$

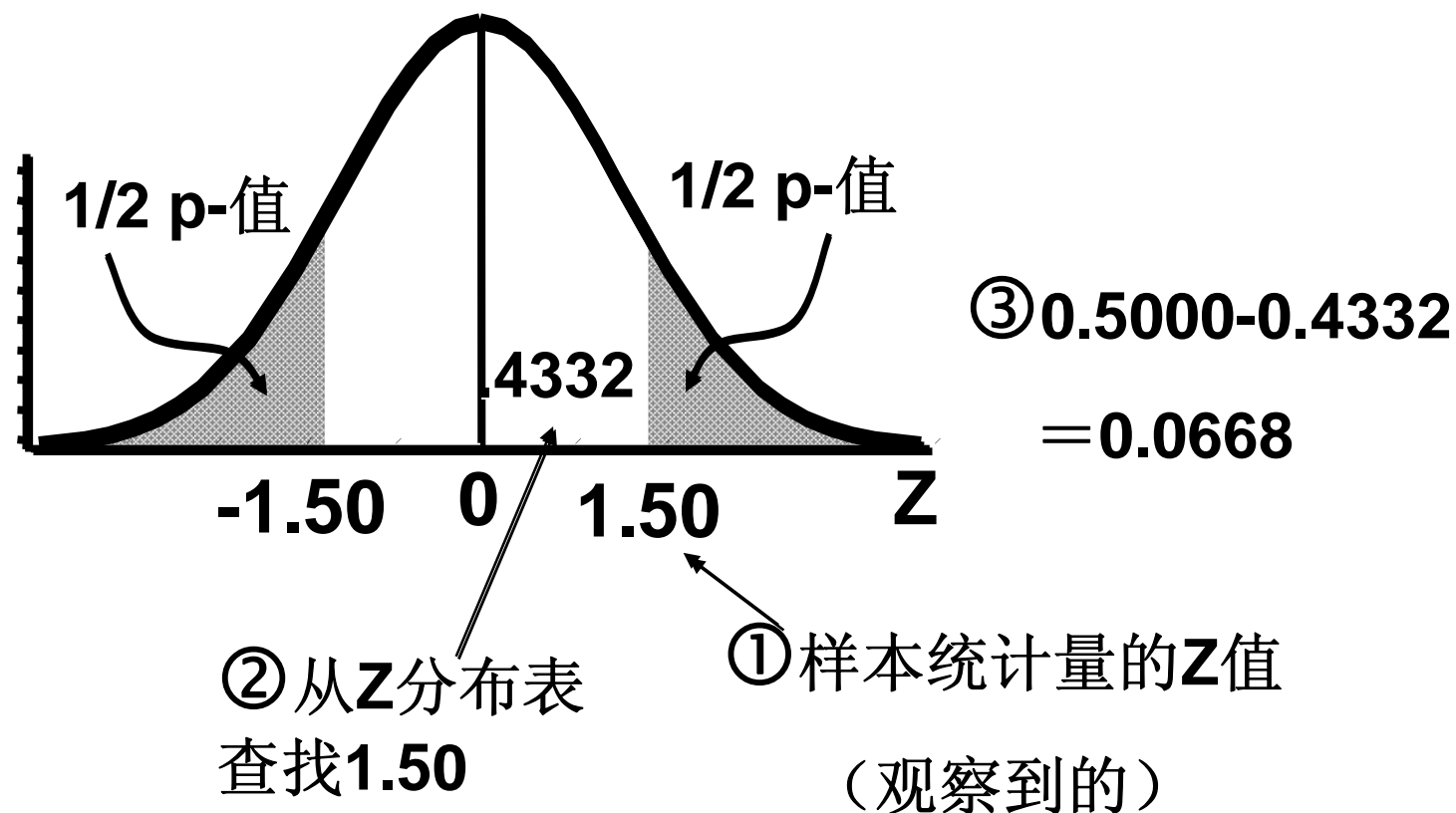
7-117

②从Z分布表
查找1.50

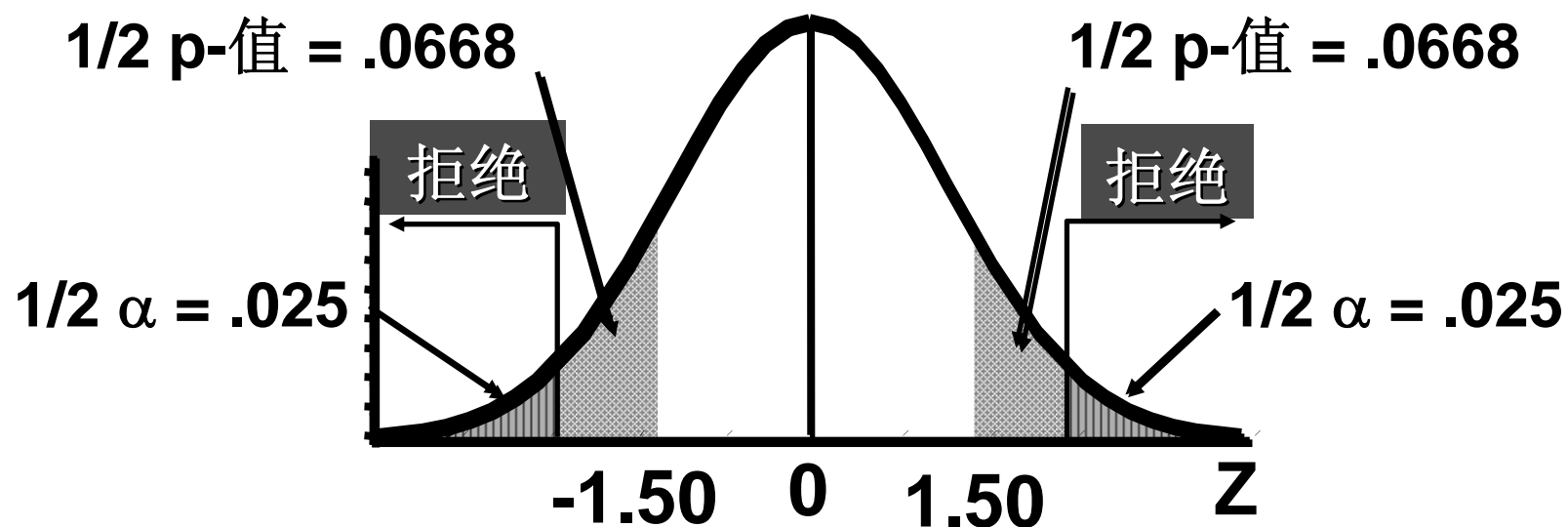
①样本统计量的Z值
(观察到的)

双尾 Z 检验 (*P*-值计算结果)

p-值为 $P(Z \leq -1.50 \text{ 或 } Z \geq 1.50)$

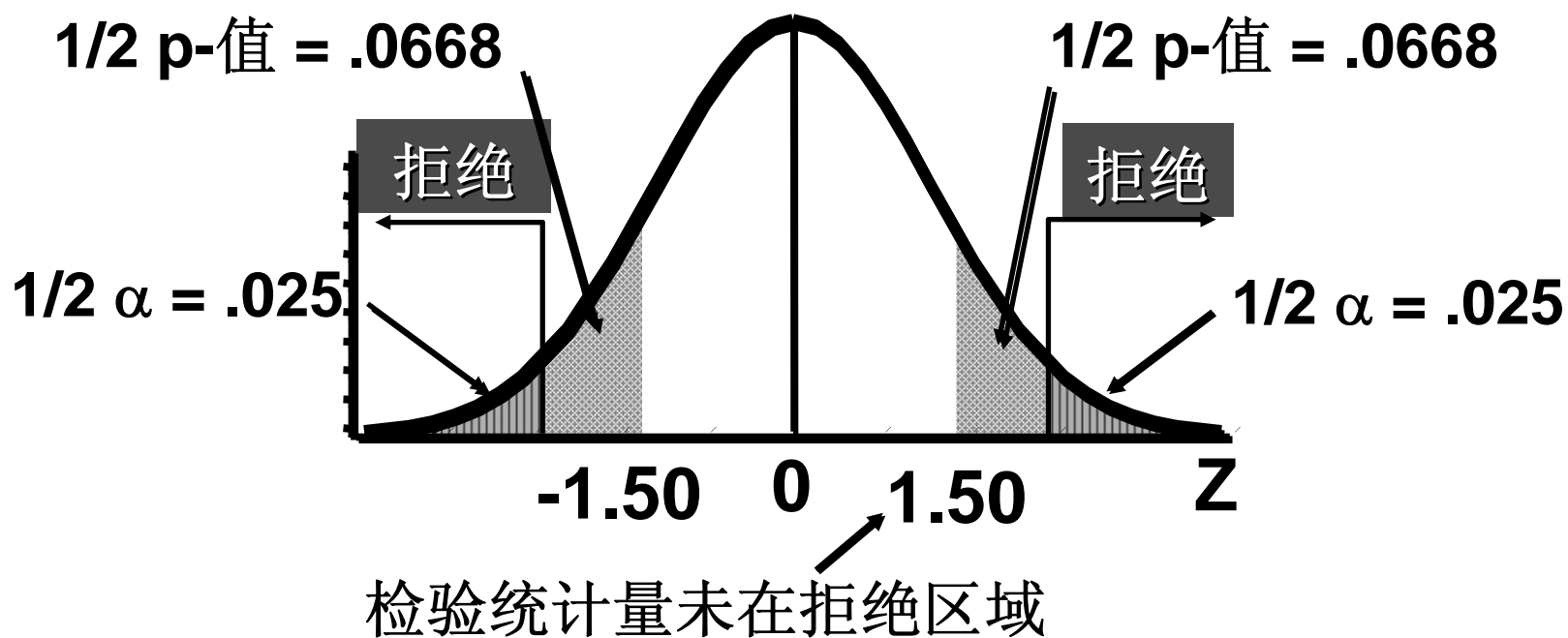


双尾 Z 检验 (P -值计算结果)



双尾 Z 检验 (*P*-值计算结果)

$2p = 0.1336 > \alpha = 0.05$, 不能拒绝 H_0



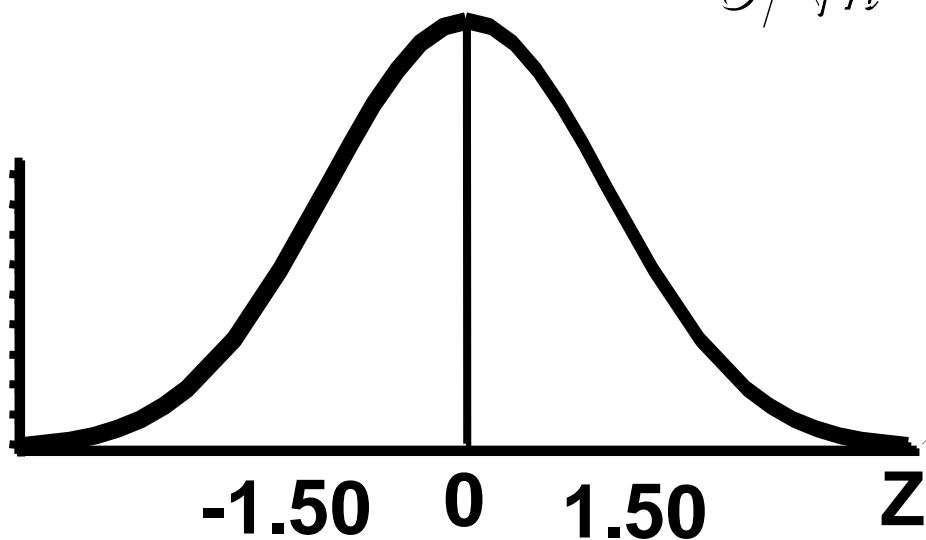
单尾 Z 检验 (P -值计算结果)

【例】欣欣儿童食品厂生产的某种盒装儿童食品，规定每盒的重量不低于**368**克。现从某天生产的一批食品中随机抽取**25**盒进行检查，测得每盒的平均重量为 $\bar{x}=372.5$ 克。企业规定每盒重量的标准差 σ 为**15**克。确定 **P -值**。



单尾 Z 检验 (P-值计算结果)

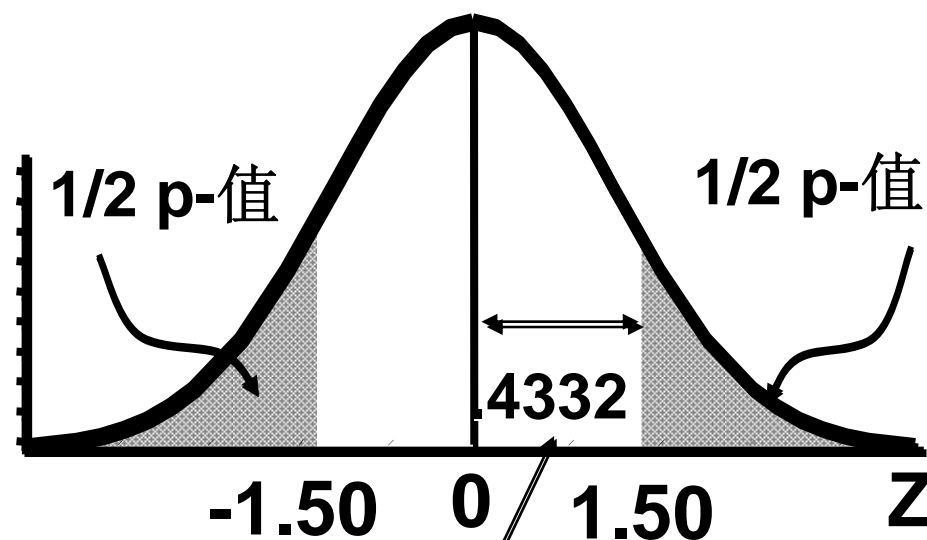
计算的检验统计量为：
$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{372.5 - 368}{15 / \sqrt{25}} = 1.5$$



① 样本统计量的Z值

双尾 Z 检验 (*P*-值计算结果)

p-值为 $P(Z \leq -1.50 \text{ 或 } Z \geq 1.50)$



注: $0.9332 - 0.5$
 $= 0.4332$

7 - 123

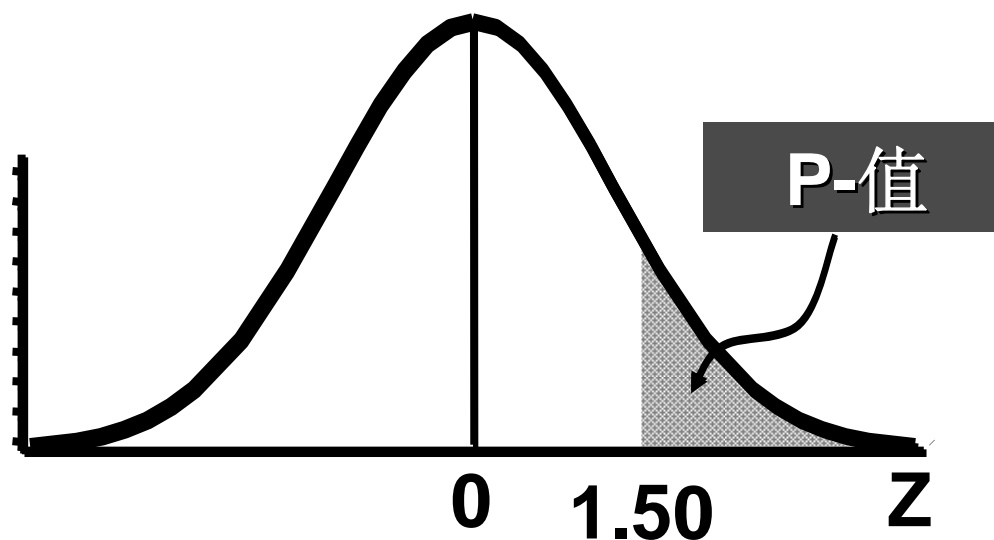
②从Z分布表
查找1.50

①样本统计量的Z值
(观察到的)

单尾 Z 检验 (*P*-值计算结果)

p-值为 $P(Z \geq 1.50)$

②用备择假
设找出方向

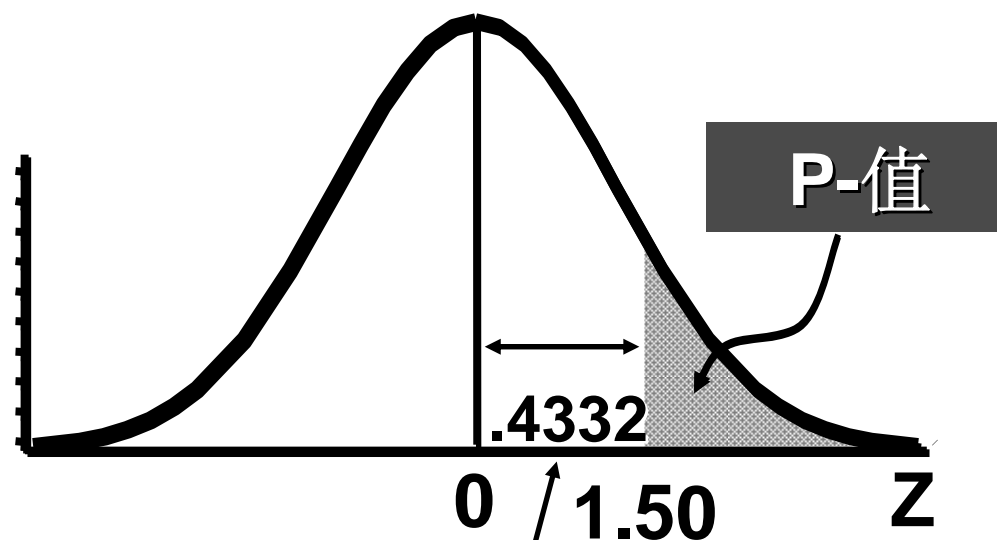


①样本统计量的 Z 值

单尾 Z 检验 (P-值计算结果)

p-值为 $P(Z \geq 1.50)$

②用备择假设找出方向



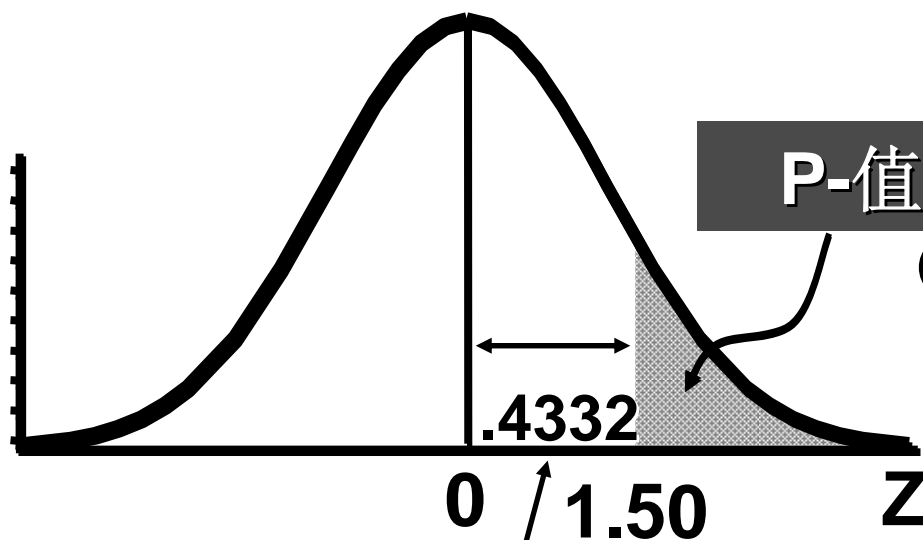
③从Z分布表
:查找1.50

①样本统计量的Z值

单尾 Z 检验 (P-值计算结果)

p-值为 $P(Z \geq 1.50)$

②用备择假设找出方向



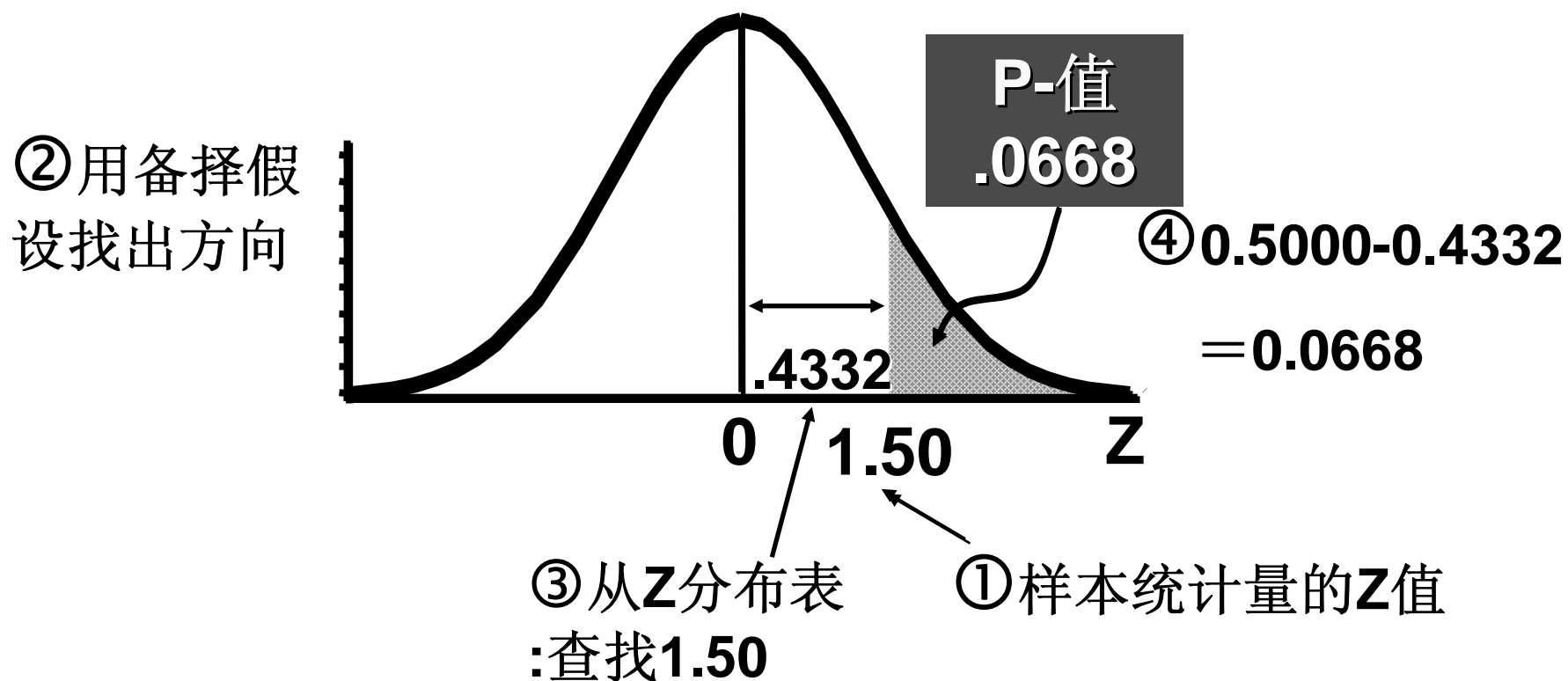
④ $0.5000 - 0.4332$
 $= 0.0668$

③从Z分布表
:查找1.50

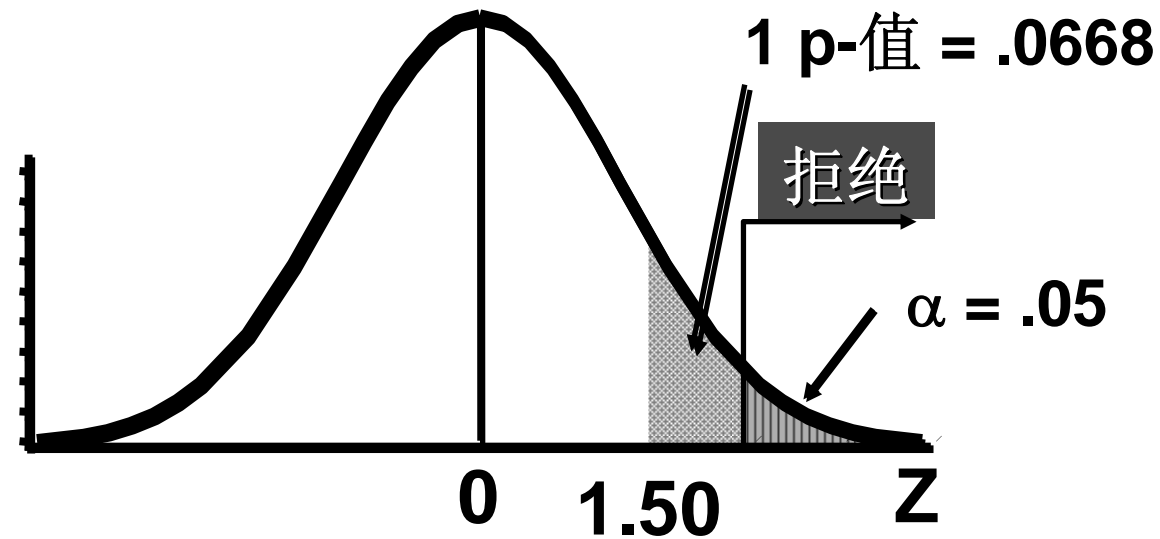
①样本统计量的Z值

单尾 Z 检验 (P-值计算结果)

p-值为 $P(Z \geq 1.50) = .0668$

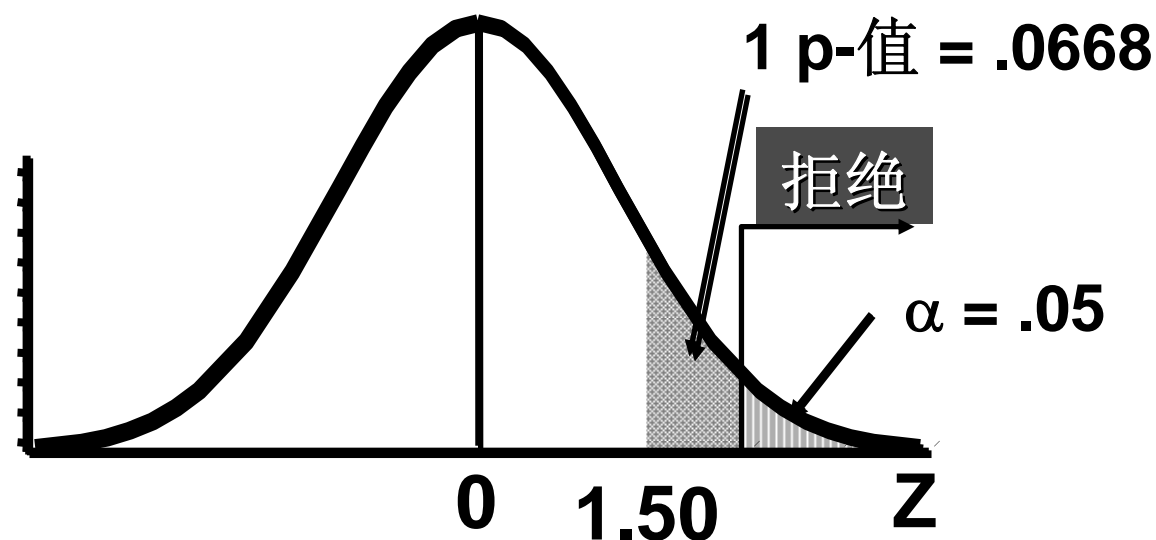


单尾 Z 检验 (P -值计算结果)



单尾 Z 检验 (P -值计算结果)

$(p\text{-值} = 0.0668) \geq (\alpha = .05)$, 不能拒绝 H_0



检验统计量未在拒绝区域

1. 假设检验的概念和类型
2. 假设检验的过程
3. 基于一个样本的假设检验问题
4. 基于两个样本的假设检验问题
5. 利用置信区间进行假设检验
6. 利用 p -值进行假设检验

结 束



第八章 方差分析

PowerPoint



第八章 方差分析

第一节 方差分析的基本问题

第二节 单因素方差分析

第三节 双因素方差分析

学习目标

1. 解释方差分析的概念
2. 解释方差分析的基本思想和原理
2. 掌握单因素方差分析的方法及应用
3. 掌握双因素方差分析的方法及应用

第一节 方差分析的基本问题

- 一. 方差分析的内容
- 二. 方差分析的原理
- 三. F 分布

什么是方差分析？

什么是方差分析? (概念要点)

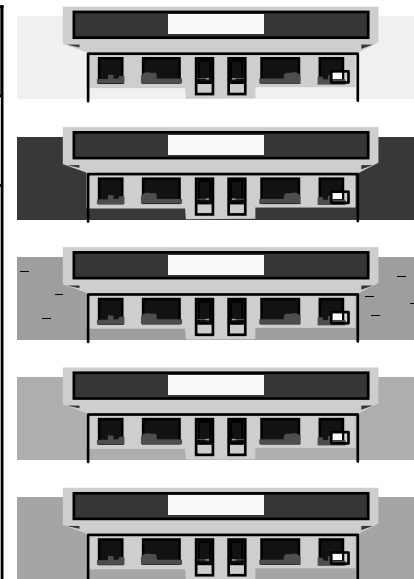
1. 检验多个总体均值是否相等
 - 通过对各观察数据误差来源的分析来判断多个总体均值是否相等
2. 变量
 - 一个定类尺度的自变量
 - 2个或多个 (k 个) 处理水平或分类
 - 一个定距或比例尺度的因变量
3. 用于分析完全随机化试验设计

什么是方差分析？ (一个例子)

【例8.1】 某饮料生产企业研制出一种新型饮料。饮料的颜色共有四种，分别为橘黄色、粉色、绿色和无色透明。这四种饮料的营养含量、味道、价格、包装等可能影响销售量的因素全部相同。现从地理位置相似、经营规模相仿的五家超级市场上收集了前一时期的销售情况，见表8-1。试分析饮料的颜色是否对销售量产生影响。

表8-1 该饮料在五家超市的销售情况

超市	无色	粉色	橘黄色	绿色
1	26.5	31.2	27.9	30.8
2	28.7	28.3	25.1	29.6
3	25.1	30.8	28.5	32.4
4	29.1	27.9	24.2	31.7
5	27.2	29.6	26.5	32.8



什么是方差分析？

（例子的进一步分析）

1. 检验饮料的颜色对销售量是否有影响，也就是检验四种颜色饮料的平均销售量是否相同
2. 设 μ_1 为无色饮料的平均销售量， μ_2 为粉色饮料的平均销售量， μ_3 为橘黄色饮料的平均销售量， μ_4 为绿色饮料的平均销售量，也就是检验下面的假设
 - $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
 - $H_1: \mu_1, \mu_2, \mu_3, \mu_4$ 不全相等
3. 检验上述假设所采用的方法就是方差分析

经济、管理类
基础课程

统计学

方差分析的基本思想和原理

方差分析的基本思想和原理 (几个基本概念)

1. 因素或因子

- 所要检验的对象称为因子
- 要分析饮料的颜色对销售量是否有影响，颜色是要检验的因素或因子

2. 水平

- 因素的具体表现称为水平
- A_1 、 A_2 、 A_3 、 A_4 四种颜色就是因素的水平

3. 观察值

- 在每个因素水平下得到的样本值
- 每种颜色饮料的销售量就是观察值

方差分析的基本思想和原理 (几个基本概念)

1. 试验

- 这里只涉及一个因素，因此称为单因素四水平的试验

2. 总体

- 因素的每一个水平可以看作是一个总体
- 比如 A_1 、 A_2 、 A_3 、 A_4 四种颜色可以看作是四个总体

3. 样本数据

- 上面的数据可以看作是从这四个总体中抽取的样本数据

方差分析的基本思想和原理

1. 比较两类误差，以检验均值是否相等
2. 比较的基础是方差比
3. 如果系统(处理)误差显著地不同于随机误差，则均值就是不相等的；反之，均值就是相等的
4. 误差是由各部分的误差占总误差的比例来测度的

方差分析的基本思想和原理 (两类误差)

1. 随机误差

- 在因素的同一水平(同一个总体)下，样本的各观察值之间的差异
- 比如，同一种颜色的饮料在不同超市上的销售量是不同的
- 不同超市销售量的差异可以看成是随机因素的影响，或者说是由于抽样的随机性所造成的，称为*随机误差*

2. 系统误差

- 在因素的不同水平(不同总体)下，各观察值之间的差异
- 比如，同一家超市，不同颜色饮料的销售量也是不同的
- 这种差异*可能*是由于抽样的随机性所造成的，*也可能*是由于颜色本身所造成的，后者所形成的误差是由系统性因素造成的，称为*系统误差*

方差分析的基本思想和原理 (两类方差)

1. 组内方差

- 因素的同一水平(同一个总体)下样本数据的方差
- 比如, 无色饮料 A_1 在5家超市销售数量的方差
- 组内方差只包含*随机误差*

2. 组间方差

- 因素的不同水平(不同总体)下各样本之间的方差
- 比如, A_1 、 A_2 、 A_3 、 A_4 四种颜色饮料销售量之间的方差
- 组间方差既包括*随机误差*, 也包括*系统误差*

方差分析的基本思想和原理 (方差的比较)

1. 如果不同颜色(水平)对销售量(结果)没有影响, 那么在组间方差中只包含有随机误差, 而没有系统误差。这时, 组间方差与组内方差就应该很接近, 两个方差的比值就会接近1
2. 如果不同的水平对结果有影响, 在组间方差中除了包含随机误差外, 还会包含有系统误差, 这时组间方差就会大于组内方差, 组间方差与组内方差的比值就会大于1
3. 当这个比值大到某种程度时, 就可以说不同水平之间存在着显著差异

方差分析中的基本假定

方差分析中的基本假定

1. 每个总体都应服从正态分布
 - 对于因素的每一个水平，其观察值是来自服从正态分布总体的简单随机样本
 - 比如，每种颜色饮料的销售量必需服从正态分布
2. 各个总体的方差必须相同
 - 对于各组观察数据，是从具有相同方差的总体中抽取的
 - 比如，四种颜色饮料的销售量的方差都相同
3. 观察值是独立的
 - 比如，每个超市的销售量都与其他超市的销售量独立

方差分析中的基本假定

1. 在上述假定条件下，判断颜色对销售量是否有显著影响，实际上也就是检验具有同方差的四个正态总体的均值是否相等的问题
2. 如果四个总体的均值相等，可以期望四个样本的均值也会很接近
 - 四个样本的均值越接近，我们推断四个总体均值相等的证据也就越充分
 - 样本均值越不同，我们推断总体均值不同的证据就越充分

方差分析中基本假定

➡ 如果原假设成立，即 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

- 四种颜色饮料销售的均值都相等
- 没有系统误差

这意味着每个样本都来自均值为 μ 、差为 σ^2 的同一正态总体

$f(X)$

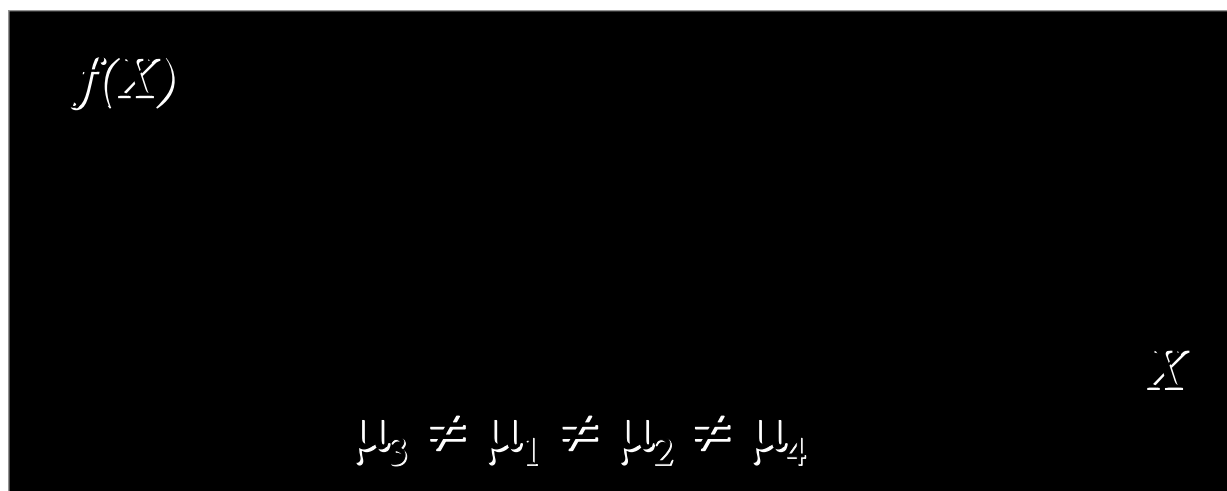
X

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

方差分析中基本假定

- ➡ 如果备择假设成立，即 $H_1: \mu_i (i=1, 2, 3, 4)$ 不全相等
- 至少有一个总体的均值是不同的
 - 有系统误差

这意味着四个样本分别来自均值不同的四个正态总体



第二节 单因素方差分析

- 一. 单因素方差分析的步骤
- 二. 方差分析中的多重比较
- 三. 单因素方差分析中的其他问题

单因素方差分析的数据结构

观察值 (j)	因素(A) i			
	水平 A_1	水平 A_2	...	水平 A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
:	:	:	:	:
:	:	:	:	:
n	x_{n1}	x_{n2}	...	x_{nk}

单因素方差分析的步骤

- 提出假设
- 构造检验统计量
- 统计决策

提出假设

1. 一般提法

- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (因素有 k 个水平)
- $H_1: \mu_1, \mu_2, \dots, \mu_k$ 不全相等

2. 对前面的例子

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
 - 颜色对销售量没有影响
- $H_1: \mu_1, \mu_2, \mu_3, \mu_4$ 不全相等
 - 颜色对销售量有影响

构造检验的统计量

1. 为检验 H_0 是否成立，需确定检验的统计量
2. 构造统计量需要计算
 - 水平的均值
 - 全部观察值的总均值
 - 离差平方和
 - 均方(MS)

构造检验的统计量 (计算水平的均值)

1. 假定从第 i 个总体中抽取一个容量为 n_i 的简单随机样本，第 i 个总体的样本均值为该样本的全部观察值总和除以观察值的个数
2. 计算公式为

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad (i = 1, 2, \dots, k)$$

式中： n_i 为第 i 个总体的样本观察值个数
 x_{ij} 为第 i 个总体的第 j 个观察值

构造检验的统计量 (计算全部观察值的总均值)

1. 全部观察值的总和除以观察值的总个数
2. 计算公式为

$$\bar{x} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}}{n} = \frac{\sum_{i=1}^k n_i \bar{x}_i}{n}$$

式中： $n = n_1 + n_2 + \cdots + n_k$

构造检验的统计量 (前例计算结果)

表8-2 四种颜色饮料的销售量及均值

超市 (j)	水平A (i)				
	无色(A_1)	粉色(A_2)	橘黄色(A_3)	绿色(A_4)	
1	26.5	31.2	27.9	30.8	
2	28.7	28.3	25.1	29.6	
3	25.1	30.8	28.5	32.4	
4	29.1	27.9	24.2	31.7	
5	27.2	29.6	26.5	32.8	
合计	136.6	147.8	132.2	157.3	573.9
水平均值	$\bar{x}_1=27.32$	$\bar{x}_2=29.56$	$\bar{x}_3=26.44$	$\bar{x}_4=31.46$	总均值
观察值个数	$n_1=5$	$n_2=5$	$n_3=5$	$n_4=5$	$\bar{x}=28.695$

构造检验的统计量 (计算总离差平方和 SST)

1. 全部观察值 x_{ij} 与总平均值 $\bar{\bar{x}}$ 的离差平方和
2. 反映全部观察值的离散状况
3. 其计算公式为

$$SST = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2$$

- 前例的计算结果:

$$\begin{aligned} SST &= (26.5-28.695)^2 + (28.7-28.695)^2 + \dots + (32.8-28.695)^2 \\ &= 115.9295 \end{aligned}$$

构造检验的统计量 (计算误差项平方和 SSE)

1. 每个水平或组的各样本数据与其组平均值的离差平方和
2. 反映每个样本各观察值的离散状况，又称组内离差平方和
3. 该平方和反映的是随机误差的大小
4. 计算公式为

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

- 前例的计算结果： $SSE = 39.084$

构造检验的统计量 (计算水平项平方和 SSA)

1. 各组平均值 \bar{x}_i ($i = 1, 2, \dots, k$) 与总平均值 $\bar{\bar{x}}$ 的离差平方和
2. 反映各总体的样本均值之间的差异程度，又称组间平方和
3. 该平方和既包括随机误差，也包括系统误差
4. 计算公式为

$$SSA = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2$$

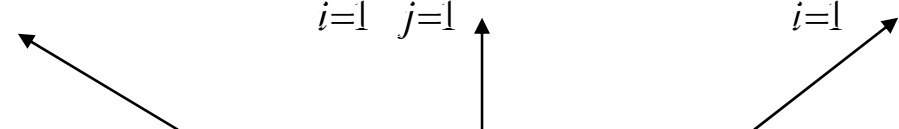
- 前例的计算结果： $SSA = 76.8455$

构造检验的统计量 (三个平方和的关系)

➡ 总离差平方和(SST)、误差项离差平方和(SSE)、水平项离差平方和(SSA)之间的关系

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{\bar{x}})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2 + \sum_{i=1}^k n_i (\bar{x}_i - \bar{\bar{x}})^2$$

$SST = SSE + SSA$



构造检验的统计量 (三个平方和的作用)

1. ***SST***反映了全部数据总的误差程度；***SSE***反映了随机误差的大小；***SSA***反映了随机误差和系统误差的大小
2. 如果原假设成立，即 $H_1 = H_2 = \dots = H_k$ 为真，则表明没有系统误差，组间平方和***SSA***除以自由度后的均方与组内平方和***SSE***和除以自由度后的均方差异就不会太大；如果组间均方显著地大于组内均方，说明各水平(总体)之间的差异不仅有随机误差，还有系统误差
3. 判断因素的水平是否对其观察值有影响，实际上就是比较组间方差与组内方差之间差异的大小
4. 为检验这种差异，需要构造一个用于检验的统计量

构造检验的统计量 (计算均方 MS)

1. 各离差平方和的大小与观察值的多少有关，为了消除观察值多少对离差平方和大小的影响，需要将其平均，这就是均方，也称为方差
2. 计算方法是用离差平方和除以相应的自由度
3. 三个平方和的自由度分别是
 - SST 的自由度为 $n-1$ ，其中 n 为全部观察值的个数
 - SSA 的自由度为 $k-1$ ，其中 k 为因素水平(总体)的个数
 - SSE 的自由度为 $n-k$

构造检验的统计量 (计算均方 MS)

1. SSA 的均方也称组间方差，记为 MSA ，计算公式为

$$MSA = \frac{SSA}{k-1} \quad \text{前例的计算结果: } MSA = \frac{76.8455}{4-1} = 25.6152$$

2. SSE 的均方也称组内方差，记为 MSE ，计算公式为

$$MSE = \frac{SSE}{n-k} \quad \text{前例的计算结果: } MSE = \frac{39.084}{20-4} = 2.4428$$

构造检验的统计量 (计算检验的统计量 F)

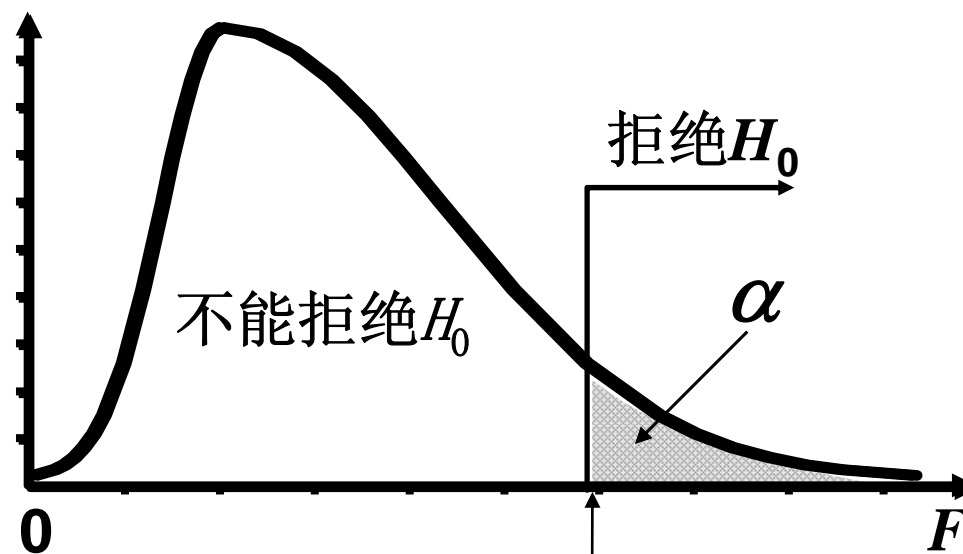
1. 将 MSA 和 MSE 进行对比，即得到所需要的检验统计量 F
2. 当 H_0 为真时，二者的比值服从分子自由度为 $k-1$ 、分母自由度为 $n-k$ 的 F 分布，即

$$F = \frac{MSA}{MSE} \sim F(k-1, n-k)$$

$$\text{前例的计算结果: } F = \frac{25.6152}{2.4428} = 10.486$$

构造检验的统计量 (F 分布与拒绝域)

如果均值相等,
 $F = MSA/MSE \rightarrow 1$



$F_{\alpha}(k-1, n-k)$

F 分布

统计决策

- ➡ 将统计量的值 F 与给定的显著性水平 α 的临界值 F_{α} 进行比较，作出接受或拒绝原假设 H_0 的决策
- 根据给定的显著性水平 α ，在 F 分布表中查找与第一自由度 $df_1=k-1$ 、第二自由度 $df_2=n-k$ 相应的临界值 F_{α}
 - 若 $F > F_{\alpha}$ ，则拒绝原假设 H_0 ，表明均值之间的差异是显著的，所检验的因素(A)对观察值有显著影响
 - 若 $F \leq F_{\alpha}$ ，则不能拒绝原假设 H_0 ，表明所检验的因素(A)对观察值没有显著影响

单因素方差分析表 (基本结构)

方差来源	平方和 <i>SS</i>	自由度 <i>df</i>	均方 <i>MS</i>	<i>F</i> 值
组间(因素影响)	<i>SSA</i>	<i>k-1</i>	<i>MSA</i>	$\frac{MSA}{MSE}$
组内(误差)	<i>SSE</i>	<i>n-k</i>	<i>MSE</i>	
总和	<i>SST</i>	<i>n-1</i>		

单因素方差分析 (Excel 的输出结果)

方差分析：单因素方差分析

SUMMARY

组	计数	求和	平均	方差
列 1	5	136.6	27.32	2.672
列 2	5	147.8	29.56	2.143
列 3	5	132.2	26.44	3.298
列 4	5	157.3	31.46	1.658

方差分析

差异源	SS	df	MS	F	P-value	F crit
组间	76.8455	3	25.615	10.486	0.0005	3.2389
组内	39.084	16	2.4428			
总计	115.93	19				

单因素方差分析 (一个例子)

【例】为了对几个行业的服务质量进行评价，消费者协会在零售业、旅游业、航空公司、家电制造业分别抽取了不同的样本，其中零售业抽取7家，旅游业抽取了6家，航空公司抽取5家、家电制造业抽取了5家，然后记录了一年中消费者对总共23家服务企业投诉的次数，结果如表9.7。试分析这四个行业的服务质量是否有显著差异？($\alpha=0.05$)

单因素方差分析 (一个例子)

消费者对四个行业的投诉次数				
观察值 (j)	行业(A)			
	零售业	旅游业	航空公司	家电制造业
1	57	62	51	70
2	55	49	49	68
3	46	60	48	63
4	45	54	55	69
5	54	56	47	60
6	53	55		
7	47			

单因素方差分析 (计算结果)

解：设四个行业被投诉次数的均值分别为， μ_1 、 μ_2 、 μ_3 、 μ_4 ，则需要检验如下假设

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ (四个行业的服务质量无显著差异)
- $H_1: \mu_1, \mu_2, \mu_3, \mu_4$ 不全相等 (有显著差异)
- Excel输出的结果如下

差异源	SS	自由度	MS	F	P-值	临界值
组间	845.2174	3	281.7391	14.78741	3.31E-05	3.127354
组内	362	19	19.05263			
总和	1207.217	22				

- 结论：拒绝 H_0 。四个行业的服务质量有显著差异

方差分析中的多重比较

方差分析中的多重比较 (作用)

1. 多重比较是通过对总体均值之间的配对比较来进一步检验到底哪些均值之间存在差异
2. 多重比较方法有多种，这里介绍Fisher提出的最小显著差异方法，简写为***LSD***，该方法可用于判断到底哪些均值之间有差异
3. ***LSD***方法是对检验两个总体均值是否相等的 t 检验方法的总体方差估计加以修正(用***MSE***来代替)而得到的

方差分析中的多重比较 (步骤)

1. 提出假设

- $H_0: \mu_i = \mu_j$ (第*i*个总体的均值等于第*j*个总体的均值)
- $H_1: \mu_i \neq \mu_j$ (第*i*个总体的均值不等于第*j*个总体的均值)

2. 检验的统计量为

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t(n - k)$$

3. 若 $|t| \geq t_{\alpha/2}$, 拒绝 H_0 ; 若 $|t| < t_{\alpha/2}$, 不能拒绝 H_0

方差分析中的多重比较

(基于统计量 $\bar{x}_i - \bar{x}_j$ 的 LSD 方法)

1. 通过判断样本均值之差的大小来检验 H_0
2. 检验的统计量为： $\bar{x}_i - \bar{x}_j$
3. 检验的步骤为
 - 提出假设
 - $H_0: \mu_i = \mu_j$ (第*i*个总体的均值等于第*j*个总体的均值)
 - $H_1: \mu_i \neq \mu_j$ (第*i*个总体的均值不等于第*j*个总体的均值)
 - 计算 LSD

$$LSD = t_{\alpha/2} \cdot \sqrt{MSE \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

- 若 $|\bar{x}_i - \bar{x}_j| \geq LSD$, 拒绝 H_0 , 若 $|\bar{x}_i - \bar{x}_j| < LSD$, 不能拒绝 H_0

方差分析中的多重比较 (实例)

1. 根据前面的计算结果: $\bar{x}_1=27.3$; $\bar{x}_2=29.5$;
 $\bar{x}_3=26.4$; $\bar{x}_4=31.4$

2. 提出假设

■ $H_0: \mu_i = \mu_j$; $H_1: \mu_i \neq \mu_j$

3. 计算 LSD

$$LSD = 2.12 \sqrt{2.4428 \left(\frac{1}{5} + \frac{1}{5} \right)} = 2.096$$

方差分析中的多重比较 (实例)

$$|\bar{x}_1 - \bar{x}_2| = |27.3 - 29.5| = 2.2 > 2.096$$

颜色1与颜色2的销售量有显著差异

$$|\bar{x}_1 - \bar{x}_3| = |27.3 - 26.4| = 0.9 < 2.096$$

颜色1与颜色3的销售量没有显著差异

$$|\bar{x}_1 - \bar{x}_4| = |27.3 - 31.4| = 4.1 > 2.096$$

颜色1与颜色4的销售量有显著差异

$$|\bar{x}_2 - \bar{x}_3| = |29.5 - 26.4| = 3.1 > 2.096$$

颜色2与颜色3的销售量有显著差异

$$|\bar{x}_2 - \bar{x}_4| = |29.5 - 31.4| = 1.9 < 2.096$$

颜色2与颜色4的销售量没有显著差异

$$|\bar{x}_3 - \bar{x}_4| = |26.4 - 31.4| = 5 > 2.096$$

颜色3与颜色4的销售量有显著差异

第三节 双因素方差分析

- 一. 双因素方差分析的基本问题
- 二. 双因素方差分析的数据结构
- 三. 双因素方差分析的步骤
- 四. 一个应用实例

双因素方差分析的基本问题

双因素方差分析 (概念要点)

1. 分析两个因素(因素A和因素B)对试验结果的影响
2. 分别对两个因素进行检验，分析是一个因素在起作用，还是两个因素都起作用，还是两个因素都不起作用
3. 如果A和B对试验结果的影响是相互独立的，分别判断因素A和因素B对试验指标的影响，这时的双因素方差分析称为无交互作用的双因素方差分析
4. 如果除了A和B对试验结果的单独影响外，因素A和因素B的搭配还会对销售量产生一种新的影响，这时的双因素方差分析称为有交互作用的双因素方差分析
5. 对于无交互作用的双因素方差分析，其结果与对每个因素分别进行单因素方差分析的结果相同

双因素方差分析的基本假定

1. 每个总体都服从正态分布
 - 对于因素的每一个水平，其观察值是来自正态分布总体的简单随机样本
2. 各个总体的方差必须相同
 - 对于各组观察数据，是从具有相同方差的总体中抽取的
3. 观察值是独立的

双因素方差分析的数据结构

因素A (i)	因素(B)j				平均值 $\bar{x}_{i.}$
	B_1	B_2	...	B_r	
A_1	x_{11}	x_{12}	...	x_{1k}	$\bar{x}_{1.}$
A_2	x_{21}	x_{22}	...	x_{2k}	$\bar{x}_{2.}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_k	x_{r1}	x_{r2}	...	x_{rk}	$\bar{x}_{k.}$
平均值 $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.r}$	$\bar{\bar{x}}$

8 - 54

$$x_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, r)$$

双因素方差分析的数据结构

➡ $\bar{x}_{i.}$ 是因素A的第*i*个水平下各观察值的平均值

$$\bar{x}_{i.} = \frac{\sum_{j=1}^r x_{ij}}{r} \quad (i=1,2,\dots,k)$$

➡ $\bar{x}_{.j}$ 是因素B的第*j*个水平下的各观察值的均值

$$\bar{x}_{.j} = \frac{\sum_{i=1}^k x_{ij}}{k} \quad (j=1,2,\dots,r)$$

➡ $\bar{\bar{x}}$ 是全部 kr 个样本数据的总平均值

$$\bar{\bar{x}} = \frac{\sum_{i=1}^k \sum_{j=1}^r x_{ij}}{kr}$$

双因素方差分析的步骤

提出假设

1. 对因素A提出的假设为

- $H_0: \mu_1 = \mu_2 = \dots = \mu_i = \dots = \mu_k$ (μ_i 为第*i*个水平的均值)
- $H_1: \mu_i (i=1,2, \dots, k)$ 不全相等

2. 对因素B提出的假设为

- $H_0: \mu_1 = \mu_2 = \dots = \mu_j = \dots = \mu_r$ (μ_j 为第*j*个水平的均值)
- $H_1: \mu_j (j=1,2,\dots,r)$ 不全相等

构造检验的统计量

1. 为检验 H_0 是否成立，需确定检验的统计量
2. 构造统计量需要计算
 - 总离差平方和
 - 水平项平方和
 - 误差项平方和
 - 均方

构造检验的统计量 (计算总离差平方和 SST)

1. 全部观察值 x_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, r$) 与总平均值 $\bar{\bar{x}}$ 的离差平方和
2. 反映全部观察值的离散状况
3. 计算公式为

$$SST = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2$$

构造检验的统计量 (计算SSA、SSB和SSE)

1. 因素A的离差平方和SSA

$$SSA = \sum_{i=1}^k \sum_{j=1}^r (\bar{x}_{i.} - \bar{\bar{x}})^2$$

2. 因素B的离差平方和SSB

$$SSB = \sum_{i=1}^k \sum_{j=1}^r (\bar{x}_{.j} - \bar{\bar{x}})^2$$

3. 误差项平方和SSE

$$SSE = \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2$$

构造检验的统计量 (各平方和的关系)

- ➡ 总离差平方和(SST)、水平项离差平方和 (SSA 和 SSB)、误差项离差平方和(SSE) 之间的关系

$$\begin{aligned} & \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{\bar{x}})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^r (\bar{x}_{i.} - \bar{\bar{x}})^2 + \sum_{i=1}^k \sum_{j=1}^r (\bar{x}_{.j} - \bar{\bar{x}})^2 + \sum_{i=1}^k \sum_{j=1}^r (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}})^2 \end{aligned}$$

$$SST = SSA + SSB + SSE$$

构造检验的统计量 (计算均方 MS)

1. 各离差平方和的大小与观察值的多少有关，为消除观察值多少对离差平方和大小的影响，需要将其平均，这就是均方，也称为方差
2. 计算方法是用离差平方和除以相应的自由度
3. 三个平方和的自由度分别是
 - 总离差平方和 SST 的自由度为 $kr-1$
 - 因素 A 的离差平方和 SSA 的自由度为 $k-1$
 - 因素 B 的离差平方和 SSB 的自由度为 $r-1$
 - 随机误差平方和 SSE 的自由度为 $(k-1) \times (r-1)$

构造检验的统计量 (计算均方 MS)

1. 因素A的均方，记为 MSA ，计算公式为

$$MSA = \frac{SSA}{k-1}$$

2. 因素B的均方，记为 MSB ，计算公式为

$$MSB = \frac{SSB}{r-1}$$

3. 随机误差项的均方，记为 MSE ，计算公式为

$$MSE = \frac{SSE}{(k-1)(r-1)}$$

构造检验的统计量 (计算检验的统计量 F)

1. 为检验因素 A 的影响是否显著，采用下面的统计量

$$F'_A = \frac{MSA}{MSE} \sim F(k-1, (k-1)(r-1))$$

2. 为检验因素 B 的影响是否显著，采用下面的统计量

$$F'_B = \frac{MSB}{MSE} \sim F(r-1, (k-1)(r-1))$$

统计决策

- ➡ 将统计量的值 F 与给定的显著性水平 α 的临界值 F_α 进行比较，作出接受或拒绝原假设 H_0 的决策
 - 根据给定的显著性水平 α 在 F 分布表中查找相应的临界值 F_α
 - 若 $F_A \geq F_\alpha$ ，则拒绝原假设 H_0 ，表明均值之间的差异是显著的，即所检验的因素(A)对观察值有显著影响
 - 若 $F_B \geq F_\alpha$ ，则拒绝原假设 H_0 ，表明均值之间有显著差异，即所检验的因素(B)对观察值有显著影响

双因素方差分析表 (基本结构)

方差来源	平方和 SS	自由度 df	均方 MS	F 值
因素A	SSA	$k-1$	MSA	F_A
因素B	SSB	$r-1$	MSB	F_B
误差	SSE	$(k-1) \times (r-1)$	MSE	
总和	SST	$kr-1$		

双因素方差分析 (一个例子)

【例】有四个品牌的彩电在五个地区销售，为分析彩电的品牌(因素A)和销售地区(因素B)对销售量是否有影响，对每个品牌在各地区的销售量取得以下数据，见下表。试分析品牌和销售地区对彩电的销售量是否有显著影响？

不同品牌的彩电在各地区的销售量数据

品牌 (因素A)	销售地区(因素B)				
	B_1	B_2	B_3	B_4	B_5
A_1	365	350	343	340	323
A_2	345	368	363	330	333
A_3	358	323	353	343	308
A_4	288	280	298	260	298

双因素方差分析 (提出假设)

1. 对因素A提出的假设为

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
(品牌对销售量没有影响)
- $H_1: \mu_i (i = 1, 2, \dots, 4)$ 不全相等
(品牌对销售量有影响)

2. 对因素B提出的假设为

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
(地区对销售量没有影响)
- $H_1: \mu_j (j = 1, 2, \dots, 5)$ 不全相等
(地区对销售量有影响)

双因素方差分析 (Excel 输出的结果)

差异源	SS	df	MS	F	P-value	F crit
行(因素A)	13004.55	3	4334.85	18.10777	9.46E-05	3.4903
列(因素B)	2011.7	4	502.925	2.100846	0.143665	3.2592
误差	2872.7	12	239.3917			
总和	17888.95	19				

结论:

- $F_A = 18.10777 > F_{\alpha} = 3.4903$, 拒绝原假设 H_0 , 说明彩电的品牌对销售量有显著影响
- $F_B = 2.100846 < F_{\alpha} = 3.2592$, 接受原假设 H_0 , 说明销售地区对彩电的销售量没有显著影响

本章小结

1. 方差分析(**ANOVA**)的概念
2. 方差分析思想和原理
3. 方差分析中的基本假设
4. 用**Excel**进行方差分析

结 束



第九章 列联分析

PowerPoint



第九章 列联分析

第一节 列联表

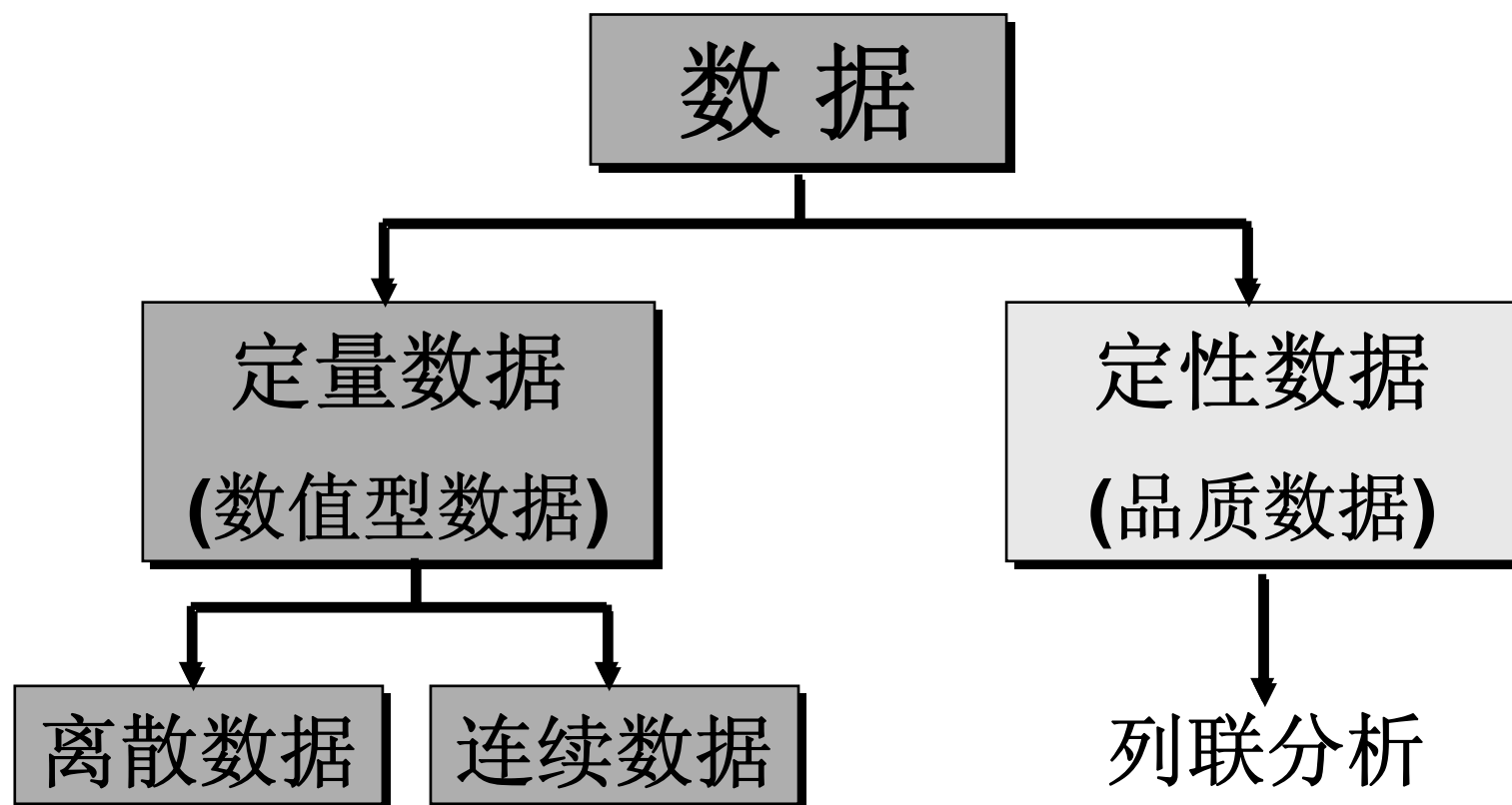
第二节 χ^2 分布与 χ^2 检验

第三节 列联表中的相关测量

学习目标

1. 解释列联表
2. 进行 χ^2 检验
 - 一致性检验
 - 独立性检验
3. 测度列联表中的相关性

数据的类型与列联分析



品质数据

1. 品质随机变量的结果表现为类别
 - 例如：性别 (男, 女)
2. 各类别用符号或数字代码来测度
3. 使用定类或定序尺度
 - 你吸烟吗？
 - 1.是； 2.否
 - 你赞成还是反对这一改革方案？
 - 1.赞成； 2.反对
4. 对品质数据的描述和分析通常使用列联表
5. 可使用 χ^2 检验

第一节 列联表

- 一. 列联表的构造
- 二. 列联表的分布

列联表的构造

列联表 (概念要点)

1. 由两个以上的变量进行交叉分类的频数分布表
2. 行变量的类别用 r 表示, r_i 表示第 i 个类别
3. 列变量的类别用 c 表示, c_j 表示第 j 个类别
4. 每种组合的观察频数用 f_{ij} 表示
5. 表中列出了行变量和列变量的所有可能的组合, 所以称为列联表
6. 一个 r 行 c 列的列联表称为 $r \times c$ 列联表

列联表的结构 (2×2 列联表)

一个 2×2 列联表

行 (r_i) \ 列 (c_j)	列 (c_j)		合计
	$j=1$	$j=2$	
$i=1$	f_{11}	f_{12}	$f_{11}+f_{12}$
$i=2$	f_{21}	f_{22}	$f_{21}+f_{22}$
合计	$f_{11}+f_{21}$	$f_{12}+f_{22}$	n

列联表的结构

($r \times c$ 列联表的一般表示)

r 行 c 列的列联表

列(c_j) 行(r_i)	列(c_j)			合计
	$j=1$	$j=2$...	
$i=1$	f_{11}	f_{12}	...	r_1
$i=2$	f_{21}	f_{22}	...	r_2
:	:	:	:	:
合计	c_1	c_2	...	n

f_{ij} 表示第 i 行第 j 列的观察频数

列联表 (一个实际例子)

【例】一个集团公司在四个不同的地区设有分公司，现该集团公司欲进行一项改革，此项改革可能涉及到各分公司的利益，故采用抽样调查方式，从四个分公司共抽取420个样本单位(人)，了解职工对此项改革的看法，调查结果如下表

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

列联表的分布

观察值的分布 (概念要点)

1. 边缘分布

- 行边缘分布

- 行观察值的合计数的分布
- 例如，赞成改革方案的共有279人，反对改革方案的141人

- 列边缘分布

- 列观察值的合计数的分布
- 例如，四个分公司接受调查的人数分别为100人，120人，90人，110人

2. 条件分布与条件频数

- 变量 X 条件下变量 Y 的分布，或在变量 Y 条件下变量 X 的分布
- 每个具体的观察值称为条件频数

观察值的分布 (图示)

	一分公司	二分公司	三分公司	四分公司	合计
赞成该方案	68	75	57	79	279
反对该方案	32	75	33	31	141
合计	100	120	90	110	420

条件频数

行边缘分布

列边缘分布

百分比分布 (概念要点)

1. 条件频数反映了数据的分布，但不适合进行对比
2. 为在相同的基数上进行比较，可以计算相应的百分比，称为百分比分布
 - 行百分比：行的每一个观察频数除以相应的行合计数 (f_{ij} / r_i)
 - 列百分比：列的每一个观察频数除以相应的列合计数 (f_{ij} / c_j)
 - 总百分比：每一个观察值除以观察值的总个数 (f_{ij} / n)

百分比分布 (图示)

	行百分比	列百分比				总百分比
	一分公司	二分公司	三分公司	四分公司	合计	
赞成该方案	24.4%	26.9%	20.4%	28.3%	66.4%	
	68.0%	62.5%	63.3%	71.8%	—	
	16.2%	17.8%	13.6%	18.8%	—	
反对该方案	22.7%	31.9%	23.4%	22.0%	33.6%	
	32.0%	37.5%	36.7%	28.2%	—	
	7.6%	10.7%	7.9%	7.4%	—	
合计	23.8%	28.6%	21.4%	26.2%	100%	

期望频数的分布 (概念要点)

1. 假定行变量和列变量是独立的
2. 一个实际频数 f_{ij} 的期望频数 e_{ij} ，是总频数的个数 n 乘以该实际频数 f_{ij} 落入第 i 行和第 j 列的概率，即

$$e_{ij} = n \cdot \left(\frac{r_i}{n} \right) \cdot \left(\frac{c_j}{n} \right) = \frac{r_i c_j}{n}$$

期望频数的分布 (算例)

➡ 例如，第1行和第1列的实际频数为 f_{11} ，它落在第1行的概率估计值为该行的频数之和 r_1 除以总频数的个数 n ，即： r_1/n ；它落在第1列的概率的估计值为该列的频数之和 c_1 除以总频数的个数 n ，即： c_1/n 。根据概率的乘法公式，该频数落在第1行和第1列的概率应为

$$\left(\frac{r_1}{n}\right) \cdot \left(\frac{c_1}{n}\right)$$

由于观察频数的总数为 n ，所以 f_{11} 的期望频数 e_{11} 应为

$$e_{11} = n \cdot \left(\frac{r_1}{n}\right) \cdot \left(\frac{c_1}{n}\right) = \frac{r_1 c_1}{n} = \frac{279 \times 100}{420} = 66.43 \approx 66$$

期望频数的分布 (算例)

➡根据上述公式计算的前例的期望频数

		一分公司	二分公司	三分公司	四分公司
赞成该 方案	实际频数	68	75	57	79
	期望频数	66	80	60	73
反对该 方案	实际频数	32	75	33	31
	期望频数	34	40	30	37

第二节 χ^2 分布与 χ^2 检验

- 一. χ^2 统计量
- 二. χ^2 检验

χ^2 统计量

χ^2 统计量 (要点)

1. 用于检验列联表中变量之间是否存在显著性差异，或者用于检验变量之间是否独立
2. 计算公式为

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

其自由度为 $(r-1)(c-1)$

式中： f_{ij} — 列联表中第 i 行第 j 列类别的实际频数

e_{ij} — 列联表中第 i 行第 j 列类别的期望频数

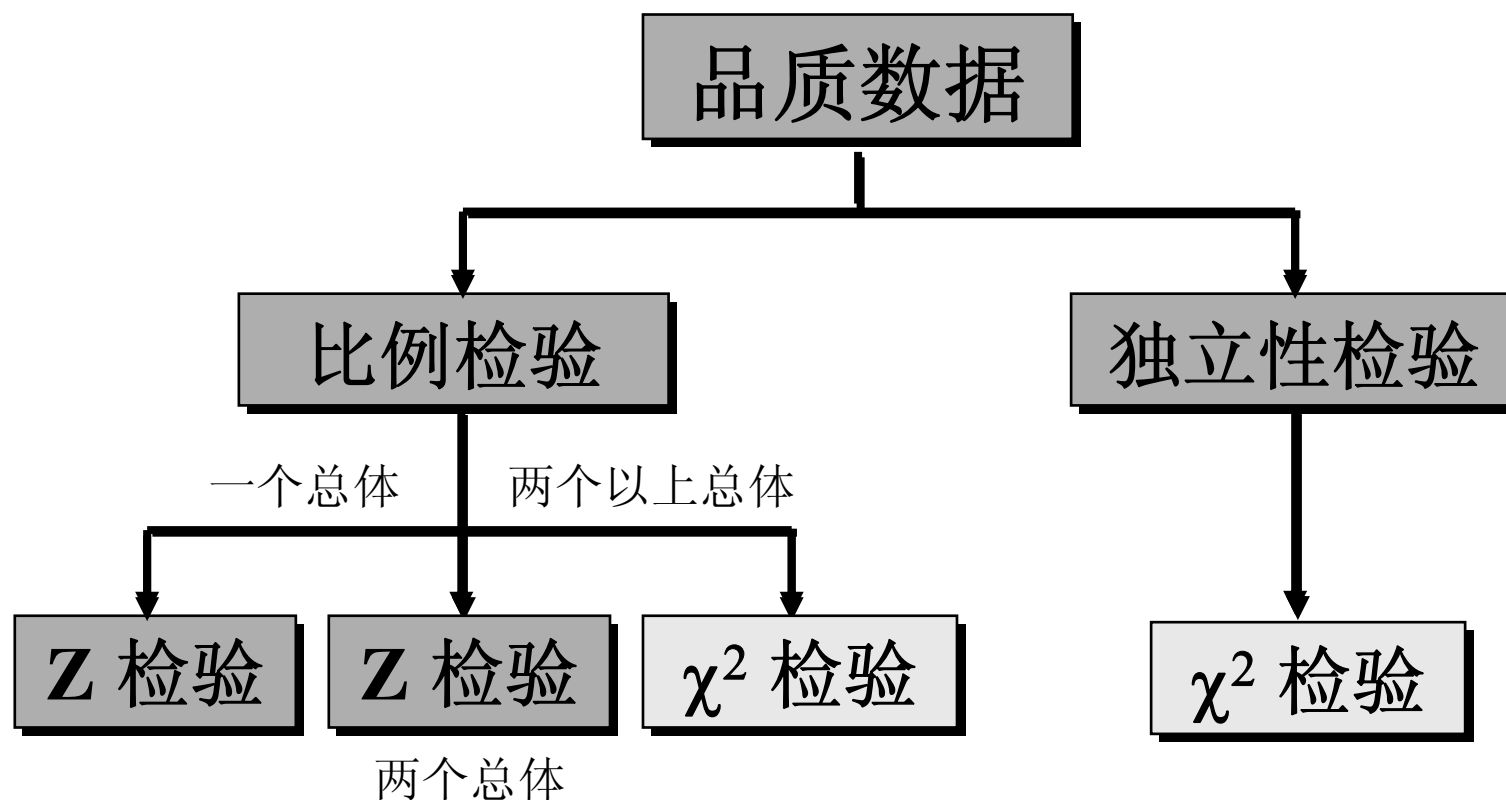
χ^2 统计量 (算例)

实际频数 (f_{ij})	期望频数 (e_{ij})	$f_{ij} - e_{ij}$	$(f_{ij} - e_{ij})^2$	$\frac{(f_{ij} - e_{ij})^2}{f}$
68	66	2	4	0.0606
75	80	-5	25	0.3125
57	60	-3	9	0.1500
79	73	6	36	0.4932
32	34	-2	4	0.1176
45	40	5	25	0.6250
33	30	3	9	0.3000
31	37	-6	36	0.9730

$$\chi^2 = \sum \frac{(f - e)^2}{e} = 3.0319$$

χ^2 检验

品质数据的假设检验



一致性检验 (要点)

1. 检验列联表中目标变量之间是否存在显著性差异
2. 检验的步骤为
 - 提出假设
 - $H_0: P_1 = P_2 = \dots = P_j$ (目标变量的各个比例一致)
 - $H_1: P_1, P_2, \dots, P_j$ 不全相等 (各个比例不一致)
 - 计算检验的统计量
$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$
 - 进行决策
 - 根据显著性水平 α 和自由度 $(r-1)(c-1)$ 查出临界值 χ_α^2
 - 若 $\chi^2 \geq \chi_\alpha^2$, 拒绝 H_0 ; 若 $\chi^2 < \chi_\alpha^2$, 接受 H_0

一致性检验 (实例)

【例】续前例，检验职工的态度是否与所在单位有关？
($\alpha=0.1$)

1. 提出假设

- $H_0: P_1 = P_2 = P_3 = P_4$ (赞成比例一致)
- $H_1: P_1, P_2, P_3, P_4$ 不全相等 (赞成比例不一致)

2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = 3.0319$$

3. 根据显著性水平 $\alpha=0.1$ 和自由度 $(2-1)(4-1)=3$ 查出相应的临界值 $\chi_{\alpha}^2=6.251$ 。由于 $\chi^2=3.0319 < \chi_{\alpha}^2=6.251$ ，接受 H_0

独立性检验 (要点)

1. 检验列联表中的行变量与列变量之间是否独立
2. 检验的步骤为

- 提出假设

- H_0 : 行变量与列变量独立
- H_1 : 行变量与列变量不独立

- 计算检验的统计量

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- 进行决策

- 根据显著性水平 α 和自由度 $(r-1)(c-1)$ 查出临界值 χ_{α}^2
- 若 $\chi^2 \geq \chi_{\alpha}^2$, 拒绝 H_0 ; 若 $\chi^2 < \chi_{\alpha}^2$, 接受 H_0

独立性检验 (实例)

【例】一种原料来自三个不同的地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。检验各地区与原料之间是否存在依赖关系 ($\alpha = 0.05$)

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

独立性检验 (实例)

1. 提出假设

- H_0 : 地区与原料等级之间独立
- H_1 : 地区与原料等级之间不独立

2. 计算检验的统计量

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = 19.82$$

3. 根据显著性水平 $\alpha=0.05$ 和自由度 $(3-1)(3-1)=4$ 查出相应的临界值 $\chi_{\alpha}^2=9.488$ 。由于 $\chi^2=19.82 > \chi_{\alpha}^2=9.448$, 拒绝 H_0

第三节 列联表中的相关测量

- 一. ϕ 相关系数
- 二. 列联相关系数
- 三. V 相关系数

列联表中的相关测量 (一般问题)

1. 品质相关

- 对品质数据(定类和定序数据)之间相关程度的测度

2. 列联表变量的相关属于品质相关

3. 列联表相关测量的指标主要有

- ϕ 相关系数
- 列联相关系数
- V 相关系数

ϕ 相关系数 (要点)

1. 测度 2×2 列联表中数据相关程度的一个量
2. 对于 2×2 列联表, ϕ 系数的值在 $0 \sim 1$ 之间
3. ϕ 相关系数计算公式为

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

$$\text{式中: } \chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

n 为实际频数的总个数, 即样本容量

ϕ 相关系数 (原理分析)

一个简化的 2×2 列联表

因素 Y	因素 X		合计
	x_1	x_2	
y_1	a	b	$a + b$
y_2	c	d	$c + d$
合计	$a + c$	$b + d$	n

ϕ 相关系数 (原理分析)

1. 列联表中每个单元格的期望频数分别为

$$e_{11} = \frac{(a+b)(a+c)}{n} \quad e_{21} = \frac{(a+c)(c+d)}{n}$$

$$e_{12} = \frac{(a+b)(b+d)}{n} \quad e_{22} = \frac{(b+d)(c+d)}{n}$$

2. 将各期望频数代入 χ^2 的计算公式得

$$\begin{aligned} \chi^2 &= \frac{(a-e_{11})^2}{e_{11}} + \frac{(b-e_{12})^2}{e_{12}} + \frac{(c-e_{21})^2}{e_{21}} + \frac{(d-e_{22})^2}{e_{22}} \\ &= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \end{aligned}$$

ϕ 相关系数 (原理分析)

3. 将 χ^2 入 ϕ 相关系数的计算公式得

$$\phi = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

- ad 等于 bc , $\phi = 0$, 表明变量 X 与 Y 之间独立
- 若 $b=0$, $c=0$, 或 $a=0$, $d=0$, 意味着各观察频数全部落在对角线上, 此时 $|\phi| = 1$, 表明变量 X 与 Y 之间完全相关

4. 列联表中变量的位置可以互换, ϕ 的符号没有实际意义, 故取绝对值即可

列联相关系数 (要点)

1. 用于测度大于 2×2 列联表中数据的相关程度
2. 计算公式为

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- C 的取值范围是 $0 \leq C < 1$
- $C = 0$ 表明列联表中的两个变量独立
- C 的数值大小取决于列联表的行数和列数，并随行数和列数的增大而增大
- 根据不同行和列的列联表计算的列联系数不便于比较

V 相关系数 (要点)

1. 计算公式为

$$V = \sqrt{\frac{\chi^2}{n \min[(r-1), (c-1)]}}$$

式中： $\min[(r-1), (c-1)]$ 表示取 $(r-1), (c-1)$ 中较小的一个

2. V 的取值范围是 $0 \leq V \leq 1$
3. $V = 0$ 表明列联表中的两个变量独立
4. $V = 1$ 表明列联表中的两个变量完全相关
5. 不同行和列的列联表计算的列联系数不便于比较
6. 当列联表中有一维为2， $\min[(r-1), (c-1)] = 1$, 此时

$$V = \phi$$

ϕ 、 C 、 V 的比较

1. 同一个列联表， ϕ 、 C 、 V 的结果会不同
2. 不同的列联表， ϕ 、 C 、 V 的结果也不同
3. 在对不同列联表变量之间的相关程度进行比较时，不同列联表中的行与行、列与列的个数要相同，并且采用同一种系数

列联表中的相关测量 (一个实例)

【例】一种原料来自三个不同地区，原料质量被分成三个不同等级。从这批原料中随机抽取500件进行检验，结果如下表。分别计算 ϕ 系数、C系数和V系数，并分析相关程度

地区	一级	二级	三级	合计
甲地区	52	64	24	140
乙地区	60	59	52	171
丙地区	50	65	74	189
合计	162	188	150	500

列联表中的相关测量 (一个实例)

解：已知 $n=500$ ，根据前面的计算 $\chi^2=19.82$ ，列联表为 3×3

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{19.82}{500}} = 0.199$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{19.82}{19.82 + 500}} = 0.195$$

$$V = \sqrt{\frac{\chi^2}{n \min[(r-1), (c-1)]}} = \sqrt{\frac{19.82}{500(2)}} = 0.141$$

结论：三个系数均不高，表明产地和原料等级之间的相关程度不高

本章小结

1. 解释列联表
2. 计算期望频数
3. 进行 χ^2 检验
 - 一致性检验
 - 独立性检验
4. 对列联表进行相关分析
5. 用**Excel**进行 χ^2 检验

结 束



第十章 相关与回归分析

PowerPoint



第十章 相关与回归分析

- 第一节 变量间的相关关系
- 第二节 一元线性回归
- 第三节 多元线性回归
- 第四节 可化为线性回归的曲线回归

学习目标

1. 掌握相关系数的含义、计算方法和应用
2. 掌握一元线性回归的基本原理和参数的最小二乘估计方法
3. 掌握回归方程的显著性检验
4. 利用回归方程进行预测
4. 掌握多元线性回归分析的基本方法
5. 了解可化为线性回归的曲线回归
6. 用 Excel 进行回归分析

第一节 变量间的相关关系

- 一. 变量相关的概念
- 二. 相关系数及其计算

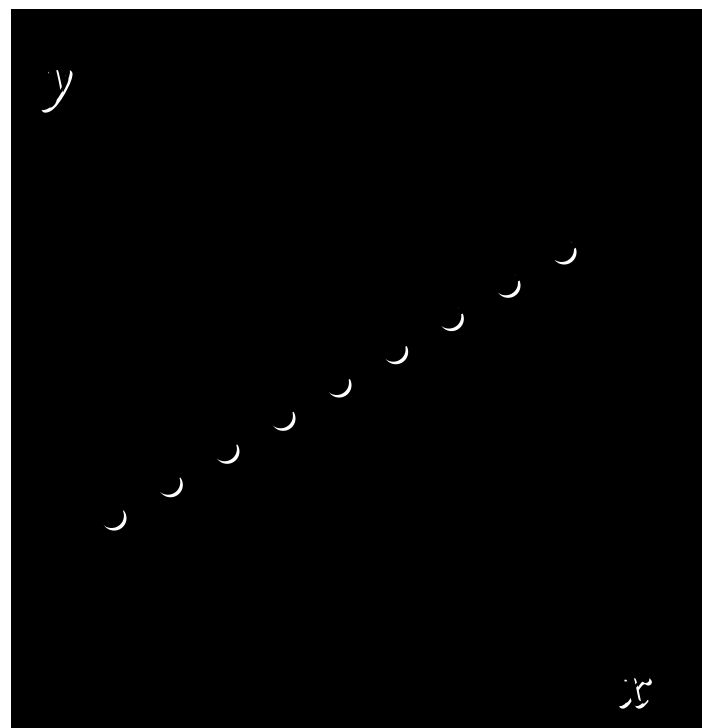
经济、管理类
基础课程

统计学

变量相关的概念

变量间的关系 (函数关系)

1. 是一一对应的确定关系
2. 设有两个变量 x 和 y ，变量 y 随变量 x 一起变化，并完全依赖于 x ，当变量 x 取某个数值时， y 依确定的关系取相应的值，则称 y 是 x 的函数，记为 $y = f(x)$ ，其中 x 称为自变量， y 称为因变量
3. 各观测点落在一条线上



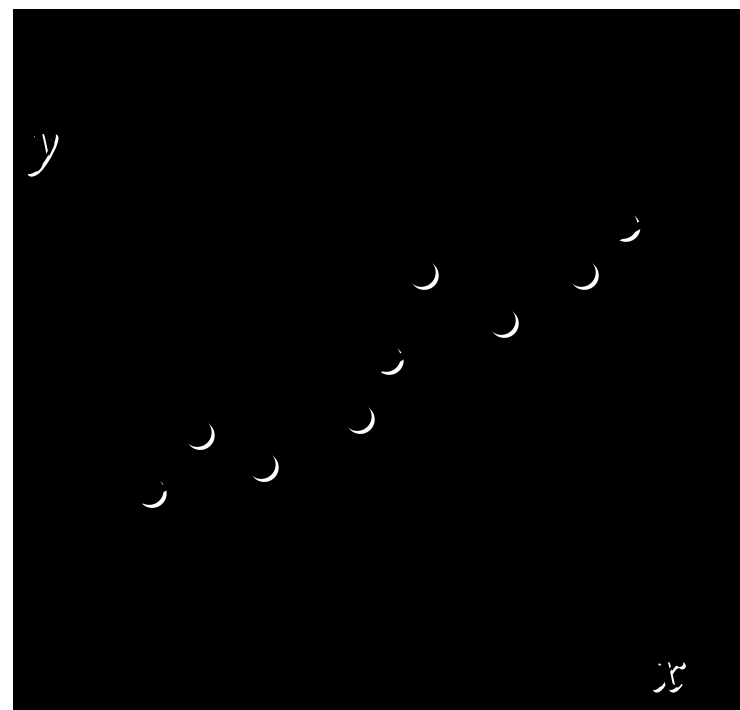
变量间的关系 (函数关系)

➡ 函数关系的例子

- 某种商品的销售额(y)与销售量(x)之间的关系可表示为 $y = p x$ (p 为单价)
- 圆的面积(S)与半径之间的关系可表示为 $S = \pi R^2$
- 企业的原材料消耗额(y)与产量(x_1)、单位产量消耗(x_2)、原材料价格(x_3)之间的关系可表示为 $y = x_1 x_2 x_3$

变量间的关系 (相关关系)

1. 变量间关系不能用函数关系精确表达
2. 一个变量的取值不能由另一个变量唯一确定
3. 当变量 x 取某个值时，变量 y 的取值可能有几个
4. 各观测点分布在直线周围

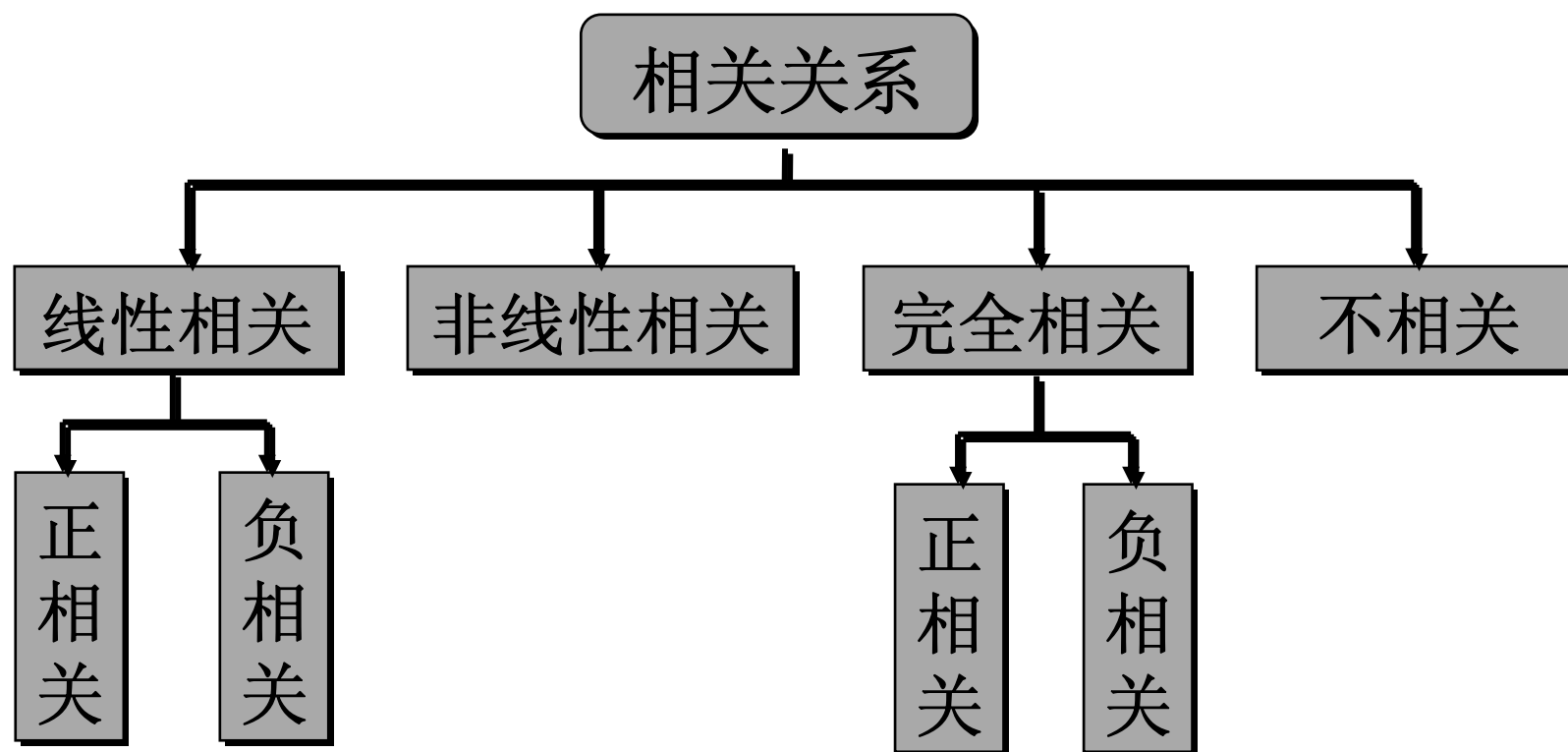


变量间的关系 (相关关系)

➡ 相关关系的例子

- 商品的消费量(y)与居民收入(x)之间的关系
- 商品销售额(y)与广告费支出(x)之间的关系
- 粮食亩产量(y)与施肥量(x_1)、降雨量(x_2)、温度(x_3)之间的关系
- 收入水平(y)与受教育程度(x)之间的关系
- 父亲身高(y)与子女身高(x)之间的关系

相关关系的类型



相关关系的图示



经济、管理类
基础课程

统计学

相关系数及其计算

相关关系的测度 (相关系数)

1. 对变量之间关系密切程度的度量
2. 对两个变量之间线性相关程度的度量称为简单相关系数
3. 若相关系数是根据总体全部数据计算的，称为总体相关系数，记为 ρ
4. 若是根据样本数据计算的，则称为样本相关系数，记为 r

相关关系的测度 (相关系数)

➡ 样本相关系数的计算公式

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}}$$

或化简为

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}}$$

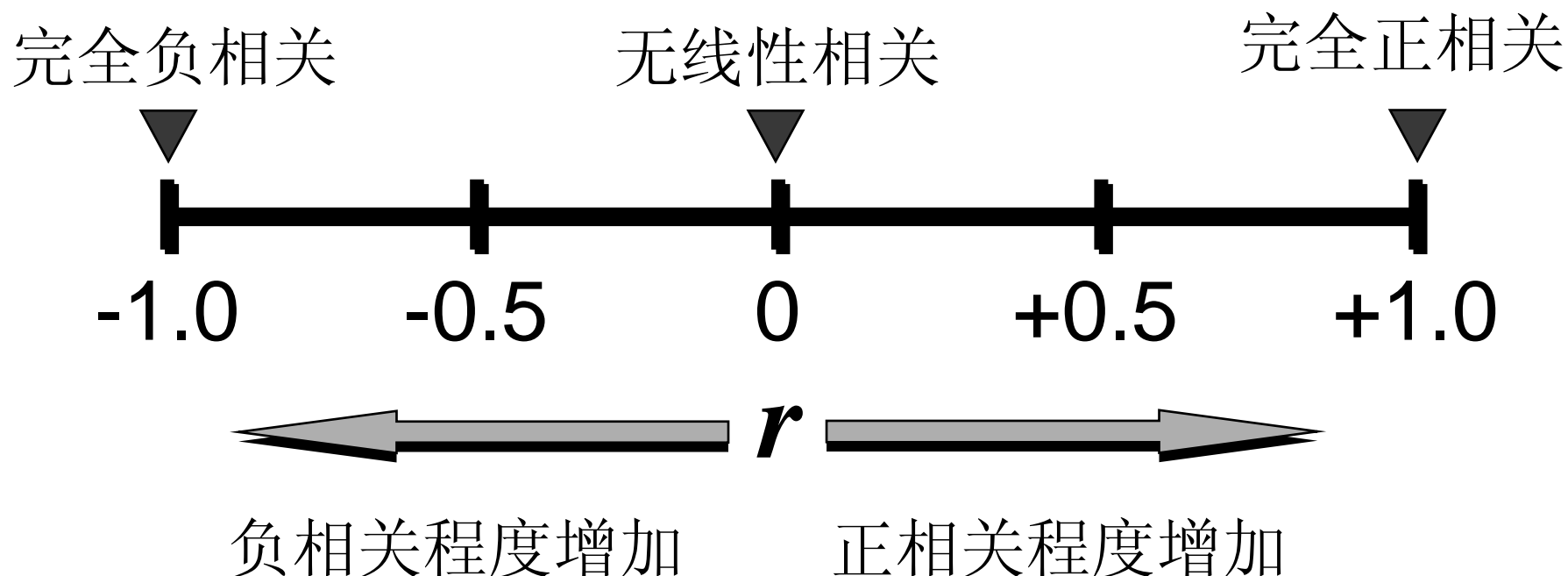
相关关系的测度

(相关系数取值及其意义)

1. r 的取值范围是 $[-1,1]$
2. $|r|=1$ ，为完全相关
 - $r=1$ ，为完全正相关
 - $r=-1$ ，为完全负正相关
3. $r=0$ ，不存在线性相关关系相关
4. $-1 \leq r < 0$ ，为负相关
5. $0 < r \leq 1$ ，为正相关
6. $|r|$ 越趋于1表示关系越密切； $|r|$ 越趋于0表示关系越不密切

相关关系的测度

(相关系数取值及其意义)



相关关系的测度 (相关系数计算例)

【例10.1】在研究我国人均消费水平的问题中，把全国人均消费额记为 y ，把人均国民收入记为 x 。我们收集到1981~1993年的样本数据 (x_i, y_i) ， $i=1,2,\dots, 13$ ，数据见表10-1，计算相关系数。

表10-1 我国人均国民收入与人均消费金额数据

单位:元

年份	人均 国民收入	人均 消费金额	年份	人均 国民收入	人均 消费金额
1981	393.8	249	1988	1068.8	643
1982	419.14	267	1989	1169.2	690
1983	460.86	289	1990	1250.7	713
1984	544.11	329	1991	1429.5	803
1985	668.29	406	1992	1725.9	947
1986	737.73	451	1993	2099.5	1148
1987	859.97	513			

相关关系的测度 (计算结果)

解：根据样本相关系数的计算公式有

$$\begin{aligned} r &= \frac{n \sum xy - \sum x \sum y}{\sqrt{n \sum x^2 - (\sum x)^2} \cdot \sqrt{n \sum y^2 - (\sum y)^2}} \\ &= \frac{13 \times 915617399 - 128275 \times 7457}{\sqrt{13 \times 1607332377 - (128275)^2} \cdot \sqrt{13 \times 5226399 - (7457)^2}} \\ &= 0.9987 \end{aligned}$$

人均国民收入与人均消费金额之间的相关系数为 **0.9987**

相关系数的显著性检验 (概念要点)

1. 检验两个变量之间是否存在线性相关关系
2. 等价于对回归系数 β_1 的检验
3. 采用 t 检验
4. 检验的步骤为
 - 提出假设: $H_0: \rho = 0$; $H_1: \rho \neq 0$
 - 计算检验的统计量: $t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$
 - 确定显著性水平 α , 并作出决策
 - 若 $|t| > t_{\alpha/2}$, 拒绝 H_0
 - 若 $|t| < t_{\alpha/2}$, 接受 H_0

相关系数的显著性检验 (实例)

➡ 对前例计算的相关系数进行显著性检验($\alpha=0.05$)

1. 提出假设: $H_0: \rho = 0$; $H_1: \rho \neq 0$

2. 计算检验的统计量

$$t = \frac{0.9987 \cdot \sqrt{13-2}}{\sqrt{1-0.9987^2}} = 64.9809$$

3. 根据显著性水平 $\alpha=0.05$, 查 t 分布表得 $t_{\alpha/2}(n-2)=2.201$

- 由于 $|t|=64.9809 > t_{\alpha/2}(13-2)=2.201$, 拒绝 H_0 , 人均消费金额与人均国民收入之间的相关关系显著

相关系数的显著性检验 (相关系数检验表的使用)

1. 若 $|r|$ 大于表上的 $\alpha=5\%$ 相应的值，小于表上 $\alpha=1\%$ 相应的值，称变量 x 与 y 之间有显著的线性关系
2. 若 $|r|$ 大于表上 $\alpha=1\%$ 相应的值，称变量 x 与 y 之间有十分显著的线性关系
3. 若 $|r|$ 小于表上 $\alpha=5\%$ 相应的值，称变量 x 与 y 之间没有明显的线性关系
4. 根据前例的 $r=0.9987 > \alpha=5\%(n-2)=0.553$ ，表明人均消费金额与人均国民收入之间有十分显著的线性相关关系

第二节 一元线性回归

- 一. 一元线性回归模型
- 二. 参数的最小二乘估计
- 三. 回归方程的显著性检验
- 四. 预测及应用

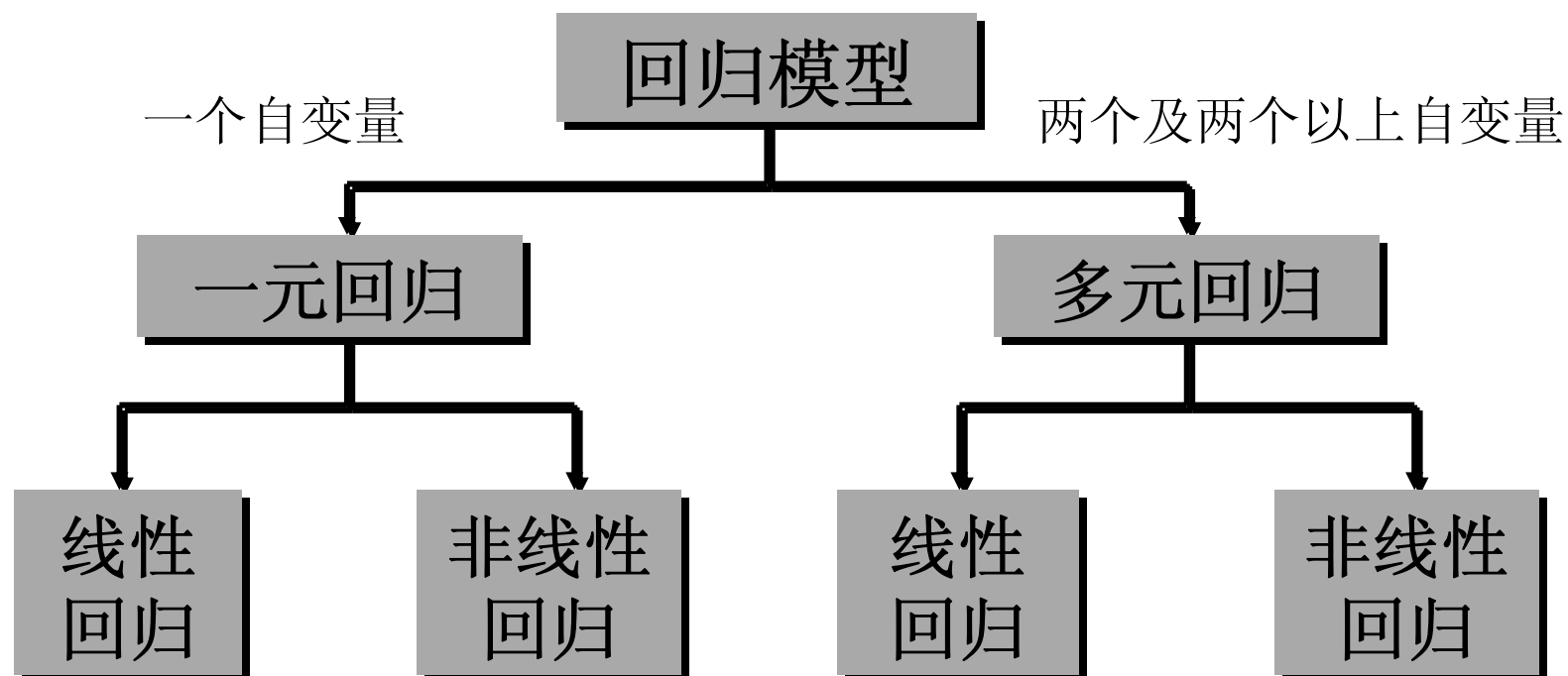
什么是回归分析？ (内容)

1. 从一组样本数据出发，确定变量之间的数学关系式
2. 对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著
3. 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度

回归分析与相关分析的区别

1. 相关分析中，变量 x 变量 y 处于平等的地位；回归分析中，变量 y 称为因变量，处在被解释的地位， x 称为自变量，用于预测因变量的变化
2. 相关分析中所涉及的变量 x 和 y 都是随机变量；回归分析中，因变量 y 是随机变量，自变量 x 可以是随机变量，也可以是非随机的确定变量
3. 相关分析主要是描述两个变量之间线性关系的密切程度；回归分析不仅可以揭示变量 x 对变量 y 的影响大小，还可以由回归方程进行预测和控制

回归模型的类型



经济、管理类
基础课程

统计学

回归模型与回归方程

回归模型

1. 回答“变量之间是什么样的关系？”
2. 方程中运用
 - 1 个数字的因变量(响应变量)
 - 被预测的变量
 - 1 个或多个数字的或分类的自变量 (解释变量)
 - 用于预测的变量
3. 主要用于预测和估计

一元线性回归模型 (概念要点)

1. 当只涉及一个自变量时称为一元回归，若因变量 y 与自变量 x 之间为线性关系时称为一元线性回归
2. 对于具有线性关系的两个变量，可以用一条线性方程来表示它们之间的关系
3. 描述因变量 y 如何依赖于自变量 x 和误差项 ε 的方程称为回归模型

一元线性回归模型 (概念要点)

➡ 对于只涉及一个自变量的简单线性回归模型可表示为

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- 模型中, y 是 x 的线性函数(部分)加上误差项
- 线性部分反映了由于 x 的变化而引起的 y 的变化
- 误差项 ε 是随机变量
 - 反映了除 x 和 y 之间的线性关系之外的随机因素对 y 的影响
 - 是不能由 x 和 y 之间的线性关系所解释的变异性
- β_0 和 β_1 称为模型的参数

一元线性回归模型 (基本假定)

1. 误差项 ε 是一个期望值为0的随机变量，即 $E(\varepsilon)=0$ 。对于一个给定的 x 值， y 的期望值为 $E(y)=\beta_0+\beta_1x$
2. 对于所有的 x 值， ε 的方差 σ^2 都相同
3. 误差项 ε 是一个服从正态分布的随机变量，且相互独立。即 $\varepsilon \sim N(0, \sigma^2)$
 - 独立性意味着对于一个特定的 x 值，它所对应的 ε 与其他 x 值所对应的 ε 不相关
 - 对于一个特定的 x 值，它所对应的 y 值与其他 x 所对应的 y 值也不相关

回归方程 (概念要点)

1. 描述 y 的平均值或期望值如何依赖于 x 的方程称为回归方程
2. 简单线性回归方程的形式如下

$$E(y) = \beta_0 + \beta_1 x$$

- 方程的图示是一条直线，因此也称为直线回归方程
- β_0 是回归直线在 y 轴上的截距，是当 $x=0$ 时 y 的期望值
- β_1 是直线的斜率，称为回归系数，表示当 x 每变动一个单位时， y 的平均变动值

估计(经验)的回归方程

1. 总体回归参数 β_0 和 β_1 是未知的，必需利用样本数据去估计
2. 用样本统计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 代替回归方程中的未知参数 β_0 和 β_1 ，就得到了估计的回归方程
3. 简单线性回归中估计的回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

其中： $\hat{\beta}_0$ 是估计的回归直线在 y 轴上的截距， $\hat{\beta}_1$ 是直线的斜率，它表示对于一个给定的 x 的值，是 y 的估计值，也表示 x 每变动一个单位时， y 的平均变动值

参数 β_0 和 β_1 的最小二乘估计

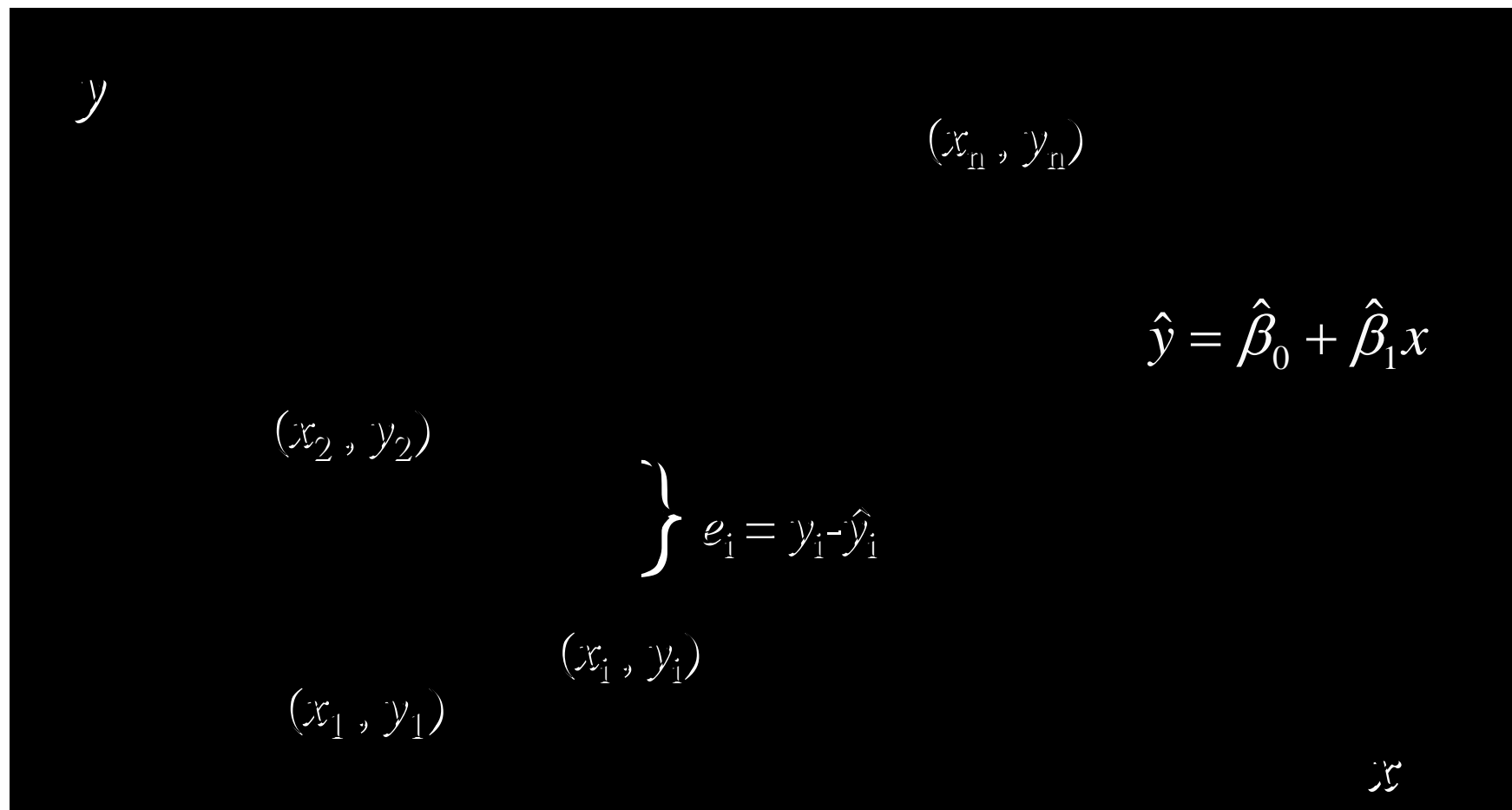
最小二乘法 (概念要点)

1. 使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方法。即

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{y})^2 = \sum_{i=1}^n e_i^2 = \text{最小}$$

2. 用最小二乘法拟合的直线来代表 x 与 y 之间的关系与实际数据的误差比其他任何直线都小

最小二乘法 (图示)



最小二乘法

($\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的计算公式)

➡ 根据最小二乘法的要求，可得求解 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准方程如下

$$\begin{cases} \hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

估计方程的求法 (实例)

【例】根据例10.1中的数据，配合人均消费金额对人均国民收入的回归方程

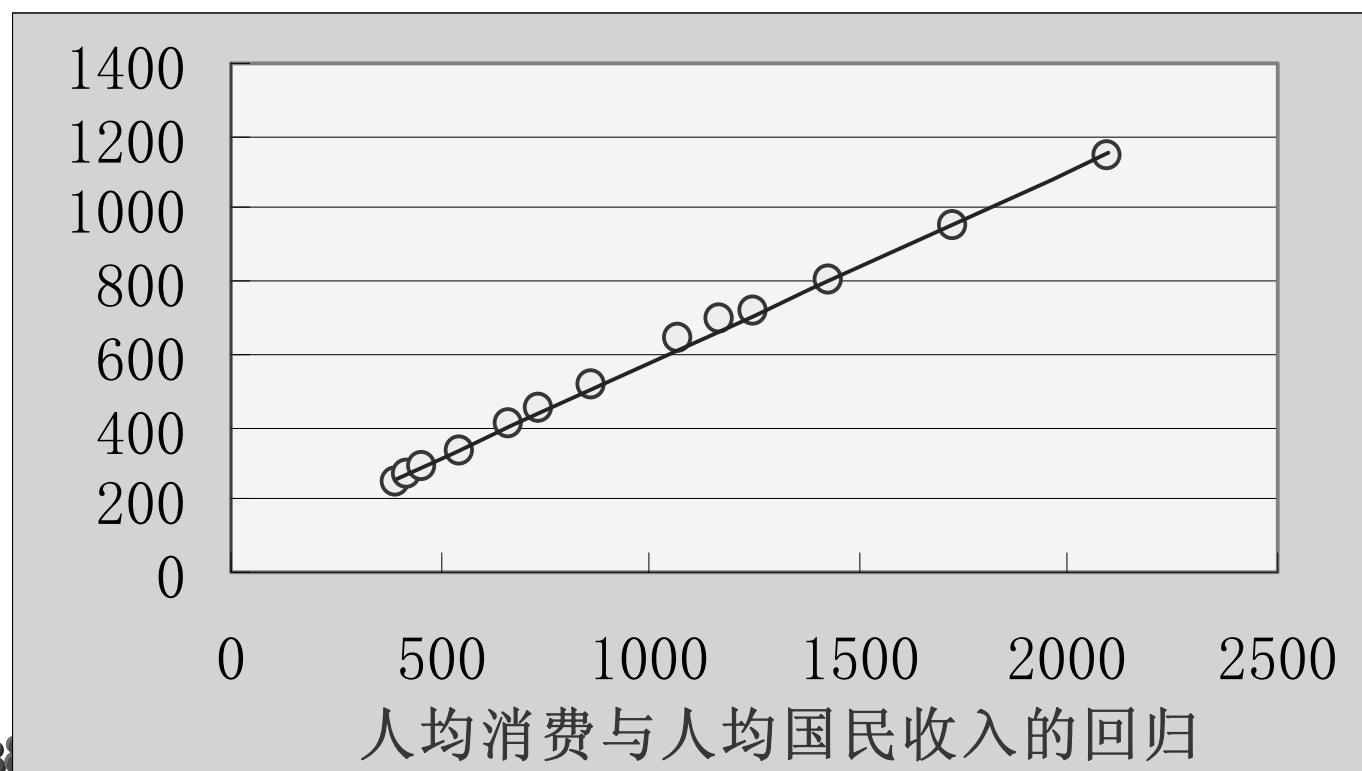
根据 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的求解公式得

$$\begin{cases} \hat{\beta}_1 = \frac{13 \times 9156173.99 - 12827.5 \times 7457}{13 \times 16073323.77 - (12827.5)^2} \\ \hat{\beta}_0 = 573.61538 - 0.52638 \times 986.73077 \end{cases}$$
$$\begin{cases} \hat{\beta}_1 = 0.526378 \\ \hat{\beta}_0 = 54.2229 \end{cases}$$

估计(经验)方程

人均消费金额对人均国民收入的回归方程为

$$\hat{y} = 54.22286 + 0.52638 x$$



估计方程的求法 (Excel的输出结果)

SUMMARY OUTPUT

回归统计

Mu1 0.998703821
R Square 0.997409322
Adjusted R 0.997173806
标准误差 14.94967766
观测值 13

$$\hat{\beta}_0 \pm t_{\alpha/2}(n-2) \cdot S_y \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\hat{\beta}_0 \quad \hat{\beta}_1$$

$$\hat{\beta}_1 \pm t_{\alpha/2}(n-2) \cdot \frac{S_y}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	54.22286392	8.99397869	6.028796	8.56501E-05	34.4272403	74.0184875
X Variable	0.52637714	0.00808855	65.07682	1.39842E-15	0.50857435	0.54417993

回归方程的显著性检验

离差平方和的分解

1. 因变量 y 的取值是不同的, y 取值的这种波动称为变差。变差来源于两个方面
 - 由于自变量 x 的取值不同造成的
 - 除 x 以外的其他因素(如 x 对 y 的非线性影响、测量误差等)的影响
2. 对一个具体的观测值来说, 变差的大小可以通过该实际观测值与其均值之差 $y - \bar{y}$ 来表示

离差平方和的分解 (图示)

y

(x_i, y_i)

$y - \hat{y}$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$y - \bar{y}$

$\hat{y} - \bar{y}$

离差分解图

x

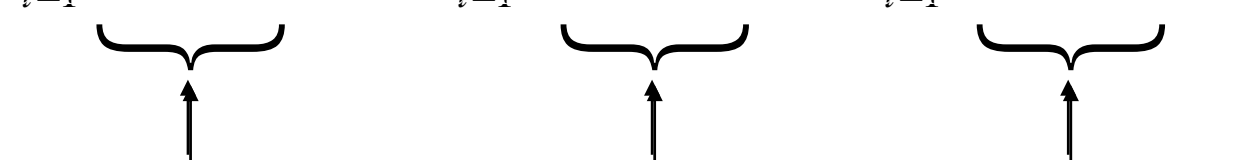
离差平方和的分解 (三个平方和的关系)

1. 从图上看有

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

2. 两端平方后求和有

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y})^2$$



总变差平方和 (SST) 回归平方和 (SSR) 残差平方和 (SSE)

$$SST = SSR + SSE$$

离差平方和的分解 (三个平方和的意义)

1. 总平方和(SST)

- 反映因变量的 n 个观察值与其均值的总离差

2. 回归平方和(SSR)

- 反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和

3. 残差平方和(SSE)

- 反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

样本决定系数 (判定系数 r^2)

1. 回归平方和占总离差平方和的比例

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

2. 反映回归直线的拟合程度
3. 取值范围在 $[0, 1]$ 之间
4. $r^2 \rightarrow 1$, 说明回归方程拟合的越好; $r^2 \rightarrow 0$, 说明回归方程拟合的越差
5. 判定系数等于相关系数的平方, 即 $r^2 = (r)^2$

回归方程的显著性检验 (线性关系的检验)

1. 检验自变量和因变量之间的线性关系是否显著
2. 具体方法是将回归离差平方和(SSR)同剩余离差平方和(SSE)加以比较, 应用 F 检验来分析二者之间的差别是否显著
 - 如果是显著的, 两个变量之间存在线性关系
 - 如果不显著, 两个变量之间不存在线性关系

回归方程的显著性检验 (检验的步骤)

1. 提出假设

- H_0 : 线性关系不显著

2. 计算检验统计量 F

$$F = \frac{SSR/1}{SSE/n-2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / 1}{\sum_{i=1}^n (y_i - \hat{y})^2 / n-2} \sim F(1, n-2)$$

3. 确定显著性水平 α ，并根据分子自由度1和分母自由度 $n-2$ 找出临界值 F_α

4. 作出决策：若 $F \geq F_\alpha$, 拒绝 H_0 ; 若 $F < F_\alpha$, 接受 H_0

回归方程的显著性检验 (方差分析表)

(续前例) Excel 输出的方差分析表

方差分析	平方和		均方			
	df	SS	MS	F	Significance F	
回归	1	946491	946491	4234.99	1.39842E-15	
残差	11	2458.42	223.493			
总计	12	948949				

估计标准误差 S_y

1. 实际观察值与回归估计值离差平方和的均方根
2. 反映实际观察值在回归直线周围的分散状况
3. 从另一个角度说明了回归直线的拟合程度
4. 计算公式为

$$S_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n y_i^2 - a \sum_{i=1}^n y_i - b \sum_{i=1}^n x_i y_i}{n-2}}$$

注：上例的计算结果为14.949678

回归系数的显著性检验 (要点)

1. 检验 x 与 y 之间是否具有线性关系，或者说，检验自变量 x 对因变量 y 的影响是否显著
2. 理论基础是回归系数 $\hat{\beta}_1$ 的抽样分布
3. 在一元线性回归中，等价于回归方程的显著性检验

回归系数的显著性检验 (样本统计量 $\hat{\beta}_1$ 的分布)

1. $\hat{\beta}_1$ 是根据最小二乘法求出的样本统计量，它有自己的分布
2. $\hat{\beta}_1$ 的分布具有如下性质
 - 分布形式：正态分布
 - 数学期望： $E(\hat{\beta}_1) = \beta_1$
 - 标准差： $\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}}$
 - 由于 σ 未知，需用其估计量 S_y 来代替得到 $\hat{\beta}_1$ 的估计的标准差

$$S_{\hat{\beta}_1} = \frac{S_y}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

回归系数的显著性检验 (样本统计量 $\hat{\beta}_1$ 的分布)

$\hat{\beta}_1$ 的抽样分布

$$S_{\hat{\beta}_1} = \frac{S_y}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$\hat{\beta}_1$

$$E(\hat{\beta}_1) = \beta_1$$

回归系数的显著性检验 (步骤)

1. 提出假设

- $H_0: \beta_1 = 0$ (没有线性关系)
- $H_1: \beta_1 \neq 0$ (有线性关系)

2. 计算检验的统计量

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

3. 确定显著性水平 α ，并进行决策

- $|t| > t_{\alpha/2}$ ，拒绝 H_0 ； $|t| < t_{\alpha/2}$ ，接受 H_0

回归系数的显著性检验 (实例)

☞ 对前例的回归系数进行显著性检验($\alpha=0.05$)

1. 提出假设

- $H_0: \beta_1 = 0$ 人均收入与人均消费之间无线性关系
- $H_1: \beta_1 \neq 0$ 人均收入与人均消费之间有线性关系

2. 计算检验的统计量

$$t = \frac{0.52638}{\sqrt{14.95^2 / 3416034827}} = 65.0758$$

3. $t=65.0758 > t_{\alpha/2}=2.201$, 拒绝 H_0 , 表明人均收入与人均消费之间有线性关系

回归系数的显著性检验 (Excel输出的结果)

SUMMARY OUTPUT

回归统计

Mul	0.998703821
R Square	0.997409322
Adjusted R	0.997173806
标准误差	14.94967766
观测值	13

$$S_{\hat{\beta}_0} = S_y \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$S_{\hat{\beta}_1} = \frac{S_y}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} = \frac{54.22286392}{8.99397869}$$

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} = \frac{0.52637714}{0.00808855}$$

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	54.22286392	8.99397869	6.028796	8.56501E-05	34.4272403	74.0184875
X Variable	0.52637714	0.00808855	65.07682	1.39842E-15	0.50857435	0.54417993

经济、管理类
基础课程

统计学

预测及应用

利用回归方程进行估计和预测

1. 根据自变量 x 的取值估计或预测因变量 y 的取值
2. 估计或预测的类型
 - 点估计
 - y 的平均值的点估计
 - y 的个别值的点估计
 - 区间估计
 - y 的平均值的置信区间估计
 - y 的个别值的预测区间估计

利用回归方程进行估计和预测 (点估计)

1. 对于自变量 x 的一个给定值 x_0 ，根据回归方程得到因变量 y 的一个估计值 \hat{y}_0
2. 点估计值有
 - y 的平均值的点估计
 - y 的个别值的点估计
3. 在点估计条件下，平均值的点估计和个别值的点估计是一样的，但在区间估计中则不同

利用回归方程进行估计和预测 (点估计)

☞ y 的平均值的点估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的平均值的一个估计值 $E(y_0)$ ，就是平均值的点估计
2. 在前面的例子中，假如我们要估计人均国民收入为2000元时，所有年份人均消费金额的的平均值，就是平均值的点估计。根据估计的回归方程得

$$\hat{y}_0 = 54.22286 + 0.52638 \times 2000 = 1160.98(\text{元})$$

利用回归方程进行估计和预测 (点估计)

☞ y 的个别值的点估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的一个个别值的估计值 \hat{y}_0 ，就是个别值的点估计
2. 比如，如果我们只是想知道1990年人均国民收入为1250.7元时的人均消费金额是多少，则属于个别值的点估计。根据估计的回归方程得

$$\hat{y}_0 = 54.22286 + 0.52638 \times 1250.7 = 712.57(\text{元})$$

利用回归方程进行估计和预测 (区间估计)

1. 点估计不能给出估计的精度，点估计值与实际值之间是有误差的，因此需要进行区间估计
2. 对于自变量 x 的一个给定值 x_0 ，根据回归方程得到因变量 y 的一个估计区间
3. 区间估计有两种类型
 - 置信区间估计
 - 预测区间估计

利用回归方程进行估计和预测 (置信区间估计)

☞ y 的平均值的置信区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的平均值 $E(y_0)$ 的估计区间，这一估计区间称为 **置信区间**
2. $E(y_0)$ 在 $1-\alpha$ 置信水平下的置信区间为

式中： S_y 为估计标准误差

利用回归方程进行估计和预测 (置信区间估计:算例)

【例】根据前例，求出人均国民收入为1250.7元时，人均消费金额95%的置信区间

解：根据前面的计算结果

$$\hat{y}_0 = 712.57, S_y = 14.95, t_{\alpha/2}(13-2) = 2.201, n = 13$$

置信区间为

$$712.57 \pm 2.201 \times 14.95 \sqrt{\frac{1}{13} + \frac{(1250.7 - 986.73077)^2}{3416034.827}}$$

$$712.57 \pm 10.265$$

人均消费金额95%的置信区间为 **702.305** 元
~**722.835**元之间

利用回归方程进行估计和预测 (预测区间估计)

☞ y 的个别值的预测区间估计

1. 利用估计的回归方程，对于自变量 x 的一个给定值 x_0 ，求出因变量 y 的一个个别值的估计区间，这一区间称为 **预测区间**
2. y_0 在 $1-\alpha$ 置信水平下的预测区间为

$$\hat{y}_0 \pm t_{\alpha/2}(n-2)S_y \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

注意！**#1**

利用回归方程进行估计和预测 (置信预测区间估计:算例)

【例】根据前例，求出1990年人均国民收入为1250.7元时，人均消费金额的95%的预测区间

解：根据前面的计算结果有

$$\hat{y}_0 = 712.57, S_y = 14.95, t_{\alpha/2}(13-2) = 2.201, n = 13$$

置信区间为

$$712.57 \pm 2.201 \times 14.95 \sqrt{1 + \frac{1}{13} + \frac{(1250.7 - 986.73077)^2}{3416034.827}}$$

712.57±34.469

人均消费金额95%的预测区间为**678.101元~747.039元**之间

影响区间宽度的因素

1. 置信水平 ($1 - \alpha$)
 - 区间宽度随置信水平的增大而增大
2. 数据的离散程度 (s)
 - 区间宽度随离散程度的增大而增大
3. 样本容量
 - 区间宽度随样本容量的增大而减小
4. 用于预测的 x_p 与 \bar{x} 的差异程度
 - 区间宽度随 x_p 与 \bar{x} 的差异程度的增大而增大

置信区间、预测区间、回归方程

y

预测上限
置信上限

置信下限
预测下限

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

\bar{x}

x_p

x

第三节 多元线性回归

- 一. 多元线性回归模型
- 二. 回归参数的估计
- 三. 回归方程的显著性检验
- 四. 回归系数的显著性检验
- 五. 多元线性回归的预测

多元线性回归模型

多元线性回归模型 (概念要点)

1. 一个因变量与两个及两个以上自变量之间的回归
2. 描述因变量 y 如何依赖于自变量 x_1, x_2, \dots, x_p 和误差项 ε 的方程称为多元线性回归模型
3. 涉及 p 个自变量的多元线性回归模型可表示为

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

- $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 是参数
- ε 是被称为误差项的随机变量
- y 是 x_1, x_2, \dots, x_p 的线性函数加上误差项 ε
- ε 说明了包含在 y 里面但不能被 p 个自变量的线性关系所解释的变异性

多元线性回归模型 (概念要点)

☞ 对于 n 组实际观察数据 $(y_i ; x_{i1}, x_{i2}, \dots, x_{ip})$, $(i=1,2,\dots,n)$, 多元线性回归模型可表示为

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + \varepsilon_2 \\ \dots\dots\dots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

多元线性回归模型 (基本假定)

1. 自变量 x_1, x_2, \dots, x_p 是确定性变量，不是随机变量
2. 随机误差项 ε 的期望值为0，且方差 σ^2 都相同
3. 误差项 ε 是一个服从正态分布的随机变量，即 $\varepsilon \sim N(0, \sigma^2)$ ，且相互独立

多元线性回归方程 (概念要点)

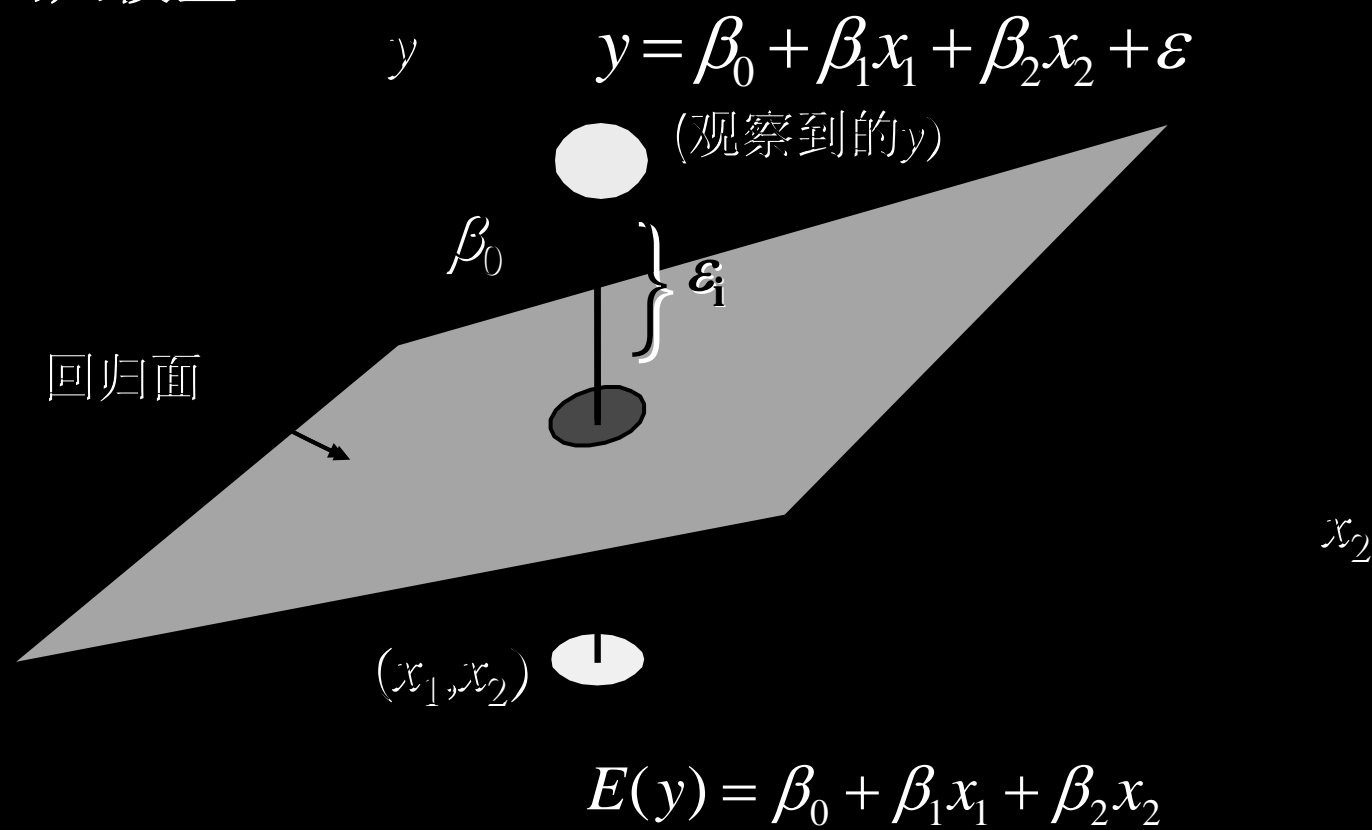
1. 描述 y 的平均值或期望值如何依赖于 x_1, x_1, \dots, x_p 的方程称为多元线性回归方程
2. 多元线性回归方程的形式为

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- $\beta_1, \beta_2, \dots, \beta_p$ 称为偏回归系数
- β_i 表示假定其他变量不变, 当 x_i 每变动一个单位时, y 的平均变动值

多元线性回归方程的直观解释

二元线性回归模型



多元线性回归的估计(经验)方程

1. 总体回归参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 是未知的, 利用样本数据去估计
2. 用样本统计量 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 代替回归方程中的未知参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 即得到估计的回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

- $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 是 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 估计值
- \hat{y} 是 y 的估计值

参数的最小二乘估计

参数的最小二乘法 (要点)

1. 使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ 。即

$$Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p) = \sum_{i=1}^n (y_i - \hat{y})^2 = \sum_{i=1}^n e_i^2 = \text{最小}$$

2. 根据最小二乘法的要求，可得求解各回归参数的标准方程如下

$$\begin{cases} \left. \frac{\partial Q}{\partial \beta_0} \right|_{\beta_0 = \hat{\beta}_0} = 0 \\ \left. \frac{\partial Q}{\partial \beta_i} \right|_{\beta_i = \hat{\beta}_i} = 0 \quad (i = 1, 2, \dots, p) \end{cases}$$

回归方程的显著性检验

多重样本决定系数 (多重判定系数 R^2)

1. 回归平方和占总离差平方和的比例

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

2. 反映回归直线的拟合程度
3. 取值范围在 $[0, 1]$ 之间
4. $R^2 \rightarrow 1$, 说明回归方程拟合的越好; $R^2 \rightarrow 0$, 说明回归方程拟合的越差
5. 等于多重相关系数的平方, 即 $R^2 = (R)^2$

修正的多重样本决定系数 (修正的多重判定系数 R^2)

1. 由于增加自变量将影响到因变量中被估计的回归方程所解释的变异性的数量，为避免高估这一影响，需要用自变量的数目去修正 R^2 的值
2. 用 n 表示观察值的数目， p 表示自变量的数目，修正的多元判定系数的计算公式可表示为

$$R_{\text{修}}^2 = 1 - \left(1 - R^2\right) \times \frac{n-1}{n-p-1}$$

回归方程的显著性检验 (线性关系的检验)

1. 检验因变量与所有的自变量和之间的是否存在一个显著的线性关系，也被称为总体的显著性检验
2. 检验方法是将回归离差平方和(SSR)同剩余离差平方和(SSE)加以比较，应用 F 检验来分析二者之间的差别是否显著
 - 如果是显著的，因变量与自变量之间存在线性关系
 - 如果不显著，因变量与自变量之间不存在线性关系

回归方程的显著性检验 (步骤)

1. 提出假设

- $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ 线性关系不显著
- $H_1: \beta_1, \beta_2, \dots, \beta_p$ 至少有一个不等于0

2. 计算检验统计量 F

$$F = \frac{SSR/p}{SSE/n-p-1} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y})^2 / n - p - 1} \sim F(p, n - p - 1)$$

3. 确定显著性水平 α 和分子自由度 p 、分母自由度 $n-p-1$ 找出临界值 F_α

4. 作出决策: 若 $F \geq F_\alpha$, 拒绝 H_0 ; 若 $F < F_\alpha$, 接受 H_0

回归系数的显著性检验 (要点)

1. 如果 F 检验已经表明了回归模型总体上是显著的，那么回归系数的检验就是用来确定每一个单个的自变量 x_i 对因变量 y 的影响是否显著
2. 对每一个自变量都要单独进行检验
3. 应用 t 检验
4. 在多元线性回归中，回归方程的显著性检验不再等价于回归系数的显著性检验

回归系数的显著性检验 (步骤)

1. 提出假设

- $H_0: \beta_i = 0$ (自变量 x_i 与 因变量 y 没有线性关系)
- $H_1: \beta_i \neq 0$ (自变量 x_i 与 因变量 y 有线性关系)

2. 计算检验的统计量 t

$$t = \frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} \sim t(n - p - 1)$$

3. 确定显著性水平 α ，并进行决策

- $|t| \geq t_{\alpha/2}$ ，拒绝 H_0 ； $|t| < t_{\alpha/2}$ ，接受 H_0

一个二元线性回归的例子

【例】一家百货公司在10个地区设有经销分公司。公司认为商品销售额与该地区的人口数和年人均收入有关，并希望建立它们之间的数量关系式，以预测销售额。有关数据如下表。试确定销售额对人口数和年人均收入的线性回归方程，并分析回归方程的拟合程度，对线性关系和回归系数进行显著性检验($\alpha=0.05$)。

销售额、人口数和年人均收入数据			
地区 编号	销售额 (万元) y	人口数 (万人) x_1	年人均收入 (元) x_2
1	33.3	32.4	1250
2	35.5	29.1	1650
3	27.6	26.3	1450
4	30.4	31.2	1310
5	31.9	29.2	1310
6	53.1	40.7	1580
7	35.6	29.8	1490
8	29.0	23.0	1520
9	35.1	28.2	1620
10	34.5	26.9	1570

一个二元线性回归的例子 (Excel 输出的结果)

SUMMARY OUTPUT

回归统计

Mul 0.968159025
R Square 0.937331897
Adjusted R Squa 0.919426725
标准误差 2.010050279
观测值 10

$$R^2_{\text{调整}} = 1 - (1 - R^2) \times \frac{n-1}{n-p-1}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-p-1}}$$

方差分析

	df	SS	MS	F	ignificance F
回归分析	2	423.01789	211.50894	52.34978	6.1612E-05
残差	7	28.282115	4.0403021		
总计	9	451.3			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-38.8251694	8.4785911	-4.579201	0.002546	-58.873837	-18.7765
X Variable 1	1.340693618	0.1433159	9.3548147	3.31E-05	1.00180562	1.679582
X Variable 2	0.022802293	0.0047542	4.7962172	0.001975	0.01156035	0.034044

一个二元线性回归的例子 (计算机输出结果解释)

1. 销售额与人口数和年人均收入的二元回归方程为

$$\hat{y} = -38.8252 + 1.341x_1 + 0.0228x_2$$

2. 多重判定系数 $R^2 = 0.9373$ ；调整后的 $R^2 = 0.9194$

3. 回归方程的显著性检验

- $F = 52.3498$ $F > F_{0.05}(2,7) = 4.74$ ，回归方程显著

4. 回归系数的显著性检验

- $t_{\beta 1} = 9.3548 > t_{\alpha 2} = 0.3646$ ，； $t_{\beta 2} = 4.7962 > t_{\alpha 2} = 2.3646$ ；
两个回归系数均显著

第三节 可化为线性回归的 曲线回归

- 一. 基本概念
- 二. 非线性模型及其线性化方法

非线性回归

1. 因变量 y 与 x 之间不是线性关系
2. 可通过变量代换转换成线性关系
3. 用最小二乘法求出参数的估计值
4. 并非所有的非线性模型都可以化为线性模型

几种常见的非线性模型

➡ 指数函数

1. 基本形式: $y = \alpha e^{\beta x}$

2. 线性化方法

- 两端取对数得: $\ln y = \ln \alpha + \beta x$
- 令: $y' = \ln y$, 则有 $y' = \ln \alpha + \beta x$

3. 图像

$$\beta > 0$$

$$\beta > 0$$

几种常见的非线性模型

➔ 幂函数

1. 基本形式: $y = \alpha x^\beta$

2. 线性化方法

- 两端取对数得: $\lg y = \lg \alpha + \beta \lg x$

- 令: $y' = \lg y$, $x' = \lg x$, 则 $y' = \lg \alpha + \beta x'$

3. 图像

$\beta > 1$ $\beta = 1$ $\beta < -1$

$0 < \beta < 1$ $-1 < \beta < 0$

几种常见的非线性模型

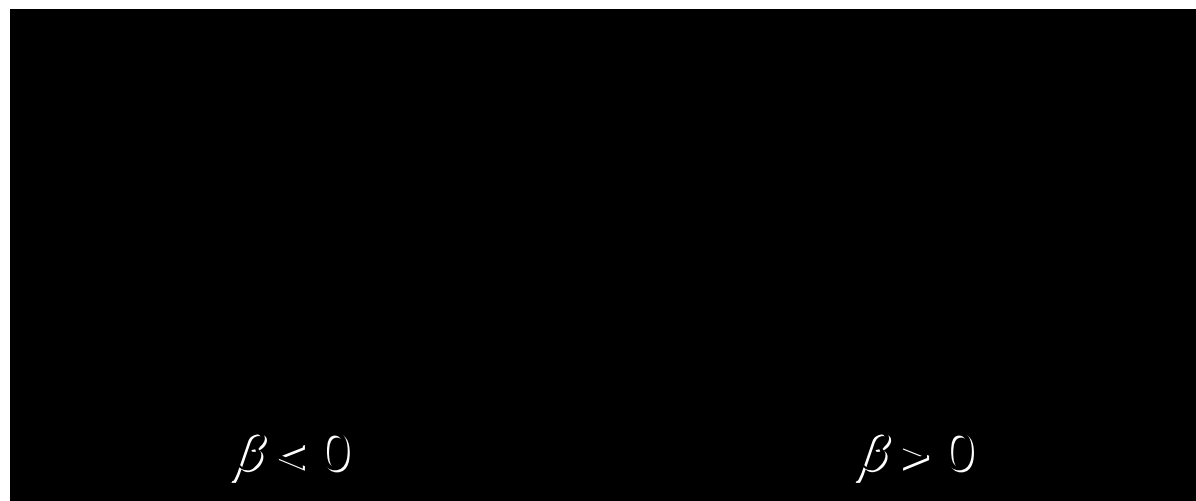
➡ 双曲线函数

1. 基本形式: $y = \frac{x}{\alpha x + \beta}$

2. 线性化方法

■ 令: $y' = 1/y$, $x' = 1/x$, 则有 $y' = \alpha + \beta x'$

3. 图像



几种常见的非线性模型

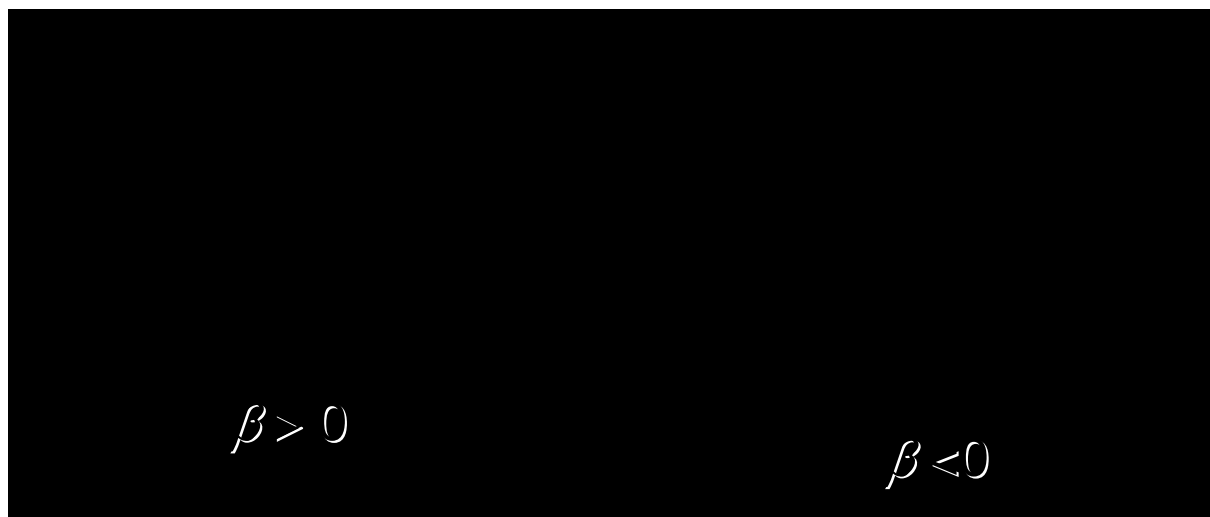
➡ 对数函数

1. 基本形式: $y = \alpha + \beta \lg x$

2. 线性化方法

■ $x' = \lg x$, 则有 $y' = \alpha + \beta x'$

3. 图像



几种常见的非线性模型

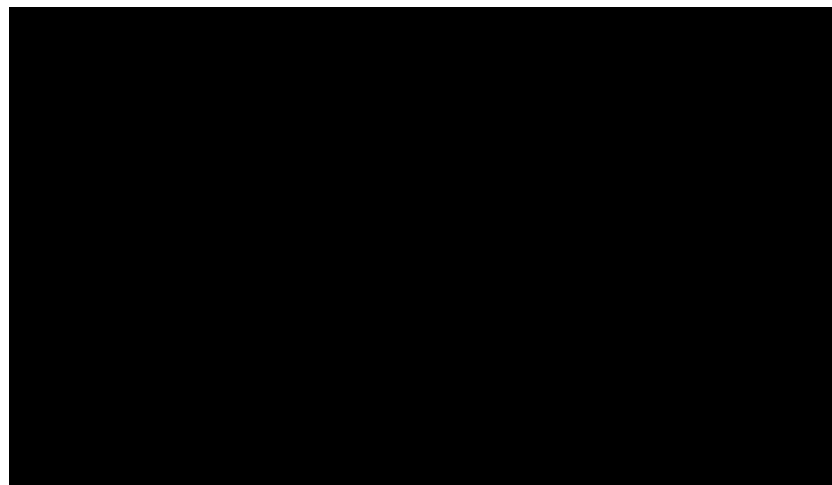
➔ S 型曲线

1. 基本形式:
$$y = \frac{1}{\alpha + \beta e^{-x}}$$

2. 线性化方法

■ 令: $y' = 1/y$, $x' = e^{-x}$, 则有 $y' = \alpha + \beta x'$

3. 图像



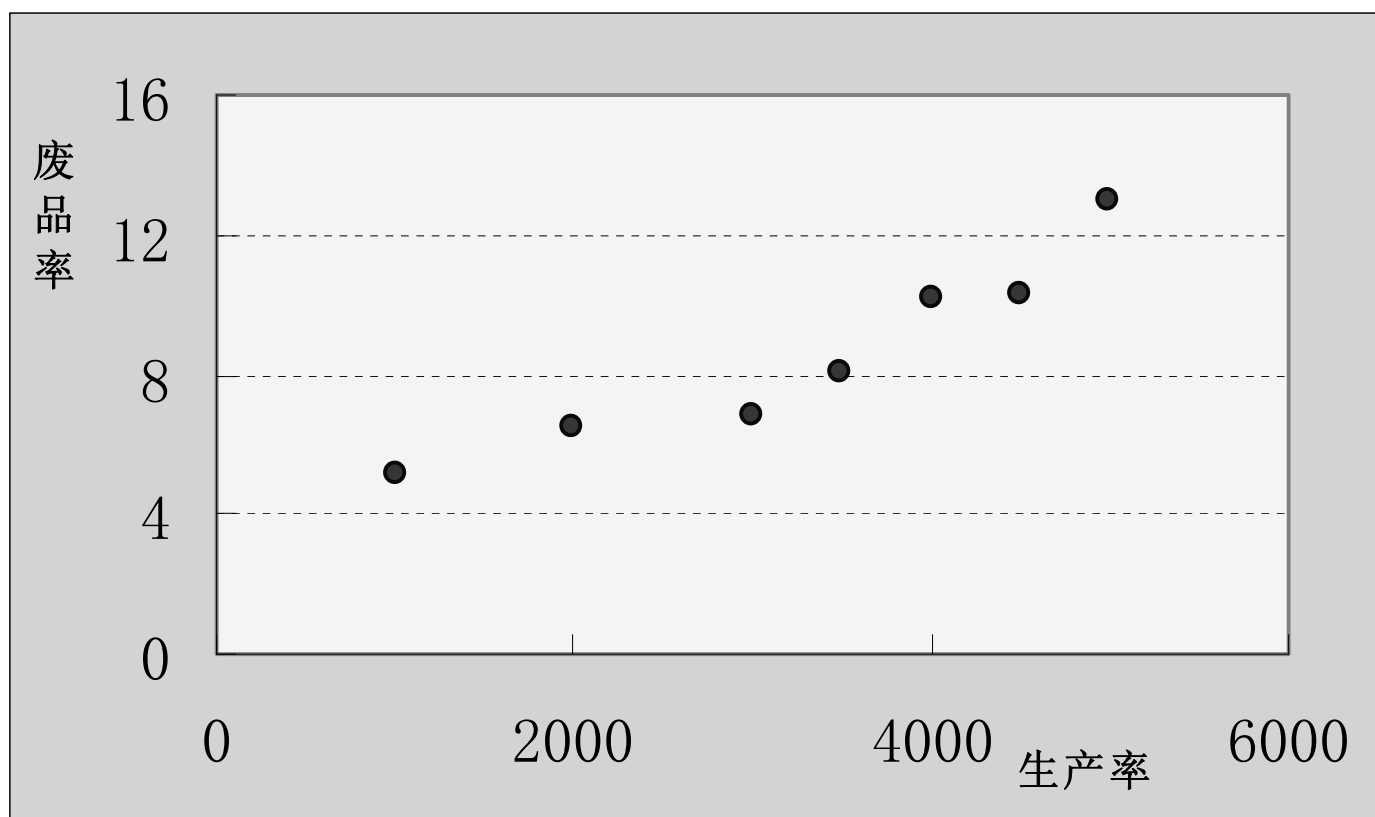
非线性回归 (实例)

【例】为研究生产率与废品率之间的关系，记录数据如下表。试拟合适当的模型。

废品率与生产率的关系							
生产率（周/单位） x	1000	2000	3000	3500	4000	4500	5000
废品率（%） y	5.2	6.5	6.8	8.1	10.2	10.3	13.0

非线性回归 (实例)

生产率与废品率的散点图



非线性回归 (实例)

1. 用线性模型: $y = \beta_0 + \beta_1 x + \varepsilon$, 有

$$y = 2.671 + 0.0018x$$

2. 用指数模型: $y = \alpha \beta^x$, 有

$$y = 4.05 \times (1.0002)^x$$

3. 比较

直线的残差平方和 = 5.3371 < 指数模型的残差平方和 = 6.11。直线模型略好于指数模型

本章小结

1. 相关系数与相关分析
2. 一元线性回归模型、回归方程与估计的回归方程
3. 多元线性回归模型、回归方程与估计的回归方程
4. 回归方程与回归系数的显著性检验
5. 非线性回归的线性化
5. 用Excel 进行回归分析

结 束



第十一章 时间序列分析

PowerPoint



第十一章 时间序列分析

第一节 时间序列的对比分析

第二节 长期趋势分析

第三节 季节变动分析

第四节 循环波动分析

学习目标

1. 掌握时间序列对比分析的方法
2. 掌握长期趋势分析的方法及应用
3. 掌握季节变动分析的原理与方法
4. 掌握循环波动的分析方法

第一节 时间序列的对比分析

- 一. 时间序列及其分类
- 二. 时间序列的水平分析
- 三. 时间序列的速度分析

经济、管理类
基础课程

统计学

时间序列及其分类

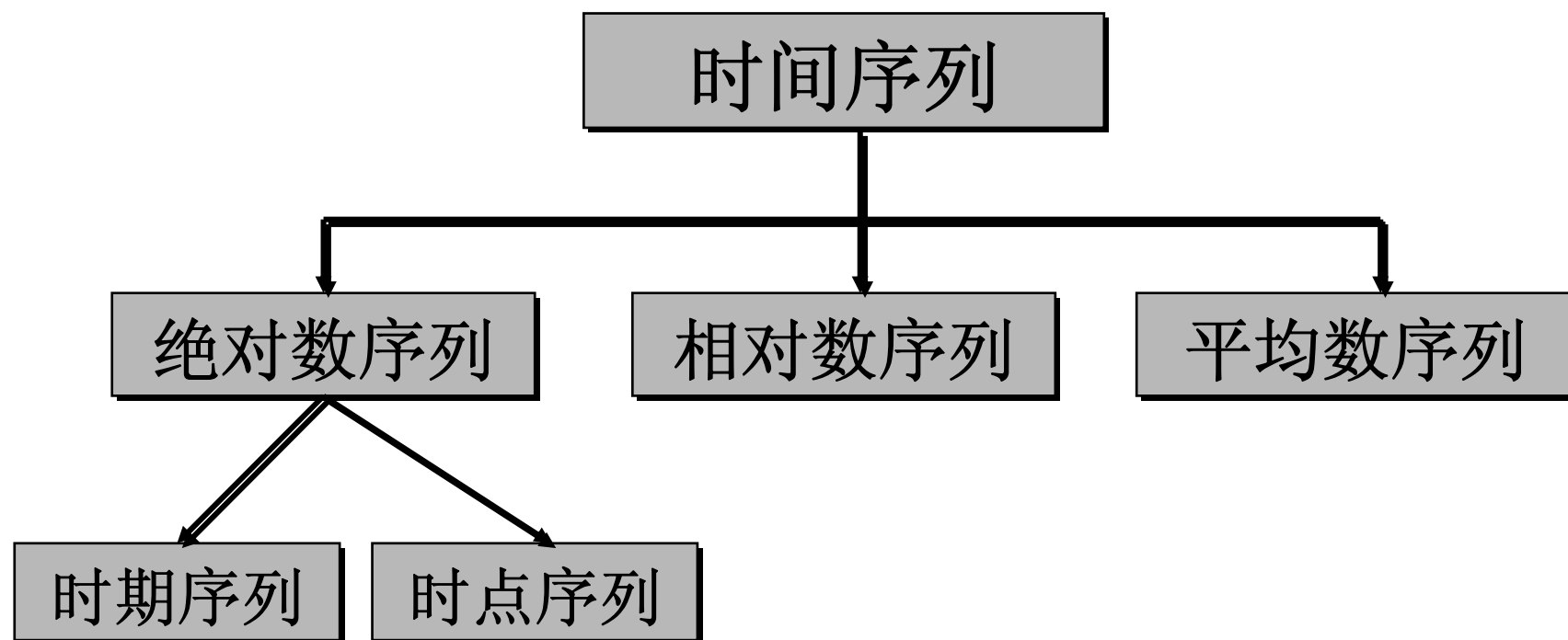
1. 同一现象在不同时间上的相继观察值排列而成的数列
2. 形式上由现象所属的时间和现象在不同时间上的观察值两部分组成
3. 排列的时间可以是年份、季度、月份或其他任何时间形式

时间序列 (一个例子)

表11-1 国内生产总值等时间序列

年 份	国内生产总值 (亿元)	年末总人口 (万人)	人口自然增长率 (‰)	居民消费水平 (元)
1990	18547.9	114333	14.39	803
1991	21617.8	115823	12.98	896
1992	26638.1	117171	11.60	1070
1993	34634.4	118517	11.45	1331
1994	46759.4	119850	11.21	1781
1995	58478.1	121121	10.55	2311
1996	67884.6	122389	10.42	2726
1997	74772.4	123626	10.06	2944
1998	79552.8	124810	9.53	3094

时间序列的分类



时间序列的分类

1. 绝对数时间序列

- 一系列绝对数按时间顺序排列而成
- 时间序列中最基本的表现形式
- 反映现象在不同时间上所达到的绝对水平
- 分为时期序列和时点序列
 - 时期序列：现象在一段时期内总量的排序
 - 时点序列：现象在某一瞬间时点上总量的排序

2. 相对数时间序列

- 一系列相对数按时间顺序排列而成

3. 平均数时间序列

- 一系列平均数按时间顺序排列而成

时间序列的水平分析

发展水平与平均发展水平 (概念要点)

1. 发展水平

- 现象在不同时间上的观察值
- 说明现象在某一时间上所达到的水平
- 表示为 Y_1, Y_2, \dots, Y_n 或 $Y_0, Y_1, Y_2, \dots, Y_n$

2. 平均发展水平

- 现象在不同时间上取值的平均数，又称序时平均数
- 说明现象在一段时期内所达到的一般水平
- 不同类型的时间序列有不同的计算方法

绝对数序列的序时平均数 (计算方法)

☞ 时期序列

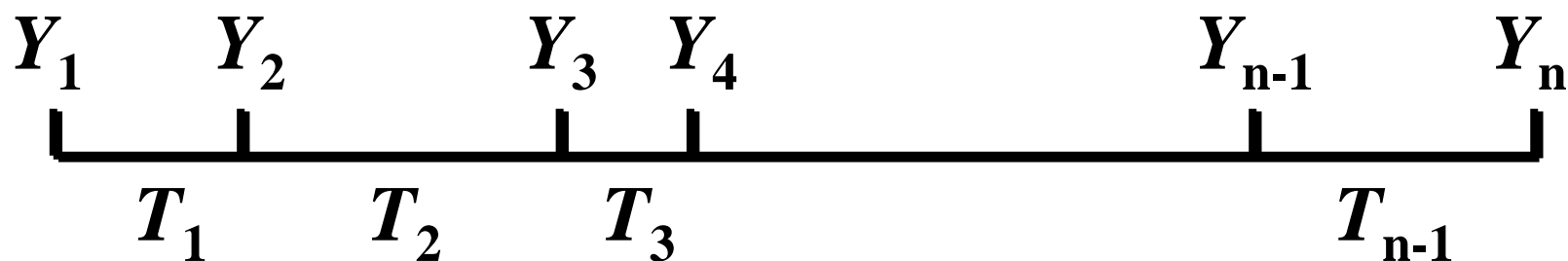
计算公式:
$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n} = \frac{\sum_{i=1}^n Y_i}{n}$$

【例11.1】 根据表11.1中的国内生产总值序列，计算各年度的平均国内生产总值

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{428885.5}{9} = 47653.94(\text{亿元})$$

绝对数序列的序时平均数 (计算方法)

☞ 时点序列— 间隔不相等



绝对数序列的序时平均数 (计算方法)

☞ 计算步骤

1. 计算出两个点值之间的平均数

$$\bar{Y}_1 = \frac{Y_1 + Y_2}{2} \quad \bar{Y}_2 = \frac{Y_2 + Y_3}{2} \quad \dots \quad \bar{Y}_{n-1} = \frac{Y_{n-1} + Y_n}{2}$$

2. 用相隔的时期长度 (T_i) 加权计算总的平均数

$$\bar{Y} = \frac{\left(\frac{Y_1 + Y_2}{2}\right)T_1 + \left(\frac{Y_2 + Y_3}{2}\right)T_2 + \dots + \left(\frac{Y_{n-1} + Y_n}{2}\right)T_{n-1}}{\sum_{i=1}^{n-1} T_i}$$

绝对数序列的序时平均数 (计算方法)

☞ 时点序列—间隔相等



当间隔相等($T_1 = T_2 = \dots = T_{n-1}$)时, 有

$$\bar{Y} = \frac{\frac{Y_1}{2} + Y_2 + \dots + Y_{n-1} + \frac{Y_n}{2}}{n - 1}$$

绝对数序列的序时平均数 (实例)

【例11.2】 设某种股票1999年各统计时点的收盘价如表11-2，计算该股票1999年的年平均价格

表11-2 某种股票1999年各统计时点的收盘价					
统计时点	1月1日	3月1日	7月1日	10月1日	12月31日
收盘价(元)	15.2	14.2	17.6	16.3	15.8

$$\bar{Y} = \frac{\left(\frac{15.2+14.2}{2}\right) \times 2 + \left(\frac{14.2+17.6}{2}\right) \times 4 + \left(\frac{17.6+16.3}{2}\right) \times 3 + \left(\frac{16.3+15.8}{2}\right) \times 3}{2+4+3+3}$$
$$= 16.0(\text{元})$$

绝对数序列的序时平均数 (实例)

【例11.3】 根据表11-1中年末总人口数序列，计算1991~1998年间的年平均人口数

$$\begin{aligned}\bar{Y} &= \frac{\frac{114333}{2} + 115823 + \dots + 123626 + \frac{124810}{2}}{9-1} \\ &= 119758.56(\text{万人})\end{aligned}$$

相对数序列的序时平均数 (计算方法)

1. 先分别求出构成相对数或平均数的分子 a_i 和分母 b_i 的平均数
2. 再进行对比, 即得相对数或平均数序列的序时平均数
3. 基本公式为

$$\bar{Y} = \frac{\bar{a}}{\bar{b}}$$

相对数序列的序时平均数 (计算方法与实例)

【例11.4】 已知1994~1998年我国的国内生产总值及构成数据如表11-3。计算1994~1998年间我国第三产业国内生产总值占全部国内生产总值的平均比重

表11-3 我国国内生产总值及其构成数据

年 份	1994	1995	1996	1997	1998
国内生产总值(亿元)	46759.4	58478.1	67884.6	74772.4	79552.8
其中：第三产业(亿元)	14930.0	17947.2	20427.5	24033.3	26104.3
比重(%)	31.9	30.7	30.1	32.1	32.8

相对数序列的序时平均数 (计算结果)

解：第三产业国内生产总值的平均数

$$\bar{a} = \frac{\sum_{i=1}^n a_i}{n} = \frac{103442.3}{5} = 20688.46(\text{亿元})$$

全部国内生产总值的平均数

$$\bar{b} = \frac{\sum_{i=1}^n b_i}{n} = \frac{327447.3}{5} = 65489.46(\text{亿元})$$

第三产业国内生产总值所占平均比重

$$\bar{Y} = \frac{\bar{a}}{\bar{b}} = \frac{20688.46}{65489.46} \times 100\% = 31.59\%$$

增长量 (概念要点)

1. 报告期水平与基期水平之差，说明现象在观察期内增长的绝对数量
2. 有逐期增长量与累积增长量之分
 - 逐期增长量
 - 报告期水平与上一期水平之差
 - 计算形式为： $\Delta_i = Y_i - Y_{i-1} \quad (i=1,2,\dots,n)$
 - 累积增长量
 - 报告期水平与某一固定时期水平之差
 - 计算形式为： $\Delta_i = Y_i - Y_0 \quad (i=1,2,\dots,n)$
3. 各逐期增长量之和等于最末期的累积增长量

平均增长量 (概念要点)

1. 观察期内各逐期增长量的平均数
2. 描述现象在观察期内平均增长的数量
3. 计算公式为

$$\begin{aligned}\text{平均增长量} &= \frac{\text{逐期增长量之和}}{\text{逐期增长量个数}} \\ &= \frac{\text{累积增长量}}{\text{观察值个数}-1}\end{aligned}$$

时间序列的速度分析

发展速度 (要点)

1. 报告期水平与基期水平之比
2. 说明现象在观察期内相对的发展变化程度
3. 有环比发展速度与定期发展速度之分

环比发展速度与定基发展速度 (要点)

1. 环比发展速度

- 报告期水平与上一期水平之比

$$R_i = \frac{Y_i}{Y_{i-1}} \quad (i = 1, 2, \dots, n)$$

2. 定基发展速度

- 报告期水平与某一固定时期水平之比

$$R_i = \frac{Y_i}{Y_0} \quad (i = 1, 2, \dots, n)$$

环比发展速度与定基发展速度 (关系)

1. 观察期内各环比发展速度的连乘积等于最末期的定基发展速度

$$\prod \frac{Y_i}{Y_{i-1}} = \frac{Y_n}{Y_0} \quad \prod \text{为连乘符号}$$

2. 两个相邻的定基发展速度，用后者除以前者，等于相应的环比发展速度

$$\frac{Y_i}{Y_0} \div \frac{Y_{i-1}}{Y_0} = \frac{Y_i}{Y_{i-1}}$$

增长速度 (要点)

1. 增长量与基期水平之比
2. 又称增长率
3. 说明现象的相对增长程度
4. 有环比增长速度与定期增长速度之分
5. 计算公式为

$$\begin{aligned}\text{增长速度} &= \frac{\text{增长量}}{\text{基期水平}} = \frac{\text{报告期水平} - \text{基期水平}}{\text{基期水平}} \\ &= \text{发展速度} - 1\end{aligned}$$

环比增长速度与定基增长速度 (要点)

1. 环比增长速度基

- 报告期水平与前一时期水平之比

$$G_i = \frac{Y_i - Y_{i-1}}{Y_{i-1}} = \frac{Y_i}{Y_{i-1}} - 1 \quad (i = 1, 2, \dots, n)$$

2. 定基增长速度

- 报告期水平与某一固定时期水平之比

$$G_i = \frac{Y_i - Y_0}{Y_0} = \frac{Y_i}{Y_0} - 1 \quad (i = 1, 2, \dots, n)$$

发展速度与增长速度的计算 (实例)

【例11.5】 根据表11-3中第三产业国内生产总值序列，计算各年的环比发展速度和增长速度，及以1994年为基期的定基发展速度和增长速度

表11-4 第三产业国内生产总值速度计算表

年 份		1994	1995	1996	1997	1998
国内生产总值(亿元)		14930.0	17947.2	20427.5	24033.3	26104.3
发展速度 (%)	环比	—	120.2	113.8	117.7	108.6
	定基	100	120.2	136.8	161.0	174.8
增长速度 (%)	环比	—	20.2	13.8	17.7	8.6
	定基	—	20.2	36.8	61.0	74.8

平均发展速度 (要点)

1. 观察期内各环比发展速度的平均数
2. 说明现象在整个观察期内平均发展变化的程度
3. 通常采用几何法(水平法)计算
4. 计算公式为

$$\begin{aligned}\bar{R} &= \sqrt[n]{\frac{Y_1}{Y_0} \times \frac{Y_2}{Y_1} \times \cdots \times \frac{Y_n}{Y_{n-1}}} = \sqrt[n]{\prod_{i=1}^n \frac{Y_i}{Y_{i-1}}} \\ &= \sqrt[n]{\frac{Y_n}{Y_0}} \quad (i = 1, 2, \cdots, n)\end{aligned}$$

平均发展速度与平均增长速度 (算例)

【例11.6】 根据表11.4中的有关数据，计算1994～1998年间我国第三产业国内生产总值的年平均发展速度和年平均增长率

➡ 平均发展速度

$$\begin{aligned}\bar{R} &= \sqrt[n]{\prod_{i=1}^n \frac{Y_i}{Y_{i-1}}} = \sqrt[4]{120.2\% \times 113.8\% \times 117.7\% \times 108.6\%} \\ &= \sqrt[4]{\frac{26104.3}{14930.0}} = 114.99\%\end{aligned}$$

➡ 平均增率

$$\bar{G} = \bar{R} - 1 = 114.99\% - 1 = 14.99\%$$

平均发展速度 (几何法的特点)

1. 从最初水平 Y_0 出发，每期按平均发展速度发展，经过 n 期后将达到最末期水平 Y_n
2. 按平均发展速度推算的最后一期的数值与最后一期的实际观察值一致
3. 只与序列的最初观察值 Y_0 和最末观察值 Y_n 有关
4. 如果关心现象在最后一期应达到的水平，采用水平法计算平均发展速度比较合适

年度化增长率 (要点)

1. 增长率以年来表示时，称为年度化增长率或年率
2. 可将月度增长率或季度增长率转换为年度增长率
3. 计算公式为

$$G_A = \left(\frac{Y_i}{Y_{i-1}} \right)^{\frac{m}{n}} - 1$$

- m 为一年中的时期个数； n 为所跨的时期总数
- 季度增长率被年度化时， $m = 4$
- 月增长率被年度化时， $m = 12$
- 当 $m = n$ 时，上述公式就是年增长率

年度化增长率 (实例)

【例11.7】 已知某地区的如下数据，计算年度化增长率

- 1) 1999年1月份的社会商品零售总额为25亿元， 2000年1月份在零售总额为30亿元
- 2) 1998年3月份财政收入总额为240亿元， 2000年6月份的财政收入总额为为300亿元
- 3) 2000年1季度完成的国内生产总值为500亿元， 2季度完成的国内生产总值为510亿元
- 4) 1997年1季度完成的国内生产总值为500亿元， 2季度完成的国内生产总值为510亿元

年度化增长率 (计算结果)

解：

- 1) 由于是月份数据，所以 **$m=12$** ；从1999年一月到2000年一月所跨的月份总数为12，所以 **$n=12$**

$$G_A = \left(\frac{30}{25} \right)^{\frac{12}{12}} - 1 = 20\%$$

即年度化增长率为**20%**，这实际上就是年增长率，因为所跨的时期总数为一年。也就是该地区社会商品零售总额的年增长率为**20%**

年度化增长率 (计算结果)

解：

2) $m=12, n=27$

年度化增长率为

$$G_A = \left(\frac{300}{240} \right)^{\frac{12}{27}} - 1 = 10.43\%$$

该地区财政收入的年增长率为**10.43%**

年度化增长率 (计算结果)

解：

- 3) 由于是季度数据，所以 $m = 4$ ，从一季度到二季度所跨的时期总数为1，所以 $n=1$

年度化增长率为

$$G_A = \left(\frac{510}{500} \right)^{\frac{4}{1}} - 1 = 8.24\%$$

即根据第一季度和第二季度数据计算的国内生产总值年增长率为**8.24%**

年度化增长率 (计算结果)

解：

- 4) **$m=4$** ，从1997年四季度到2000年四季度所跨的季度总数为12，所以 **$n=12$**

年度化增长率为

$$G_A = \left(\frac{350}{280} \right)^{\frac{4}{12}} - 1 = 7.72\%$$

即根据1998年四季度到2000年四季度的数据计算，工业增加值的年增长率为**7.72%**，这实际上就是工业增加值的年平均增长速度

速度的分析与应用 (需要注意的问题)

1. 当时间序列中的观察值出现0或负数时，不宜计算速度
2. 例如：假定某企业连续五年的利润额分别为5、2、0、-3、2万元，对这一序列计算速度，要么不符合数学公理，要么无法解释其实际意义。在这种情况下，适宜直接用绝对数进行分析
3. 在有些情况下，不能单纯就速度论速度，要注意速度与绝对水平的结合分析

速度的分析与应用 (一个例子)

【例11.8】 假定有两个生产条件基本相同的企业，各年的利润额及有关的速度值如表11-5

表11-5 甲、乙两个企业的有关资料				
年 份	甲 企 业		乙 企 业	
	利润额(万元)	增长率(%)	利润额(万元)	增长率(%)
1996	500	—	60	—
1997	600	20	84	40

速度的分析与应用 (增长1%绝对值)

1. 速度每增长一个百分点而增加的绝对量
2. 用于弥补速度分析中的局限性
3. 计算公式为

$$\text{增长1\%绝对值} = \frac{\text{逐期增长量}}{\text{环比增长速度} \times 100} = \frac{\text{前期水平}}{100}$$

甲企业增长1%绝对值 = $500/100 = 5$ 万元

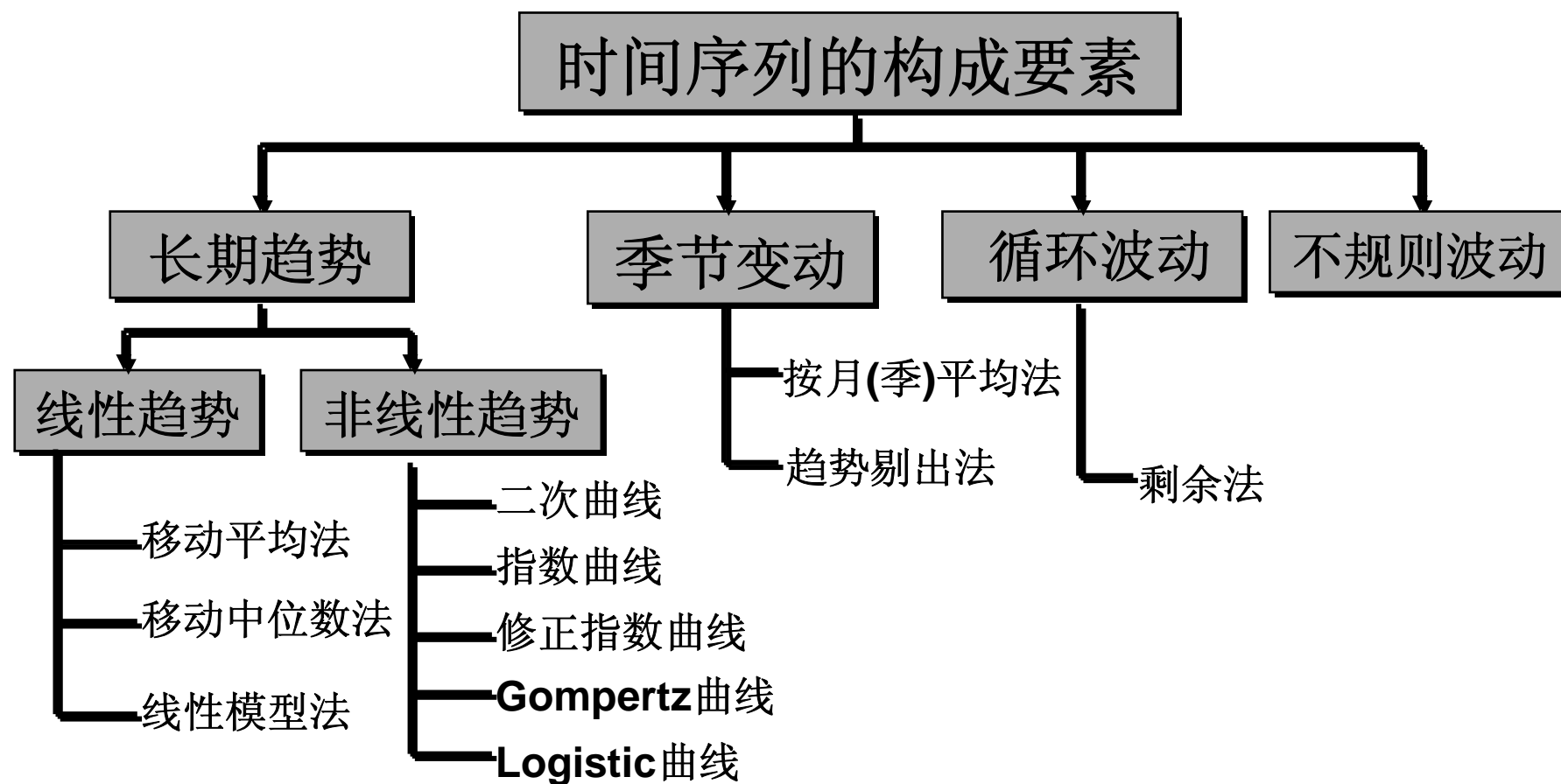
乙企业增长1%绝对值 = $60/100 = 0.6$ 万元

第二节 长期趋势分析

- 一. 时间序列的构成要素与模型
- 二. 线性趋势
- 三. 非线性趋势
- 四. 趋势线的选择

时间序列的构成要素与模型

(构成要素与测定方法)



时间序列的构成要素与模型 (要点)

1. 构成因素

- 长期趋势 (Secular trend)
- 季节变动 (Seasonal Fluctuation)
- 循环波动 (Cyclical Movement)
- 不规则波动 (Irregular Variations)

2. 模型

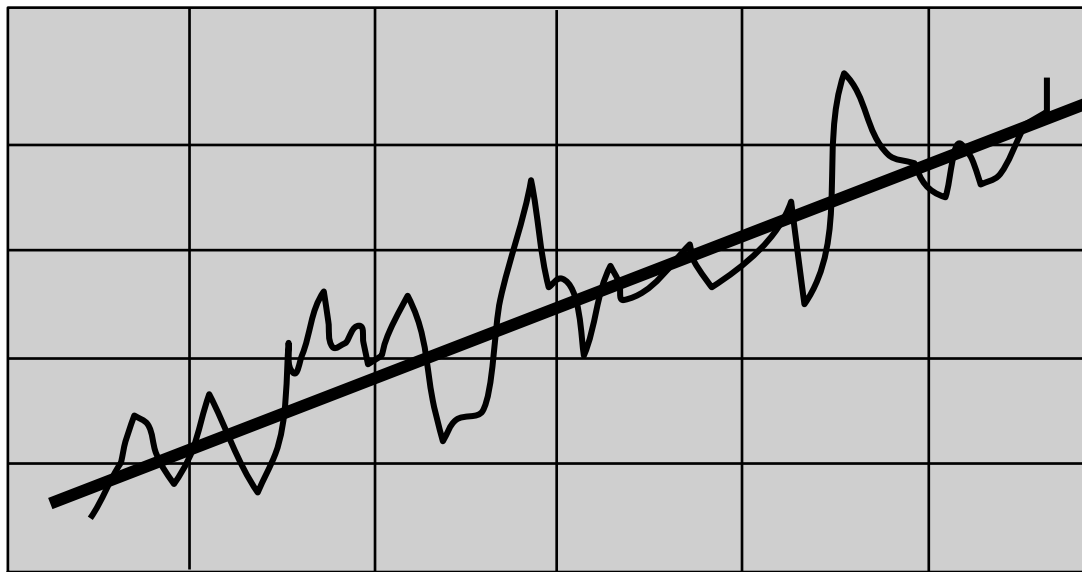
- 乘法模型: $Y_i = T_i \times S_i \times C_i \times I_i$
- 加法模型: $Y_i = T_i + S_i + C_i + I_i$

长期趋势 (概念要点)

1. 现象在较长时期内持续发展变化的一种趋向或状态
2. 由影响时间序列的基本因素作用形成
3. 时间序列的主要构成要素
4. 有线性趋势和非线性趋势



线性趋势



线性趋势

1. 现象随时间的推移呈现出稳定增长或下降的线性变化规律
2. 测定方法有
 - 移动平均法
 - 移动中位数法
 - 线性模型法

移动平均法

(Moving Average Method)

1. 测定长期趋势的一种较简单的常用方法
 - 通过扩大原时间序列的时间间隔，并按一定的间隔长度逐期移动，计算出一系列移动平均数
 - 由移动平均数形成的新的时间序列对原时间序列的波动起到修匀作用，从而呈现出现象发展的变动趋势
2. 移动步长为 $K(1 < K < n)$ 的移动平均序列为

$$\bar{Y}_i = \frac{Y_i + Y_{i+1} + \cdots + Y_{K+i-1}}{K}$$

移动平均法 (实例)

【例 11.9】 已知1981~1998年我国汽车产量数据如表11-6。分别计算三年和五年移动平均趋势值，以及三项和五项移动中位数，并作图与原序列比较

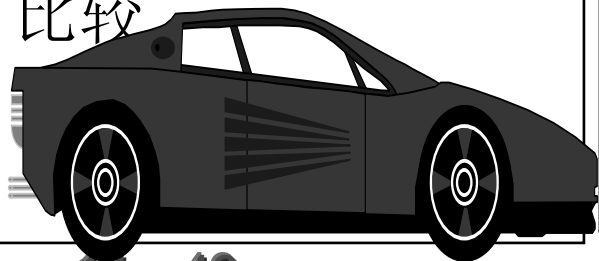


表11-6 1981~1998年我国汽车产量数据

年 份	产量(万辆)	年份	产量(万辆)
1981	17.56	1990	51.40
1982	19.63	1991	71.42
1983	23.98	1992	106.67
1984	31.64	1993	129.85
1985	43.72	1994	136.69
1986	36.98	1995	145.27
1987	47.18	1996	147.52
1988	64.47	1997	158.25
1989	58.35	1998	163.00

移动平均法 (趋势图)

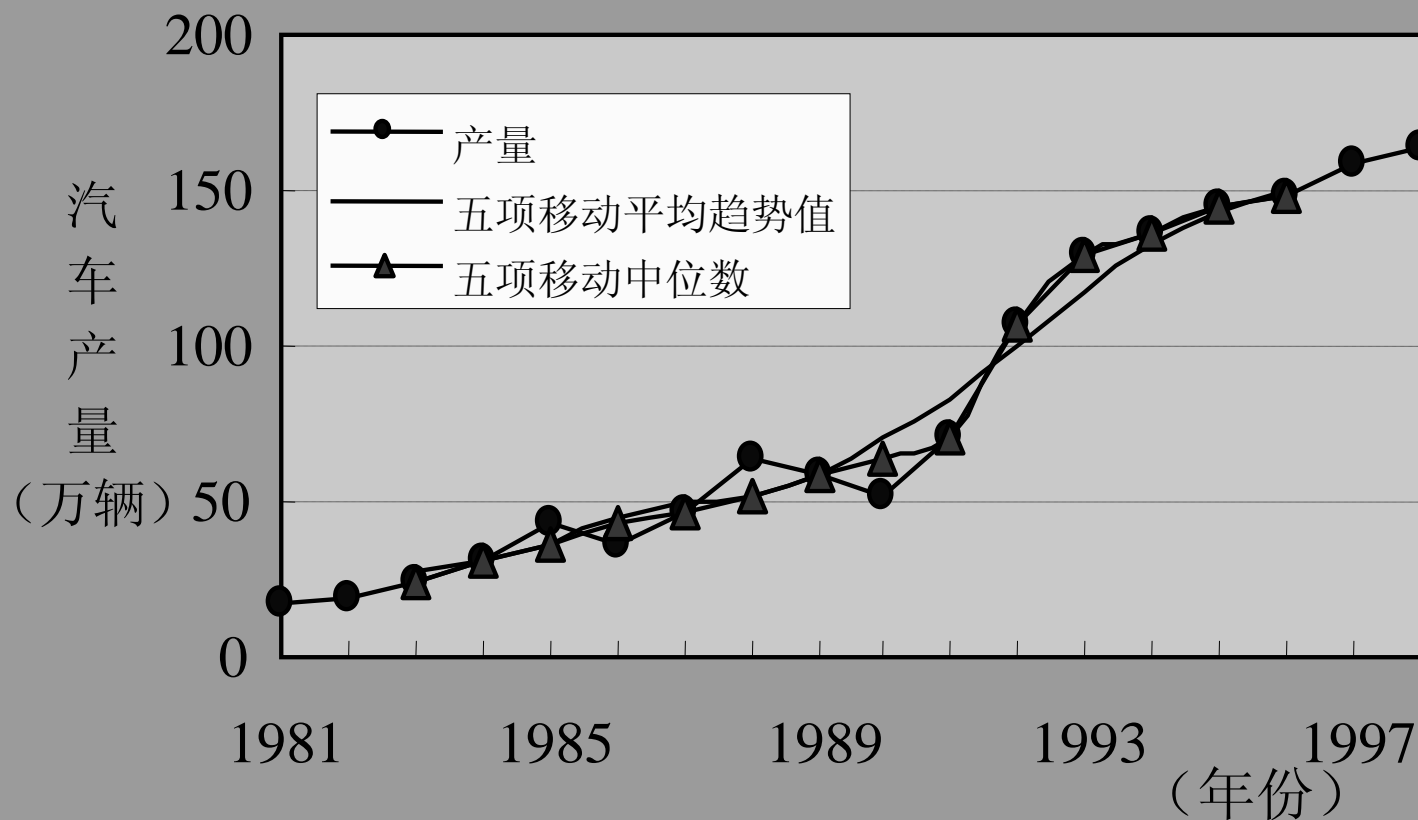


图11-1 汽车产量移动平均趋势图

移动平均法 (应注意的问题)

1. 移动平均后的趋势值应放在各移动项的中间位置
 - 对于偶数项移动平均需要进行“中心化”
2. 移动间隔的长度应长短适中
 - 如果现象的发展具有一定的周期性，应以周期长度作为移动间隔的长度
 - 若时间序列是季度资料，应采用4项移动平均
 - 若为月份资料，应采用12项移动平均

线性模型法

（概念要点与基本形式）

1. 现象的发展按线性趋势变化时，可用线性模型表示
2. 线性模型的形式为

$$\hat{Y}_t = a + bt$$

- \hat{Y}_t —时间序列的趋势值
- t —时间标号
- a —趋势线在 Y 轴上的截距
- b —趋势线的斜率，表示时间 t 变动一个单位时观察值的平均变动数量

线性模型法

(a 和 b 的最小二乘估计)

1. 趋势方程中的两个未知常数 a 和 b 按最小二乘法(Least-square Method)求得
 - 根据回归分析中的最小二乘法原理
 - 使各实际观察值与趋势值的离差平方和为最小
 - 最小二乘法既可以配合趋势直线，也可用于配合趋势曲线
2. 根据趋势线计算出各个时期的趋势值

线性模型法

(a 和 b 的最小二乘估计)

1. 根据最小二乘法得到求解 a 和 b 的标准方程为

$$\begin{cases} \sum Y = na + b \sum t \\ \sum tY = a \sum t + b \sum t^2 \end{cases} \quad \text{解得:} \quad \begin{cases} b = \frac{n \sum tY - \sum t \sum Y}{n \sum t^2 - (\sum t)^2} \\ a = \bar{Y} - b\bar{t} \end{cases}$$

2. 取时间序列的中间时期为原点时有 $\sum t=0$ ，上式可化简为

$$\begin{cases} \sum Y = na \\ \sum tY = b \sum t^2 \end{cases} \quad \text{解得:} \quad \begin{cases} a = \bar{Y} \\ b = \frac{\sum tY}{\sum t^2} \end{cases}$$

线性模型法 (实例及计算过程)

【例 11.10】 利用表11-6中的数据，根据最小二乘法确定汽车产量的直线趋势方程，计算出1981～1998年各年汽车产量的趋势值，并预测2000年的汽车产量，作图与原序列比较



表11-8 汽车产量直线趋势计算表

年份	时间标号 t	产量(万辆) Y_t	$t \times Y_t$	t^2	趋势值
1981	1	17.56	17.56	1	0.00
1982	2	19.63	39.26	4	9.50
1983	3	23.98	71.94	9	19.00
1984	4	31.64	126.56	16	28.50
1985	5	43.72	218.60	25	38.00
1986	6	36.98	221.88	36	47.50
1987	7	47.18	330.26	49	57.00
1988	8	64.47	515.76	64	66.50
1989	9	58.35	525.15	81	76.00
1990	10	51.40	514.00	100	85.50
1991	11	71.42	785.62	121	95.00
1992	12	106.67	1280.04	144	104.51
1993	13	129.85	1688.05	169	114.01
1994	14	136.69	1913.66	196	123.51
1995	15	145.27	2179.05	225	133.01
1996	16	147.52	2360.32	256	142.51
1997	17	158.25	2690.25	289	152.01
1998	18	163.00	2934.00	324	161.51
合计	171	1453.58	18411.96	2109	1453.58

线性模型法 (计算结果)

根据上表得 a 和 b 结果如下

$$\begin{cases} b = \frac{18 \times 18411.96 - 171 \times 1453.58}{18 \times 2109 - (171)^2} = 9.5004 \\ a = \frac{1453.58}{18} - 9.5004 \times \frac{171}{18} = -9.4995 \end{cases}$$

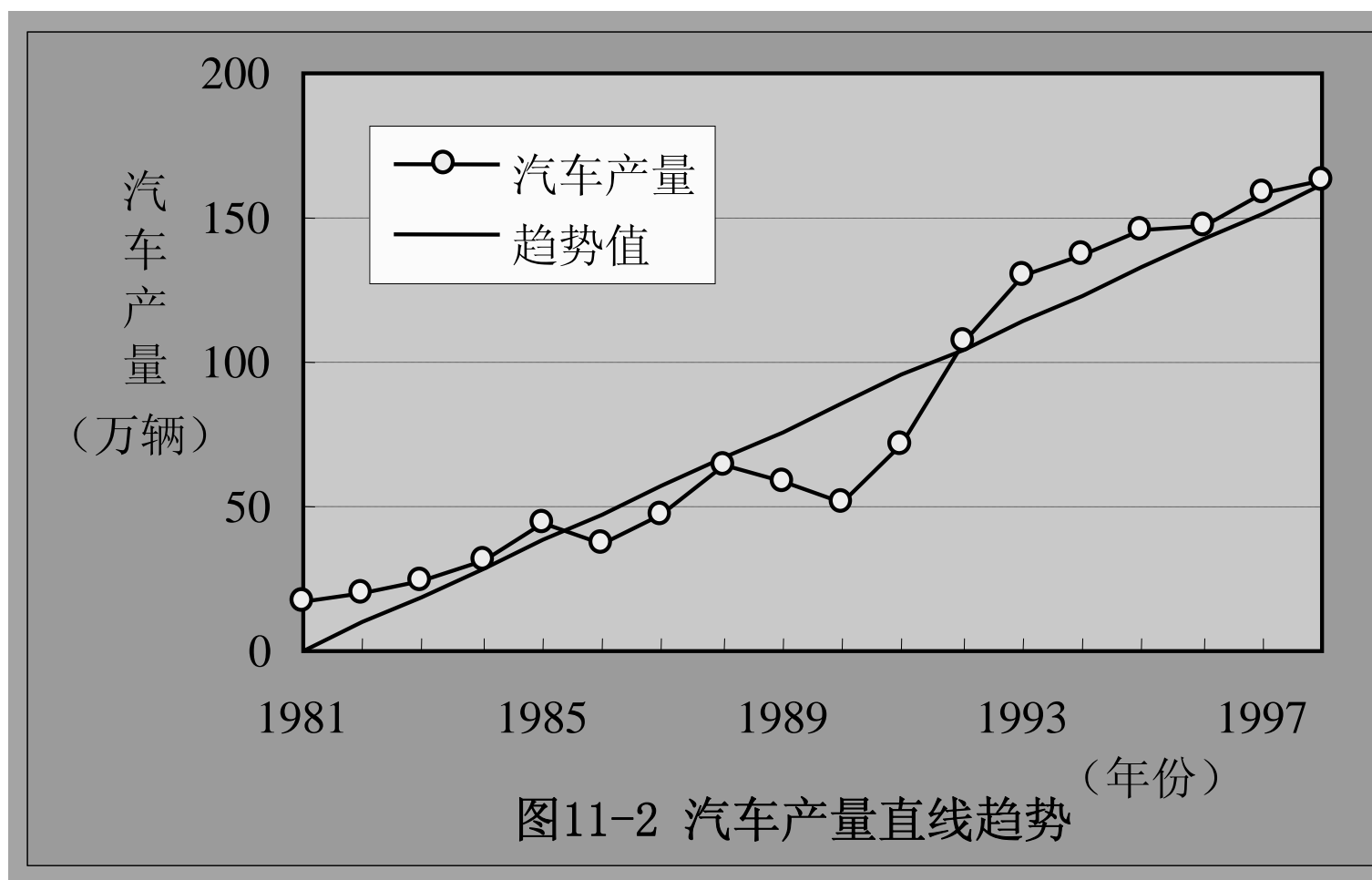
汽车产量的直线趋势方程为

$$\hat{Y}_t = -9.4995 + 9.5004 t$$

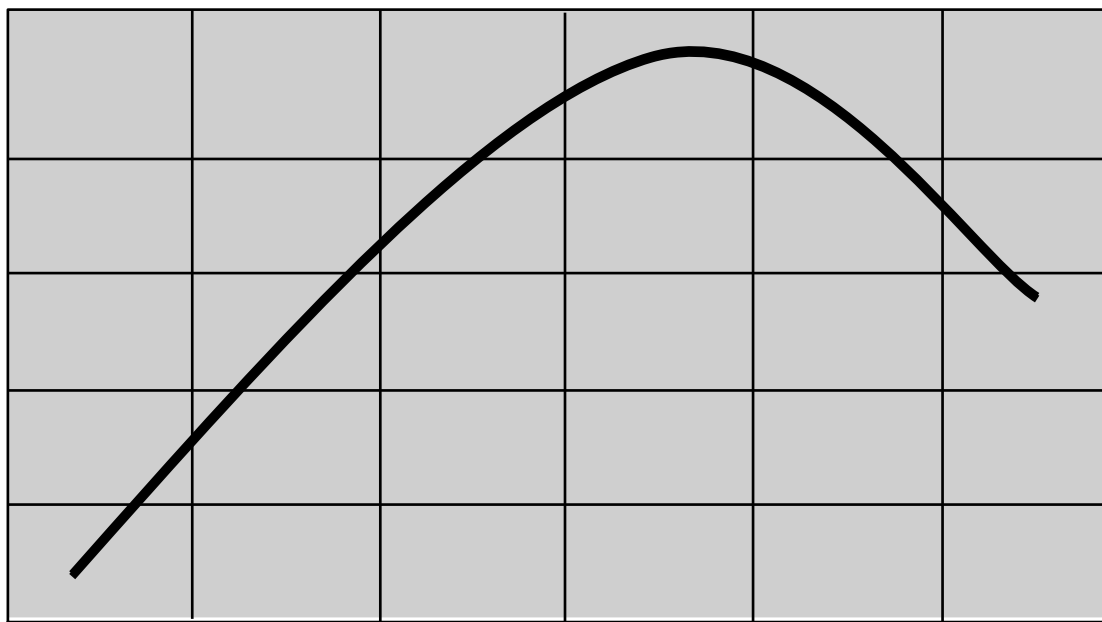
2000年汽车产量的预测值为

$$\hat{Y}_{2000} = -9.4995 + 9.5004 \times 20 = 180.51 \text{ (万辆)}$$

线性模型法 (趋势图)



非线性趋势



二次曲线 (Second Degree Curve)

1. 现象的发展趋势为抛物线形态
2. 一般形式为

$$\hat{Y}_t = a + bt + ct^2$$

- a 、 b 、 c 为未知常数
- 根据最小二乘法求得

二次曲线 (Second Degree Curve)

1. 根据最小二乘法得到求解 a 、 b 、 c 的标准方程为

$$\begin{cases} \sum Y = na + b \sum t + c \sum t^2 \\ \sum tY = a \sum t + b \sum t^2 + c \sum t^3 \\ \sum t^2 Y = a \sum t^2 + b \sum t^3 + c \sum t^4 \end{cases}$$

2. 取时间序列的中间时期为原点时有

$$\begin{cases} \sum Y = na + c \sum t^2 \\ \sum tY = b \sum t^2 \\ \sum t^2 Y = a \sum t^2 + c \sum t^4 \end{cases}$$

二次曲线 (实例)

【例 11.11】 已知我国 1978 ~ 1992 年针织内衣零售量数据如表 11-9。试配合二次曲线，计算出 1978 ~ 1992 年零售量的趋势值，并预测 1993 年的零售量，作图与原序列比较

表11- 9 1978~1992年针织内衣零售量

年 份	零售量(亿件)	年 份	零售量(亿件)
1978	7.0	1986	14.4
1979	9.1	1987	14.8
1980	9.7	1988	15.0
1981	10.8	1989	12.3
1982	11.7	1990	11.2
1983	12.1	1991	9.4
1984	13.1	1992	8.9
1985	14.3		

二次曲线 (计算过程)

表11-10 针织内衣零售量二次曲线计算表

年份	时间标号 t	零售量(亿件) Y_t	$t \times Y_t$	t^2	$t^2 Y_t$	t^4	趋势值
1978	-7	7.0	-49.0	49	343.0	2401	6.5
1979	-6	9.1	-54.6	36	327.6	1296	8.4
1980	-5	9.7	-48.5	25	242.5	625	10.0
1981	-4	10.8	-43.2	16	172.8	256	11.3
1982	-3	11.7	-35.1	9	105.3	81	12.3
1983	-2	12.1	-24.2	4	48.4	16	13.2
1984	-1	13.1	-13.1	1	13.1	1	13.7
1985	0	14.3	0	0	0	0	14.0
1986	1	14.4	14.4	1	14.4	1	14.0
1987	2	14.8	29.6	4	59.2	16	13.8
1988	3	15.0	45.0	9	135.0	81	13.3
1989	4	12.3	49.2	16	196.8	256	12.6
1990	5	11.2	56.0	25	280.0	625	11.6
1991	6	9.4	56.4	36	338.4	1296	10.3
1992	7	8.9	62.3	49	436.1	2401	8.8
合计	0	173.8	45.2	280	2712.6	9352	173.8

二次曲线 (计算结果)

根据计算表得 a 、 b 、 c 的结果如下

$$\begin{cases} 173.8 = 15a + 280c \\ 45.2 = 280b \\ 2712.6 = 280a + 9352c \end{cases} \quad \begin{cases} a = 13.9924 \\ b = 0.16143 \\ c = -0.128878 \end{cases}$$

针织内衣零售量的二次曲线方程为

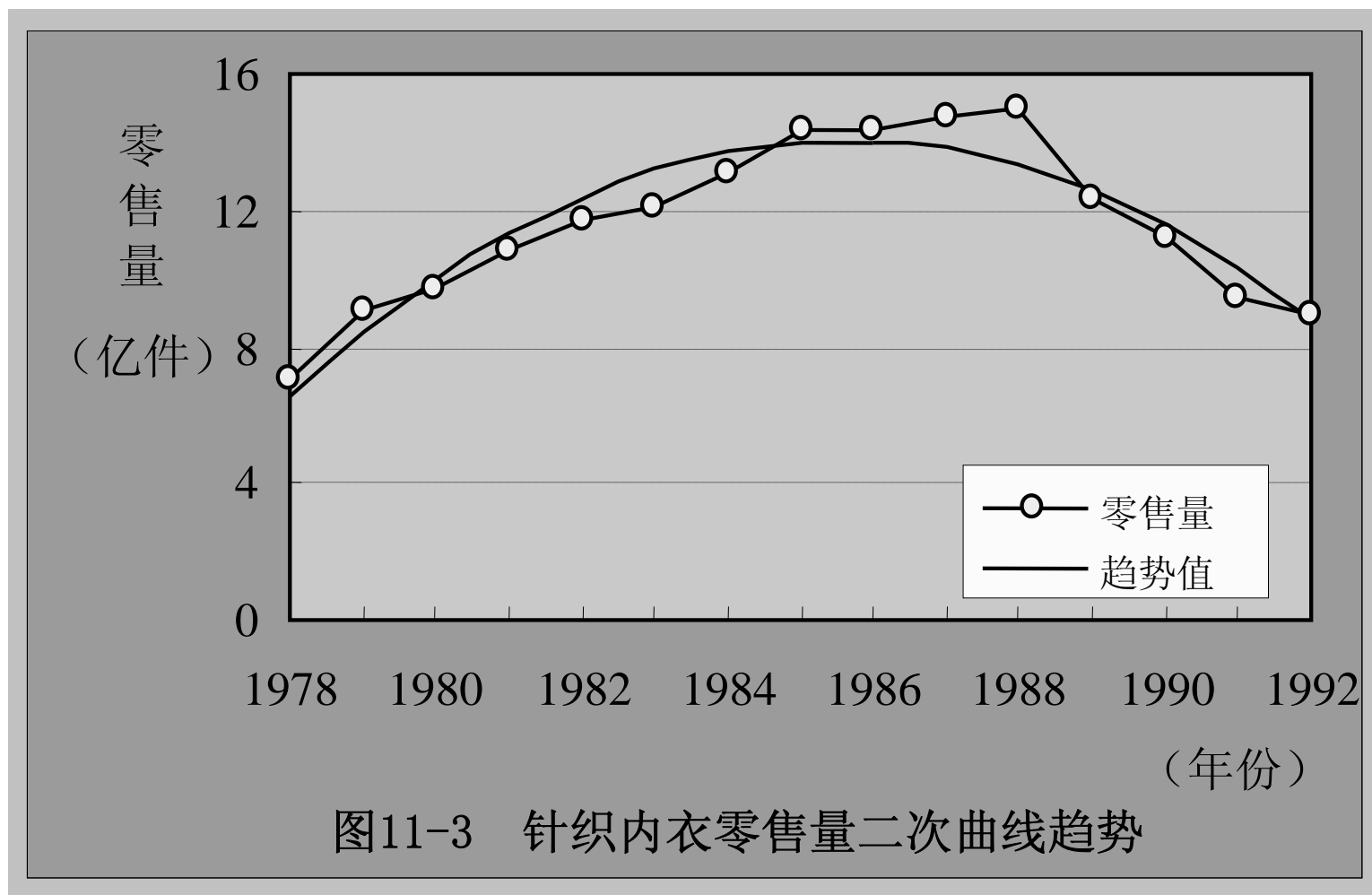
$$\hat{Y}_t = 13.9924 + 0.16143 t - 0.128878 t^2$$

1993年零售量的预测值为

$$\hat{Y}_{1993} = 13.9924 + 0.16143 \times 8 - 0.128878 \times 8^2$$

$$= 7.03 \text{ (亿件)}$$

二次曲线 (趋势图)



指数曲线 (Exponential curve)

1. 用于描述以几何级数递增或递减的现象
2. 一般形式为

$$\hat{Y}_t = ab^t$$

- a 、 b 为未知常数
- 若 $b>1$ ，增长率随着时间 t 的增加而增加
- 若 $b<1$ ，增长率随着时间 t 的增加而降低
- 若 $a>0$ ， $b<1$ ，趋势值逐渐降低到以0为极限

指数曲线

(a 、 b 的求解方法)

1. 采取“线性化”手段将其化为对数直线形式
2. 根据最小二乘法，得到求解 $\lg a$ 、 $\lg b$ 的标准方程为

$$\begin{cases} \sum \lg Y = n \lg a + \lg b \sum t \\ \sum t \lg Y = \lg a \sum t + \lg b \sum t^2 \end{cases}$$

3. 取时间序列的中间时期为原点，上式可化简为

$$\begin{cases} \sum \lg Y = n \lg a \\ \sum t \lg Y = \lg b \sum t^2 \end{cases}$$

指数曲线 (实例及计算结果)

【例11.12】根据表11-6中的资料，确定1981~1998年我国汽车产量的指数曲线方程，求出各年汽车产量的趋势值，并预测2000年的汽车产量，作图与原序列比较

$$\begin{cases} 32.459896 = 18 \lg a + 171 \lg b \\ 337.223286 = 171 \lg a + 2109 \lg b \end{cases} \quad \begin{cases} a = 17.2805 \\ b = 1.14698 \end{cases}$$

汽车产量的指数曲线方程为

$$\hat{Y}_t = 17.2805 \times (1.14698)^t$$

2000年汽车产量的预测值为

$$\hat{Y}_{2000} = 17.2805 \times (1.14698)^{20} = 268.33 (\text{万辆})$$

指数曲线 (趋势图)

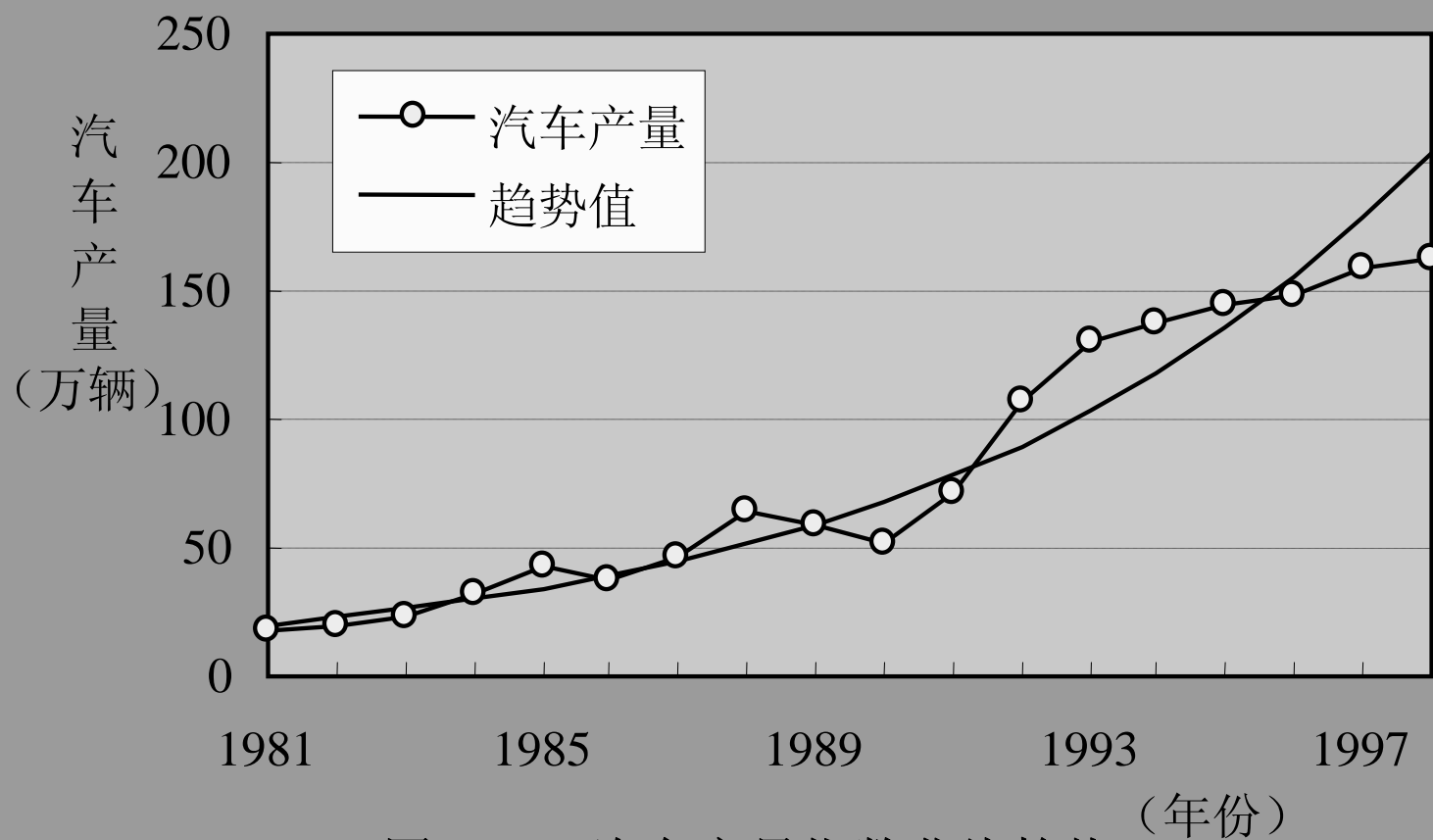


图11-4 汽车产量指数曲线趋势

指数曲线与直线的比较

1. 比一般的趋势直线有着更广泛的应用
2. 可以反应出现象的相对发展变化程度
 - 上例中， $b=1.14698$ 表示1981～1998年汽车产量趋势值的平均发展速度
3. 不同序列的指数曲线可以进行比较
 - 比较分析相对增长程度

修正指数曲线 (Modified exponential curve)

1. 在一般指数曲线的基础上增加一个常数 K
2. 一般形式为

$$\hat{Y}_t = K + ab^t$$

- K 、 a 、 b 为未知常数
 - $K > 0$, $a \neq 0$, $0 < b \neq 1$
3. 修正指数曲线用于描述的现象：初期增长迅速，随后增长率逐渐降低，最终则以 K 为增长极限

修正指数曲线

(求解 k 、 a 、 b 的三和法)

1. 趋势值 K 无法事先确定时采用
2. 将时间序列观察值等分为三个部分，每部分有 m 个时期
3. 令趋势值的三个局部总和分别等于原序列观察值的三个局部总和

修正指数曲线 (求解 k 、 a 、 b 的三和法)

1. 设观察值的三个局部总和分别为 S_1 , S_2 , S_3

$$S_1 = \sum_{t=1}^m Y_t, \quad S_2 = \sum_{t=m+1}^{2m} Y_t, \quad S_3 = \sum_{t=2m+1}^{3m} Y_t$$

2. 根据三和法求得

$$\begin{cases} b = \left(\frac{S_3 - S_2}{S_2 - S_1} \right)^{\frac{1}{m}} \\ a = (S_2 - S_1) \frac{b - 1}{b(b^m - 1)^2} \\ K = \frac{1}{m} \left(S_1 - \frac{ab(b^m - 1)}{b - 1} \right) \end{cases}$$

修正指数曲线 (实例)

【例11.13】 已知1978~1995年我国小麦单位面积产量的数据如表11-12。试确定小麦单位面积产量的修正指数曲线方程，求出各年单位面积产量的趋势值，并预测2000年的小麦单位面积产量，作图与原序列比较

表11- 12 1978~1995年小麦单位面积产量数据

年 份	单位面积产量 (公斤/公顷)	年 份	单位面积产量 (公斤/公顷)
1978	1845	1987	2985
1979	2145	1988	2970
1980	1890	1989	3045
1981	2115	1990	3195
1982	2445	1991	3105
1983	2805	1992	3331
1984	2970	1993	3519
1985	2940	1994	3426
1986	3045	1995	3542



修正指数曲线 (计算结果)

解得 K 、 a 、 b 如下

$$\begin{cases} b = \left(\frac{20118 - 17955}{17955 - 13245} \right)^{\frac{1}{6}} = 0.87836 \\ a = (17955 - 13245) \frac{0.87836 - 1}{0.87836(0.87836^6 - 1)^2} = 2230.531 \\ K = \frac{1}{6} \left(13245 - \frac{-2230.531 \times 0.87836(0.87836^6 - 1)}{0.87836 - 1} \right) = 3659.149 \end{cases}$$

修正指数曲线 (计算结果)

小麦单位面积产量的修正指数曲线方程为

$$\hat{Y}_t = 3659.149 - 2230.531 \times (0.87836)^t$$

2000年小麦单位面积产量的预测值为

$$\begin{aligned}\hat{Y}_{2000} &= 3659.149 - 2230.531 \times (0.87836)^{23} \\ &= 3546.20 \text{ (kg)}\end{aligned}$$

修正指数曲线 (趋势图)

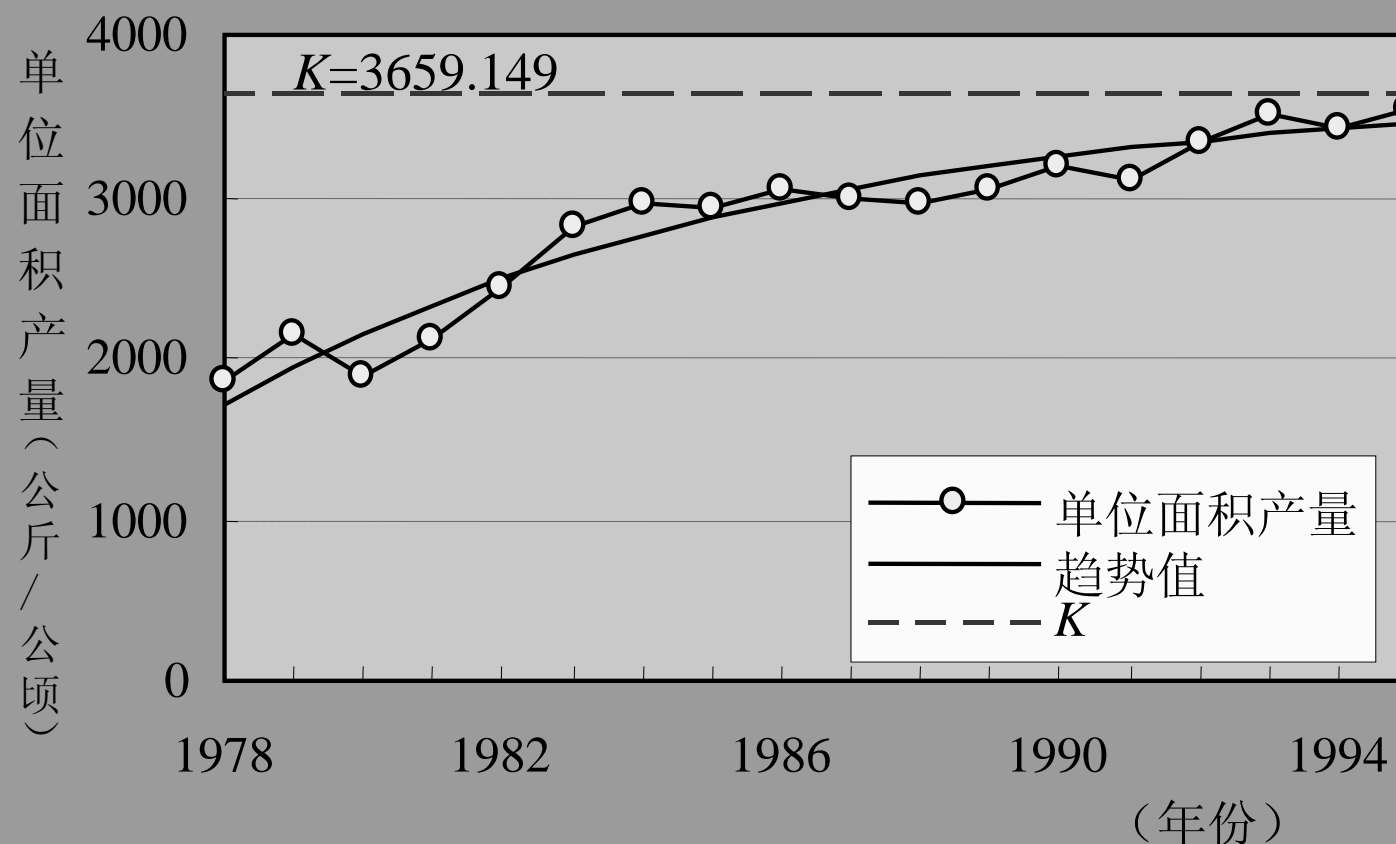


图11-5 小麦单位面积产量修正指数曲线趋势



龚珀茨曲线 (Gompertz curve)

1. 以英国统计学家和数学家 B·Gompertz 而命名
2. 一般形式为

$$\hat{Y}_t = Ka^{b^t}$$

- K 、 a 、 b 为未知常数
 - $K > 0$, $0 < a \neq 1$, $0 < b \neq 1$
3. 所描述的现象：初期增长缓慢，以后逐渐加快，当达到一定程度后，增长率又逐渐下降，最后接近一条水平线
 4. 两端都有渐近线，上渐近线为 $Y \rightarrow K$, 下渐近线为 $Y \rightarrow 0$

Gompertz曲线

(求解 k 、 a 、 b 的三和法)

1. 将其改写为对数形式

$$\lg \hat{Y}_t = \lg K + (\lg a)b^t$$

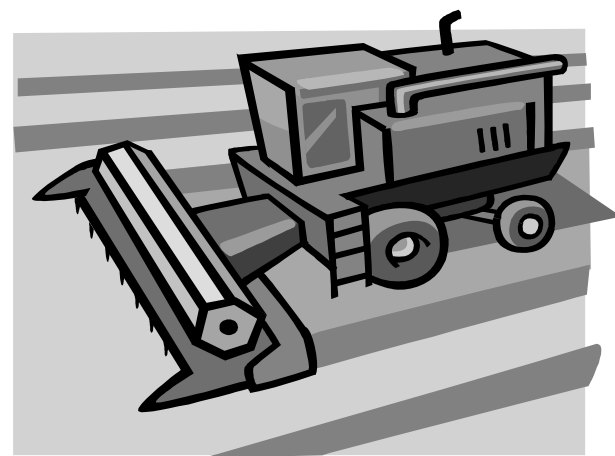
2. 仿照修正指数曲线的常数确定方法, 求出 $\lg a$ 、 $\lg K$ 、 b
3. 取 $\lg a$ 、 $\lg K$ 的反对数求得 a 和 K

令:
$$S_1 = \sum_{t=1}^m \lg Y_t, \quad S_2 = \sum_{t=m+1}^{2m} \lg Y_t, \quad S_3 = \sum_{t=2m+1}^{3m} \lg Y_t$$

则有:
$$\begin{cases} b = \left(\frac{S_3 - S_2}{S_2 - S_1} \right)^{\frac{1}{m}} \\ \lg a = (S_2 - S_1) \frac{b - 1}{b(b^m - 1)^2} \\ \lg K = \frac{1}{m} \left(S_1 - \frac{b(b^m - 1)}{b - 1} \lg a \right) \end{cases}$$

Gompertz 曲线 (实例)

【例11.14】 根据表11-12的数据，试确定小麦单位面积产量的Gompertz曲线方程，求出各年单位面积产量的趋势值，并预测2000年的小麦单位面积产量，作图与原序列比较



Gompertz曲线

(计算结果)

$$\left\{ \begin{array}{l} b = \left(\frac{21.149562 - 20.855979}{20.855979 - 20.035408} \right)^{\frac{1}{6}} = 0.842563 \\ \log a = (20.855979 - 20.035408) \frac{0.842563 - 1}{0.842563 (0.842563^6 - 1)^2} \\ a = 0.427864 \\ \log K = \frac{1}{6} \left(20.035408 - \frac{0.842563 (0.842563^6 - 1)}{0.842563 - 1} \times (-0.371750) \right) \\ K = 3566.04 \end{array} \right.$$

Gompertz 曲线 (计算结果)

小麦单位面积产量的 Gompertz 曲线方程为

$$\hat{Y}_t = 3566.04 \times (0.427864)^{0.842563^t}$$

2000年小麦单位面积产量的预测值为

$$\hat{Y}_{2000} = 3566.04 \times (0.427864)^{0.842563^{23}} = 3507(\text{kg})$$

Gompertz曲线 (趋势图)

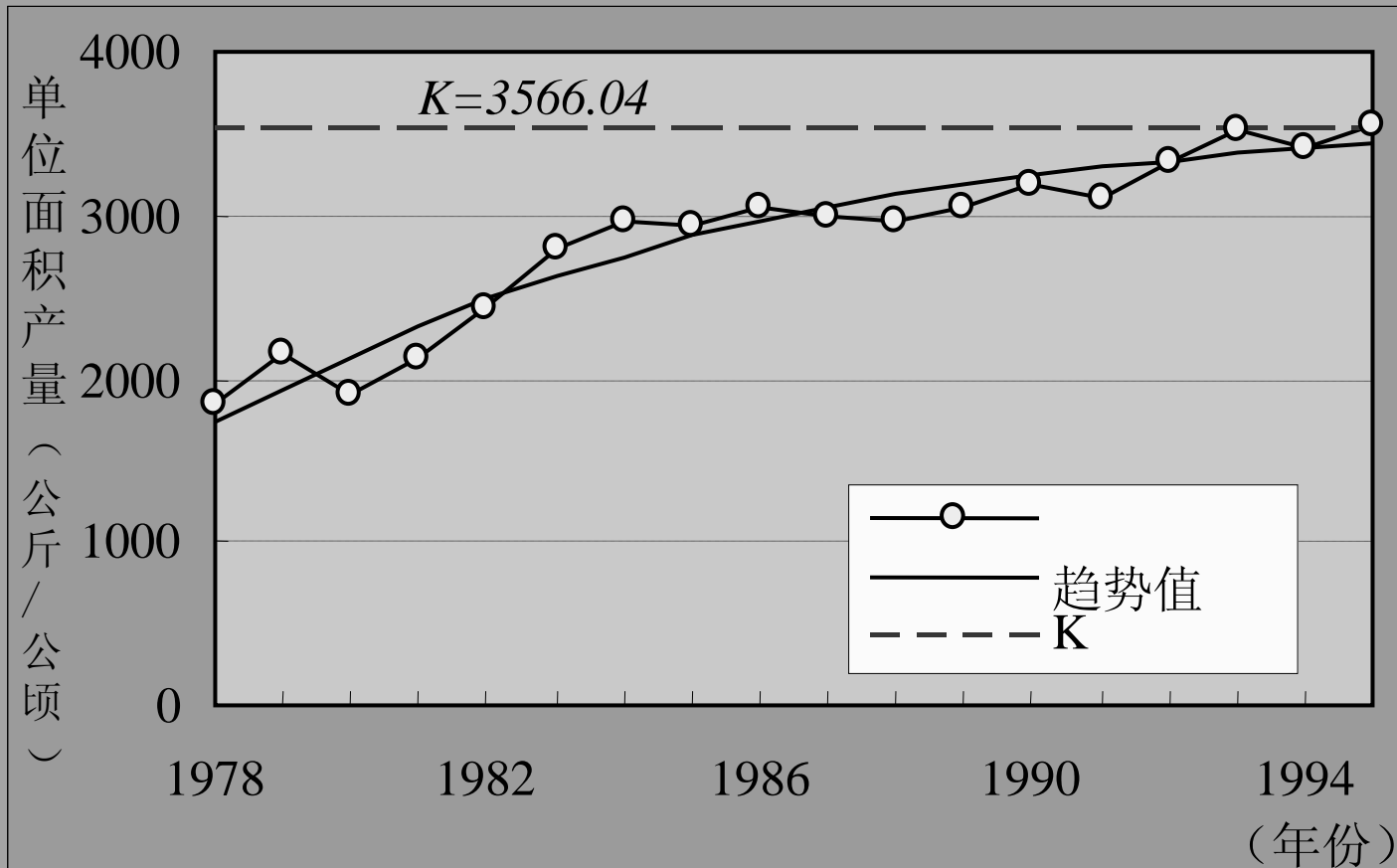


图11-6 小麦单位面积产量Gompertz曲线趋势



罗吉斯蒂曲线 (Logistic Curve)

1. 1838年比利时数学家 Verhulst所确定的名称
2. 该曲线所描述的现象的特征与Gompertz曲线类似
3. 其曲线方程为

$$\hat{Y}_t = \frac{1}{K + ab^t}$$

- K 、 a 、 b 为未知常数
- $K > 0$, $a > 0$, $0 < b \neq 1$

Logistic 曲线

(求解 k 、 a 、 b 的三和法)

1. 取观察值 Y_t 的倒数 Y_t^{-1}

- 当 Y_t^{-1} 很小时, 可乘以 10 的适当次方

2. a 、 b 、 K 的求解方程为

$$\begin{cases} b = \left(\frac{S_3 - S_2}{S_2 - S_1} \right)^{\frac{1}{m}} \\ a = (S_2 - S_1) \frac{b - 1}{b(b^m - 1)^2} \\ K = \frac{1}{m} \left(S_1 - \frac{ab(b^m - 1)}{b - 1} \right) \end{cases}$$

趋势线的选择

1. 观察散点图
2. 根据观察数据本身，按以下标准选择趋势线
 - 一次差大体相同，配合直线
 - 二次差大体相同，配合二次曲线
 - 对数的一次差大体相同，配合指数曲线
 - 一次差的环比值大体相同，配合修正指数曲线
 - 对数一次差的环比值大体相同，配合 Gompertz 曲线
 - 倒数一次差的环比值大体相同，配合 Logistic 曲线
3. 比较估计标准误差

$$S_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - m}}$$

第三节 季节变动分析

- 一. 季节变动及其测定目的
- 二. 季节变动的分析方法与原理
- 三. 季节变动的调整

季节变动及其测定目的

1. 季节变动

- 现象在一年内随着季节更换形成的有规律变动
- 各年变化强度大体相同、且每年重现
- 指任何一种周期性的变化
- 时间序列的又一个主要构成要素

2. 测定目的

- 确定现象过去的季节变化规律
- 消除时间序列中的季节因素

季节变动的分析原理

1. 将季节变动规律归纳为一种典型的季节模型
2. 季节模型由季节指数所组成
3. 季节指数的平均数等于100%
4. 根据季节指数与其平均数(100%)的偏差程度测定季节变动的程度
 - 如果现象没有季节变动，各期的季节指数等于100%
 - 如果某一月份或季度有明显的季节变化，各期的季节指数应大于或小于100%

季节变动的分析原理

☞ 季节模型

- 时间序列在各年中所呈现出的典型状态，这种状态年复一年以相同的形态出现
- 由季节指数组成，各指数刻划了现象在一个年度内各月或季的典型数量特征
- 以各个指数的平均数等于100%为条件而构成
- 如果分析的是月份数据，季节模型就由12个指数组成；若为季度数据，则由4个指数组成

季节变动的分析原理

☞ 季节指数

1. 反映季节变动的相对数
2. 以全年月或季资料的平均数为基础计算的
3. 平均数等于100%
 - 月(或季)的指数之和等于1200%(或400%)
4. 指数越远离其平均数(100%) 季节变动程度越大
5. 计算方法有按月(季)平均法和趋势剔除法

按月(季)平均法 (原理和步骤)

1. 根据原时间序列通过简单平均计算季节指数
2. 假定时间序列没有明显的长期趋势和循环波动
3. 计算季节指数的步骤
 - 计算同月(或同季)的平均数
 - 计算全部数据的总月(总季)平均数
 - 计算季节指数(S)

$$\text{季节指数}(S) = \frac{\text{同月(季)平均数}}{\text{总月(季)平均数}} \times 100\%$$

按月(季)平均法 (实例)

【例 11.15】

已知我国1978～1983年各季度的农业生产资料零售额数据如表11.15。试用按季平均法计算各季的季节指数

表11-15 1978～1983年各季度农业生产资料零售额数据

年 份	销售额(亿元)			
	一季度	二季度	三季度	四季度
1978	62.6	88.0	79.1	64.0
1979	71.5	95.3	88.5	68.7
1980	74.8	106.3	96.4	68.5
1981	75.9	106.0	95.7	69.9
1982	85.2	117.6	107.3	78.4
1983	86.5	131.1	115.4	90.3

按月(季)平均法 (计算表)

表11- 16 农业生产资料零售额季节指数计算表

年 份	销售额(亿元)				
	一季度	二季度	三季度	四季度	全年合计
1978	62.6	88.0	79.1	64.0	293.7
1979	71.5	95.3	88.5	68.7	324.0
1980	74.8	106.3	96.4	68.5	346.0
1981	75.9	106.0	95.7	69.9	347.5
1982	85.2	117.6	107.3	78.4	388.5
1983	86.5	131.1	115.4	90.3	423.3
合计	456.5	644.3	582.4	439.8	2123.0
同季平均	76.08	107.38	97.07	73.30	88.46
季节指数(%)	86.01	121.39	109.73	82.86	100.00

趋势剔除法 (原理和步骤)

1. 先将序列中的趋势予以消除，再计算季节指数
2. 计算季节指数的步骤
 - 计算移动平均趋势值(T)
 - 从序列中剔出趋势值(Y/T)
 - 按前述方法计算季节指数(S)

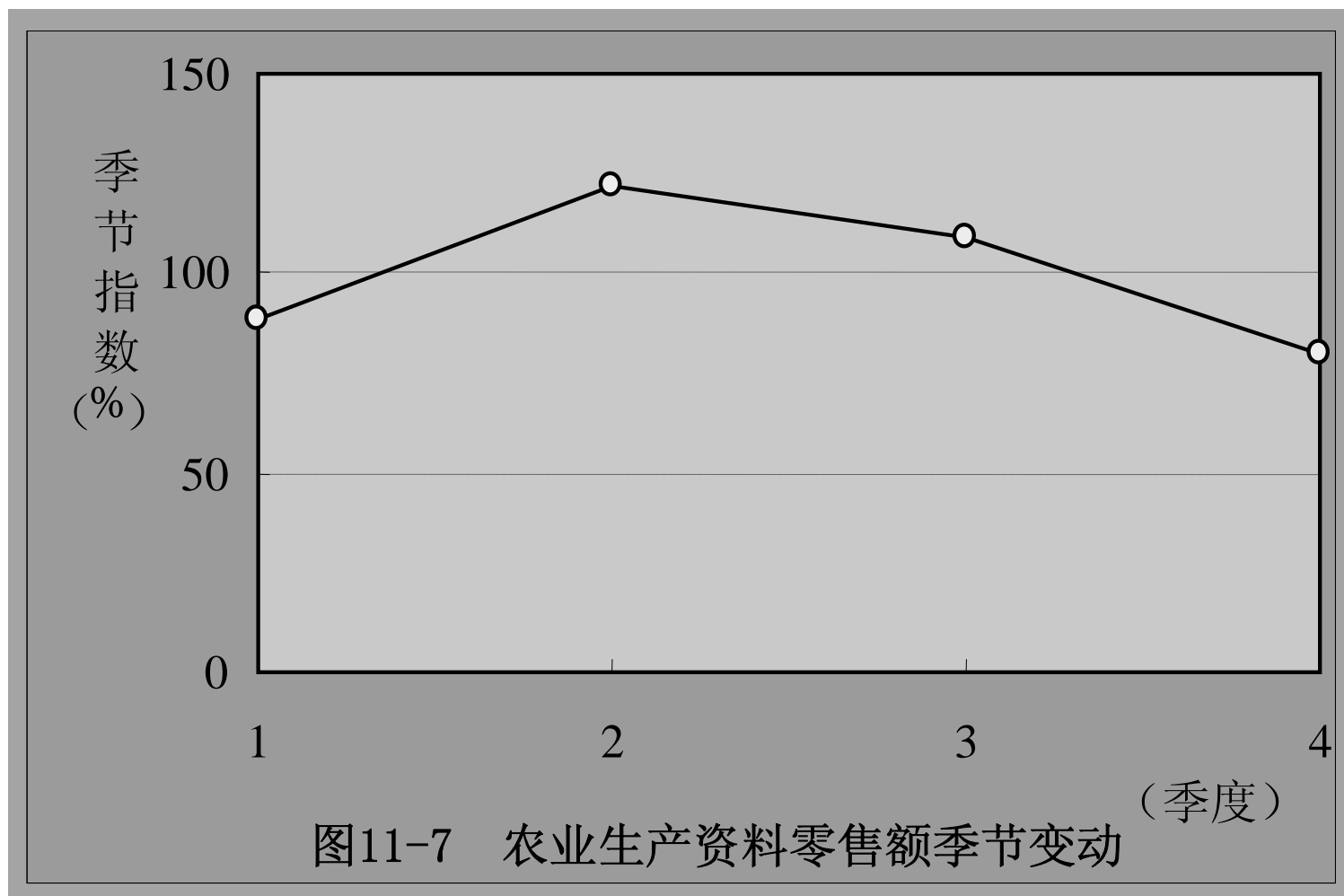
$$\text{季节指数}(S) = \frac{\text{同月(季)平均数}}{\text{总月(季)平均数}} \times 100\%$$

趋势剔除法 (续前例：计算表)

表11- 18 农业生产资料零售额季节指数计算表

年 份	销售额(亿元)				
	一季度	二季度	三季度	四季度	全年合计
1978	—	—	106.12	83.59	
1979	90.91	118.51	108.71	82.57	
1980	87.42	122.85	111.27	78.97	
1981	87.63	122.26	108.70	77.11	
1982	91.07	122.42	110.29	79.08	
1983	84.94	125.65	—	—	
合计	441.98	611.70	545.09	401.33	2000.10
同季平均	88.40	122.34	109.02	80.27	100.005
季节指数(%)	88.39	122.33	109.01	80.26	100.00

季节变动 (趋势图)

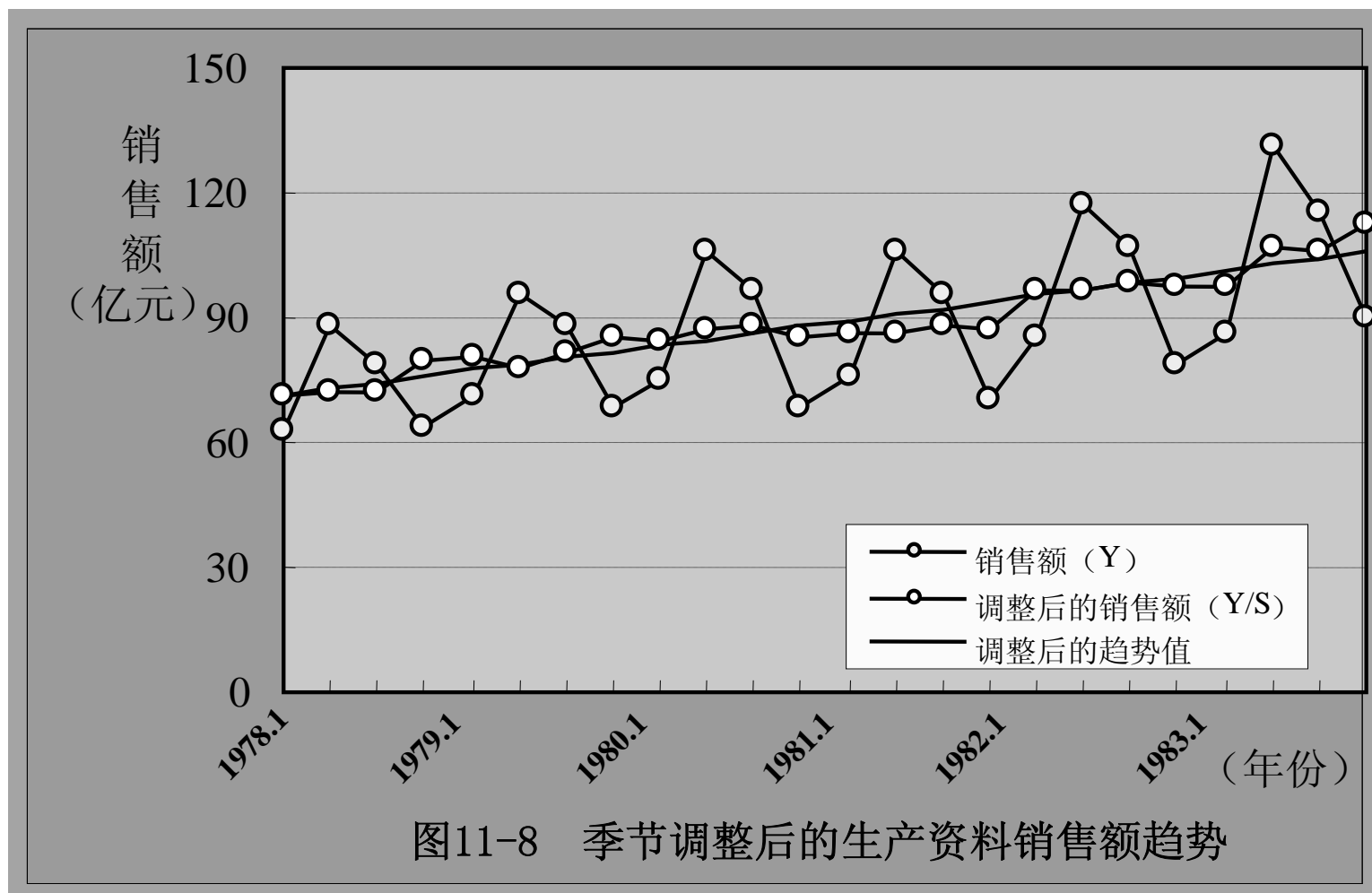


季节变动的调整 (要点和公式)

1. 将季节变动其从时间序列中予以剔除，以便观察和分析时间序列的其他特征
2. 消除季节变动的方法是将原时间序列除以相应的季节指数，计算公式为

$$\frac{Y}{S} = \frac{T \times S \times C \times I}{S} = T \times C \times I$$

季节变动的调整 (趋势图)



第四节 循环波动分析

- 一. 循环波动及其测定目的
- 二. 循环波动的分析方法

循环波动

(概念和测定目的)

1. 近乎规律性的从低至高再从高至低的周而复始的变动
2. 不同于趋势变动，它不是朝着单一方向的持续运动，而是涨落相间的交替波动
3. 不同于季节变动，其变化无固定规律，变动周期多在一年以上，且周期长短不一
4. 目的是探索现象活动的规律性

1. 采用剩余法

2. 具体计算步骤为

- 先消去季节变动，求得无季节性资料
- 再消去趋势值，求得循环及不规则波动相对数
- 将结果进行移动平均（ MA ），以消除不规则波动，即得循环波动值

$$C = MA (C \times I)$$

循环波动

(续前例：循环图)

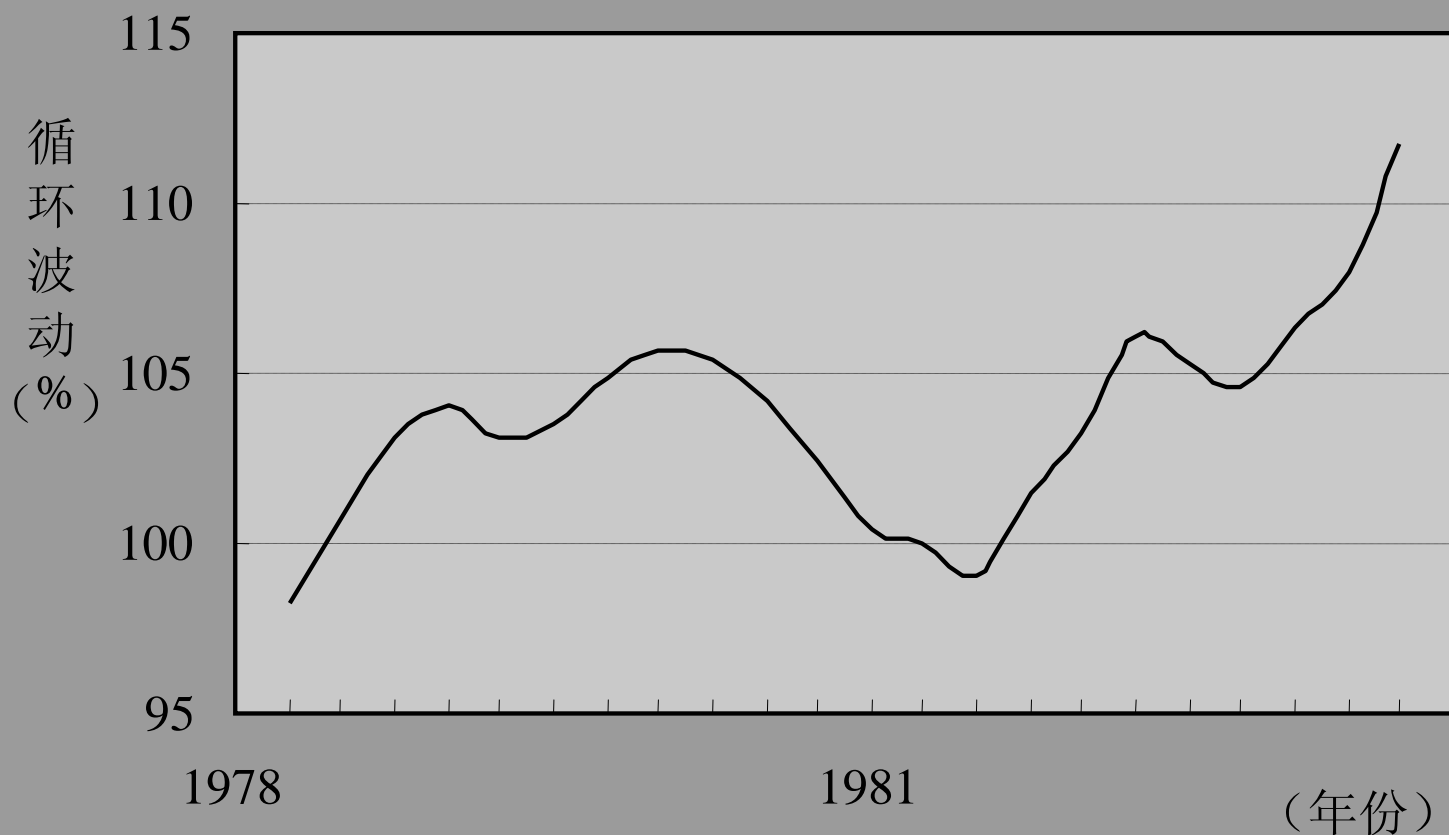


图11-9 生产资料销售额的循环波动

本章小节

1. 时间序列的概念和分类
2. 时间序列的对比分析
3. 趋势变动分析
4. 季节变动分析
5. 循环波动分析
6. 用Excel 进行季节变动分析

结 束



第十二章 指数

PowerPoint



第十二章 指数

第一节 指数编制的基本问题

第二节 加权指数

第三节 指数体系

第四节 几种常用的价格指数

学习目标

1. 掌握加权综合指数的编制方法
2. 掌握加权平均指数的编制方法
3. 利用指数体系对实际问题进行分析
4. 了解实际中常用的几种价格指数

第一节 指数编制的基本问题

- 一. 指数的性质
- 二. 指数的分类
- 三. 指数编制的基本问题

指数的概念和性质

（概念要点）

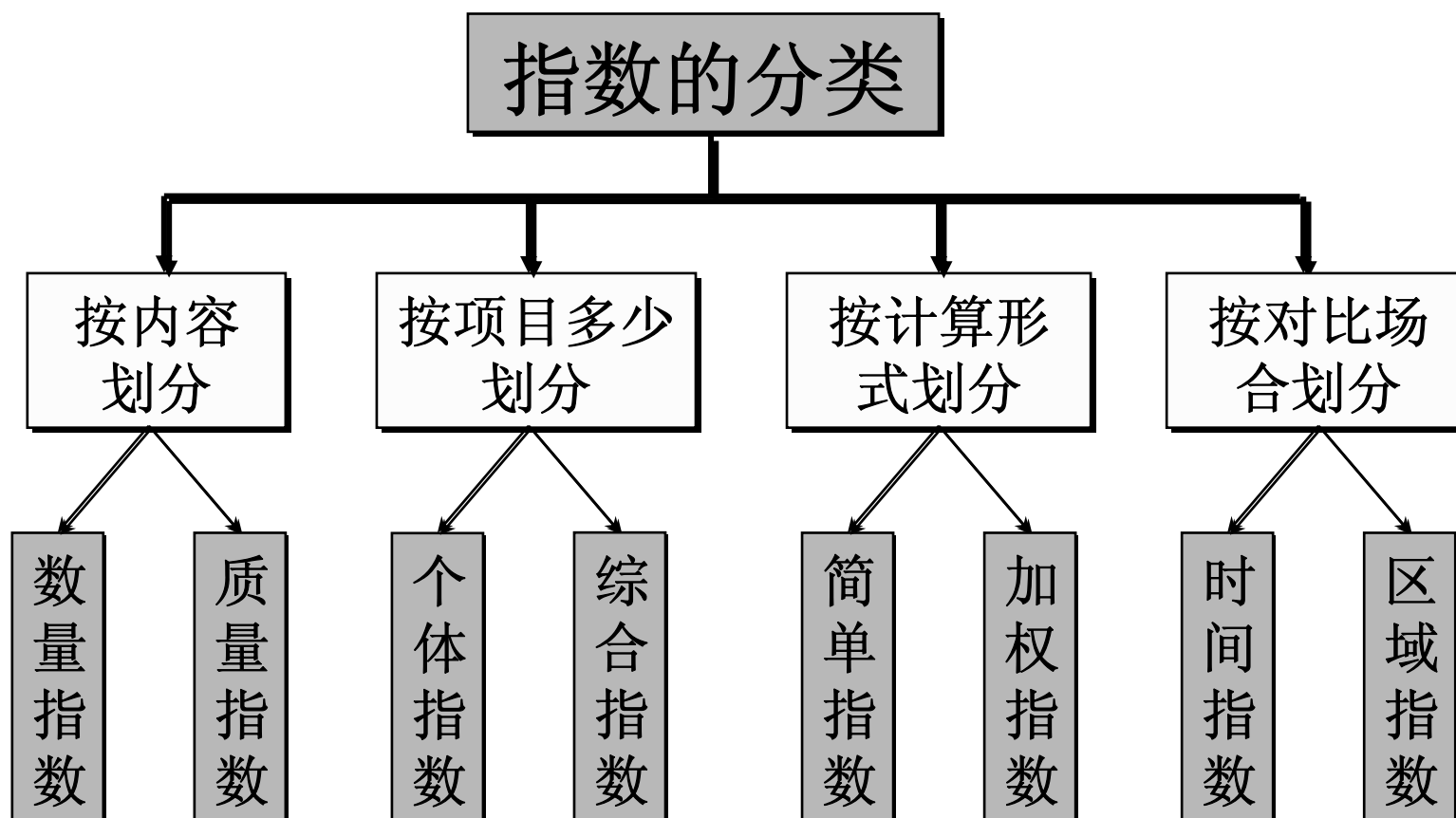
1. 指数的概念

- 广义：任何两个数值对比形成的相对数
- 狭义：用于测定总体各变量在不同场合下综合变动的一种特殊相对数

2. 指数的性质

- 相对性：总体变量在不同场合下对比形成的相对数
 - 不同时间上对比形成的指数称为时间性指数
 - 不同空间上对比形成的指数称为区域性指数
- 综合性：反映一组变量在不同场合下的综合变动
- 平均性：指数是总体水平的一个代表性数值

指数的分类



指数的分类

（数量指数与质量指数）

1. 数量指数

- 反映物量变动水平
- 如产品产量指数、商品销售量指数等

2. 质量指数

- 反映事物内含数量的变动水平
- 如价格指数、产品成本指数等

指数的分类

(个体指数与综合指数)

1. 个体指数

- 反映单一项目的变量变动
- 如一种商品的价格或销售量的变动

2. 综合指数

- 反映多个项目变量的综合变动
- 如多种商品的价格或销售量的综合变动

指数的分类 (其他)

1. 简单指数
 - 计入指数的各个项目的重要性视为相同
2. 加权指数
 - 计入指数的项目依据重要程度赋予不同的权数
3. 时间性指数
 - 总体变量在不同时间上对比形成
 - 有定基指数和环比指数之分
4. 区域性指数
 - 总体变量在不同空间上对比形成

指数编制的基本问题 (要点)

1. 样本项目的选择

- 充分性，样本容量足够大
- 代表性，样本充分反映总体的性质
- 可比性，各样本项目在定义、计算口径、计算方法、计量单位等方面一致

2. 基期的确定

- 选择正常时期或典型时期作为基期
- 报告期距基期的长短应适当

第二节 加权指数

- 一. 权数的确定
- 二. 加权综合指数
- 三. 加权平均指数

权数的确定 (要点)

1. 根据现象之间的联系确定权数
 - 计算数量指数时，应以相应的质量为权数
 - 计算质量指数时，应以相应的物量为权数
2. 确定权数的所属时期
 - 可以都是基期，也可以都是报告期或某一固定时期
 - 使用不同时期的权数，计算结果和意义不同
 - 取决于计算指数的预期目的
3. 确定权数的具体形式
 - 可以是总量形式，也可以采取比重形式
 - 主要取决于所依据的数据形式和计算方法

经济、管理类
基础课程

统计学

加权综合指数

加权综合指数 (概念要点)

1. 通过加权来测定一组项目的综合变动
2. 有加权数量指数和加权质量指数
 - 数量指数
 - 测定一组项目的数量变动
 - 如产品产量指数，商品销售量指数等
 - 质量指数
 - 测定一组项目的质量变动
 - 如价格指数、产品成本指数等
3. 因权数不同，有不同的计算公式

基期变量值加权的综合指数 (要点和计算公式)

1. 将作为权数的各变量值固定在基期
2. 也被称为拉氏指数或 L 式指数
3. 计算公式为

- 质量指数:
$$p_{1/0} = \frac{\sum p_1 q_0}{\sum p_0 q_0}$$

- 数量指数:
$$q_{1/0} = \frac{\sum p_0 q_1}{\sum p_0 q_0}$$

4. 可以消除权数变动对指数的影响

基期变量值加权的综合指数 (实例)

【例12.1】 设某粮油商店1999年和1998年三种商品的零售价格和销售量资料如表12-1。试分别以基期销售量和零售价格为权数，计算三种商品的价格综合指数和销售量综合指数。

表12-1 某粮油商店三种商品的价格和销售量

商品名称	计量单位	销售量		单价(元)	
		1998	1999	1998	1999
粳 米	公斤	1200	1500	3.6	4.0
标准粉	公斤	1500	2000	2.3	2.4
花生油	公斤	500	600	9.8	10.6

基期变量值加权的综合指数 (计算过程)

表12-2 加权综合指数计算表

商品名称	计量单位	销售量		单价(元)		销售额(元)			
		1998 q_0	1999 q_1	1998 p_0	1999 p_1	1998 p_0q_0	1999 p_1q_1	p_0q_1	p_1q_0
粳 米	kg	1200	1500	3.6	4.0	4320	6000	5400	4800
标准粉	kg	1500	2000	2.3	2.4	3450	4800	4600	3600
花生油	kg	500	600	9.8	10.6	4900	6360	5880	5300
合计	—	—	—	—	—	12670	17160	15880	13700

基期变量值加权的综合指数 (计算结果)

价格综合指数为

$$p_{1/0} = \frac{\sum_1 p_1 q_0}{\sum_1 p_0 q_0} = \frac{13700}{12670} = 108.73\%$$

销售量综合指数为

$$q_{1/0} = \frac{\sum_1 p_0 q_1}{\sum_1 p_0 q_0} = \frac{15880}{12670} = 125.34\%$$

结论：与1998年相比，三种商品的零售价格平均上涨了8.73%，销售量平均上涨了25.34%

报告期变量值加权的综合指数 (要点和计算公式)

1. 将作为权数的各变量值固定在报告期
2. 也被称为帕氏指数，或简称为 P 式指数
3. 计算公式为
 - 质量指数：
$$p_{1/0} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$
 - 数量指数：
$$q_{1/0} = \frac{\sum p_1 q_1}{\sum p_1 q_0}$$
4. 不能消除权数变动对指数的影响

报告期变量值加权的综合指数 (实例)

【例12.2】 根据表12-1中的数据资料，分别以报告期销售量和零售价格为权数计算三种商品的价格综合指数和销售量综合指数。

表12-1 某粮油商店三种商品的价格和销售量

商品名称	计量单位	销售量		单价(元)	
		1998	1999	1998	1999
粳 米	公斤	1200	1500	3.6	4.0
标准粉	公斤	1500	2000	2.3	2.4
花生油	公斤	500	600	9.8	10.6

报告期变量值加权的综合指数 (计算过程)

表12-2 加权综合指数计算表

商品名称	计量单位	销售量		单价(元)		销售额(元)			
		1998 q_0	1999 q_1	1998 p_0	1999 p_1	1998 p_0q_0	1999 p_1q_1	p_0q_1	p_1q_0
粳 米	kg	1200	1500	3.6	4.0	4320	6000	5400	4800
标准粉	kg	1500	2000	2.3	2.4	3450	4800	4600	3600
花生油	kg	500	600	9.8	10.6	4900	6360	5880	5300
合计	—	—	—	—	—	12670	17160	15880	13700

报告期变量值加权的综合指数 (计算结果)

价格综合指数为

$$p_{1/0} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{17160}{15880} = 108.06\%$$

销售量综合指数为

$$q_{1/0} = \frac{\sum p_1 q_1}{\sum p_1 q_0} = \frac{17160}{13700} = 125.26\%$$

结论：与1998年相比，三种商品的零售价格平均上涨了8.06%，销售量平均上涨了25.26

固定时期变量值加权的综合指数 (要点和计算公式)

1. 将作为权数的各变量值固定在某个具有代表性的特定时期
2. 权数不受基期和报告期的限制，使指数的编制具有较大的灵活性
3. 编制若干个时期的多个指数时，可以消除因权数不同对指数的影响
4. 生产价格指数通常采用该方法编制

固定时期变量值加权的综合指数 (实例)

【例12.3】 设某企业生产三种产品的有关资料如表12-3。
试以1990年不变价格为权数，计算各年的产品产量指数

某企业生产三种产品的有关资料					
商品名称	计量单位	销售量			1990年 不变价格 (元)
		1994	1995	1996	
甲	件	1000	960	1100	50
乙	台	120	120	125	3500
丙	箱	200	215	240	300

固定时期变量值加权的综合指数 (计算结果)

解：设1990年不变价格为 p_{90} ，各年产量分别为 q_{94} 、 q_{95} 、 q_{96} ，则各年产量指数为

$$q_{95/94} = \frac{\sum p_{90} q_{95}}{\sum p_{90} q_{94}} = \frac{532500}{530000} = 100.47\%$$

$$q_{96/95} = \frac{\sum p_{90} q_{96}}{\sum p_{90} q_{95}} = \frac{564500}{532500} = 106.01\%$$

$$q_{96/94} = \frac{\sum p_{90} q_{96}}{\sum p_{90} q_{94}} = \frac{564500}{530000} = 106.51\%$$

经济、管理类
基础课程

统计学

加权平均指数

加权平均指数 (概念要点)

1. 以某一时期的总量为权数对个体指数加权平均
2. 权数通常是两个变量的乘积
 - 可以是价值总量，如商品销售额(销售价格与销售量的乘积)、工业总产值(出厂价格与生产量的乘积)
 - 可以是其他总量，如农产品总产量(单位面积产量与收获面积的乘积)
3. 因权数所属时期的不同，有不同的计算形式

基期总量加权的平均指数 (要点和计算公式)

1. 以基期总量为权数对个体指数加权平均
2. 计算形式上采用算术平均形式

3. 计算公式为

■ 质量指数:
$$p_{1/0} = \frac{\sum \frac{p_1}{p_0} p_0 q_0}{\sum p_0 q_0}$$

■ 数量指数:
$$q_{1/0} = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0}$$

基期总量加权的平均指数 (实例)

【例12.4】 设某企业生产三种产品的有关资料如表12-4。试计算三种产品的单位成本总指数和产量总指数。

表12-4 某企业生产三种产品的有关数据

商品名称	计量单位	总成本(万元)		个体成本指数 (p_1/p_0)	个体产量指数 (q_1/q_0)
		基期 (p_0q_0)	报告期 (p_1q_1)		
甲	件	200	220	1.14	1.03
乙	台	50	50	1.05	0.98
丙	箱	120	150	1.20	1.10

基期总量加权的平均指数 (计算结果)

单位成本指数为

$$p_{1/0} = \frac{\sum \frac{p_1}{p_0} p_0 q_0}{\sum p_0 q_0} = \frac{1.14 \times 200 + 1.05 \times 50 + 1.20 \times 120}{200 + 50 + 120} = \frac{425.5}{370} = 114.73\%$$

产量总指数为

$$q_{1/0} = \frac{\sum \frac{q_1}{q_0} p_0 q_0}{\sum p_0 q_0} = \frac{1.03 \times 200 + 0.98 \times 50 + 1.10 \times 120}{200 + 50 + 120} = \frac{387}{370} = 104.59\%$$

结论：报告期与基期相比，三种产品的单位成本平均提高了14.73%，产量平均提高了4.59%

报告期总量加权的平均指数 (要点和计算公式)

1. 以报告期总量为权数对个体指数加权平均
2. 计算形式上采用调和平均形式
3. 计算公式为

- 质量指数：
$$p_{1/0} = \frac{\sum p_1 q_1}{\sum \frac{1}{p_1/p_0} p_1 q_1}$$

- 数量指数：
$$q_{1/0} = \frac{\sum p_1 q_1}{\sum \frac{1}{q_1/q_0} p_1 q_1}$$

报告期总量加权的平均指数 (实例)

【例12.5】 根据表12-4中的有关数据，用报告期总成本为权数计算三种产品的单位成本总指数和产量总指数。

表12-4 某企业生产三种产品的有关数据

商品名称	计量单位	总成本(万元)		个体成本指数 (p_1/p_0)	个体产量指数 (q_1/q_0)
		基期 (p_0q_0)	报告期 (p_1q_1)		
甲	件	200	220	1.14	1.03
乙	台	50	50	1.05	0.98
丙	箱	120	150	1.20	1.10

报告期总量加权的平均指数 (计算结果)

单位成本指数为

$$p_{1/0} = \frac{\sum p_1 q_1}{\sum \frac{1}{p_1/p_0} p_1 q_1} = \frac{220+50+150}{\frac{220}{1.14} + \frac{50}{1.05} + \frac{150}{1.20}} = \frac{420}{365.60} = 114.88\%$$

产量总指数为

$$q_{1/0} = \frac{\sum p_1 q_1}{\sum \frac{1}{q_1/q_0} p_1 q_1} = \frac{220+50+150}{\frac{220}{1.03} + \frac{50}{0.98} + \frac{150}{1.10}} = \frac{420}{400.98} = 104.74\%$$

结论：报告期与基期相比，三种产品的单位成本平均提高了14.88%，产量平均提高了4.74%

第三节 指数体系

- 一. 总量指数与指数体系
- 二. 指数体系的分析与应用

1. 由两个不同时期的总量对比

- 可以是实物总量对比，如粮食总产量指数
- 可以是价值总量对比，称为价值指数，如工业总产值、产品总成本、商品销售额指数

2. 一般形式

- 个体总量指数：
$$v_{1/0} = \frac{p_1 q_1}{p_0 q_0}$$
- 综合总量指数：
$$v_{1/0} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

指数体系 (概念要点)

1. 由总量指数及其若干个因素指数构成的数量关系式
2. 总量指数等于各因素指数的乘积
3. 总量的变动差额等于各因素指数变动差额之和
4. 两个因素指数中通常一个为数量指数，另一个为质量指数
5. 各因素指数的权数必须是不同时期的

加权综合指数体系 (要点及公式)

1. 因所用权数时期不同，有不同的指数体系
2. 比较常用的是基期权数加权的数量指数和报告期权数加权的质量指数形成的指数体系
3. 指数体系可表示为

- 相对数关系

$$\frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times \frac{\sum p_0 q_1}{\sum p_0 q_0}$$

- 绝对数关系

$$\sum p_1 q_1 - \sum p_0 q_0 = \left(\sum p_1 q_1 - \sum p_0 q_1 \right) + \left(\sum p_0 q_1 - \sum p_0 q_0 \right)$$

加权综合指数体系 (实例及应用)

【例12.6】 根据例12.1的有关数据，利用指数体系分析价格和销售量变动对销售额的影响

$$\text{销售额指数} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{17160}{12670} = 135.44\%$$

$$\text{价格指数} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{17160}{15880} = 108.06\%$$

$$\text{销售量指数} = \frac{\sum p_0 q_1}{\sum p_0 q_0} = \frac{15880}{12670} = 125.34\%$$

加权综合指数体系 (实例及应用)

销售额变动

$$= \sum p_1 q_1 - \sum p_0 q_0 = 17160 - 12670 = 4490(\text{元})$$

价格变动的影响额

$$= \left(\sum p_1 q_1 - \sum p_0 q_1 \right) = 17160 - 15880 = 1280(\text{元})$$

销售量变动的影响额

$$= \sum p_0 q_1 - \sum p_0 q_0 = 15880 - 12670 = 3210(\text{元})$$

加权综合指数体系 (实例及应用)

三者之间的相对数量关系

$$135.44\% = 108.06\% \times 125.34\%$$

三者之间的绝对数量关系

$$4490(\text{元}) = 1280(\text{元}) + 3210(\text{元})$$

结论：1999年与1998年相比，三种商品的销售额增长35.44%，增加销售额4490元。其中由于零售价格变动使销售额增长8.06%，增加销售额1280元；由于销售量变动使销售额增长25.34%，增加销售额3210元

加权平均指数体系 (要点及公式)

1. 因所用总量权数所属时期不同，有不同的指数体系
2. 常用的是基期总量加权算术平均数量指数和报告期总量加权调和平均质量指数形成的指数体系

- 相对数关系

$$\frac{\sum_1 p_1 q_1}{\sum_1 p_0 q_0} = \frac{\sum_1 \frac{q_1}{q_0} p_0 q_0}{\sum_1 p_0 q_0} \times \frac{\sum_1 p_1 q_1}{\sum_1 \frac{1}{p_1/p_0} p_1 q_1}$$

- 绝对数关系

$$\sum_1 p_1 q_1 - \sum_1 p_0 q_0 = \left(\sum_1 \frac{q_1}{q_0} p_0 q_0 - \sum_1 p_0 q_0 \right) + \left(\sum_1 p_1 q_1 - \sum_1 \frac{1}{p_1/p_0} p_1 q_1 \right)$$

利用指数体系分析平均数变动 (要点及公式)

1. 通过两个不同时期的加权算术平均数之比反映现象平均水平的变动
2. 通过对加权算术平均数的分解，分析影响平均数变动的各因素

$$\text{平均数变动指数} = \frac{\bar{x}_1}{\bar{x}_0} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0}$$

$$\text{变量影响指数} = \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1}$$

$$\text{结构影响指数} = \frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0}$$

利用指数体系分析平均数变动 (要点及公式)

1. 三个指数之间的相对数量关系

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} = \left\{ \frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} \right\} \times \left\{ \frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} \right\}$$

2. 三个指数之间的绝对数量关系

$$\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} = \left\{ \frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_1}{\sum f_1} \right\} + \left\{ \frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} \right\}$$

利用指数体系分析平均数变动 (实例)

【例12.7】某企业有三个生产车间，1998年和1999年各车间的工人数和劳动生产率资料如表12-5。试分析该企业劳动生产率的变动及其原因。

表12-5 某企业职工人数和劳动生产率资料

车间	职工人数(人)		劳动生产率(万元/人)	
	1998	1999	1998	1999
一车间	200	240	4.4	4.5
二车间	160	180	6.2	6.4
三车间	150	120	9.0	9.2

利用指数体系分析平均数变动 (计算过程)

表12-6 某企业职工人数和劳动生产率资料

车间	职工人数(人)		劳动生产率(万元/人)		总产值(万元)		
	1998 f_0	1999 f_1	1998 x_0	1999 x_1	1998 x_0f_0	1999 x_1f_1	x_1f_1
一车间	200	240	4.4	4.5	880	1080	1056
二车间	160	180	6.2	6.4	992	1152	1116
三车间	150	120	9.0	9.2	1350	1104	1080
合计	510	540	6.32	6.18	3222	3336	3252

利用指数体系分析平均数变动 (计算结果及分析)

1998年人均
劳动生产率 $\frac{\sum x_0 f_0}{\sum f_0} = \frac{3222}{510} = 6.32(\text{万元/人})$

1999年人均
劳动生产率 $\frac{\sum x_1 f_1}{\sum f_1} = \frac{3336}{540} = 6.18(\text{万元/人})$

人均劳动生
产率指数为 $\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} = \frac{6.18}{6.32} = 97.78\%$

利用指数体系分析平均数变动 (计算结果及分析)

各车间劳动生产率变动影响指数

$$\frac{\sum x_1 f_1}{\sum f_1} \div \frac{\sum x_0 f_1}{\sum f_1} = \frac{6.18}{6.02} = 102.66\%$$

各车间职工人数变动影响指数

$$\frac{\sum x_0 f_1}{\sum f_1} \div \frac{\sum x_0 f_0}{\sum f_0} = \frac{6.02}{6.32} = 95.25\%$$

三者之间的相对数量关系为

$$\mathbf{97.78\% = 102.66\% \times 95.25\%}$$

利用指数体系分析平均数变动 (计算结果及分析)

该企业人均劳动
生产率变动额

$$\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} = 6.18 - 6.32 = -0.14 (\text{万元/人})$$

各车间劳动生产
率变动影响额

$$\frac{\sum x_1 f_1}{\sum f_1} - \frac{\sum x_0 f_1}{\sum f_1} = 6.18 - 6.02 = 0.16 (\text{万元/人})$$

各车间职工人
数变动影响额

$$\frac{\sum x_0 f_1}{\sum f_1} - \frac{\sum x_0 f_0}{\sum f_0} = 6.02 - 6.32 = -0.3 (\text{万元/人})$$

三者之间的
关系为

$$-0.14 = 0.16 - 0.3$$

利用指数体系分析平均数变动 (计算结果及分析)

结论

1. 1999年同1998年相比，该企业三个车间的劳动生产率均有所提高，但企业总的劳动生产率却下降了2.22%，人均下降0.14万元
2. 各车间劳动生产率的提高使企业总的生产率提高了2.66%，人均提高0.16万元
3. 各车间职工人数结构的变化，使企业总的劳动生产率下降了4.75%，人均下降0.3万元

第四节 几种常用的价格指数

- 一. 零售价格指数
- 二. 消费价格指数
- 三. 股票价格指数

零售价格指数 (Retail Price Index)

1. 反映城乡商品零售价格变动趋势的一种经济指数
2. 编制中的主要问题
 - 代表规格品的选择
 - 典型地区的选择
 - 商品价格的确定
 - 权数的确定
 - 计算公式
$$P_{1/0} = \frac{\sum kw}{\sum w}$$

消费价格指数 (Consumer Price Index)

1. 世界各国普遍编制的一种指数
 - 不同国家对这一指数赋予的名称不一致
 - 我国称之为居民消费价格指数
2. 反映一定时期内城乡居民所购买的生活消费品价格和服务项目价格的变动趋势和程度
3. 可就城乡分别编制
4. 计算公式
$$p_{1/0} = \frac{\sum kw}{\sum w}$$

消费价格指数 (作用)

1. 反映生活消费品价格和服务价格的变动趋势和程度
2. 反映通货膨胀状况

$$\text{通货膨胀率} = \frac{\text{报告期消费价格指数} - \text{基期消费价格指数}}{\text{基期消费价格指数}} \times 100\%$$

3. 反映货币购买力变动

$$\text{货币购买力指数} = \frac{1}{\text{居民消费价格指数}} \times 100\%$$

4. 反映对职工实际工资的影响

$$\text{实际工资} = \frac{\text{名义工资}}{\text{消费价格指数}}$$

股票价格指数 (Stock Price Index)

1. 反映股票市场上多种股票价格变动趋势
2. 用“点”(point)表示
3. 计算公式为

$$P_{1/0} = \frac{\sum p_{1i} q_i}{\sum p_{0i} q_i}$$

股票价格指数 (Stock Price Index)

- ☞ 世界主要证券交易所的股票价格指数
 - 道·琼斯股票价格指数和标准普尔股票价格指数；伦敦金融时报FTSE指数；法兰克福DAX指数；巴黎CAC指数；瑞士的苏黎士SMI指数；日本的日京指数；香港的恒生指数
 - 我国上海和深圳两个证券交易所
 - 上交所的综合指数和30指数
 - 深交所的成分股指数和综合指数

本章小节

1. 指数的概念与分类
2. 加权综合指数的计算
3. 加权平均指数的计算
4. 利用指数体系进行分析
5. 几种常用的价格指数

结 束

