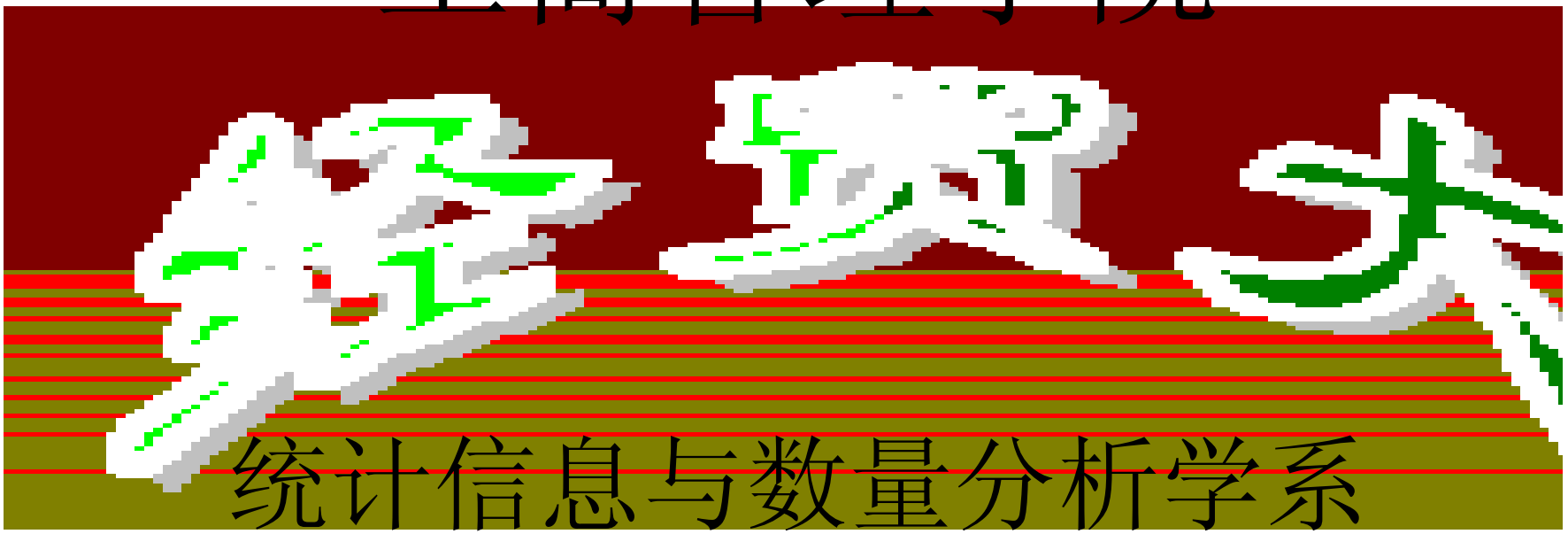
应用统计

对外经济贸易大学

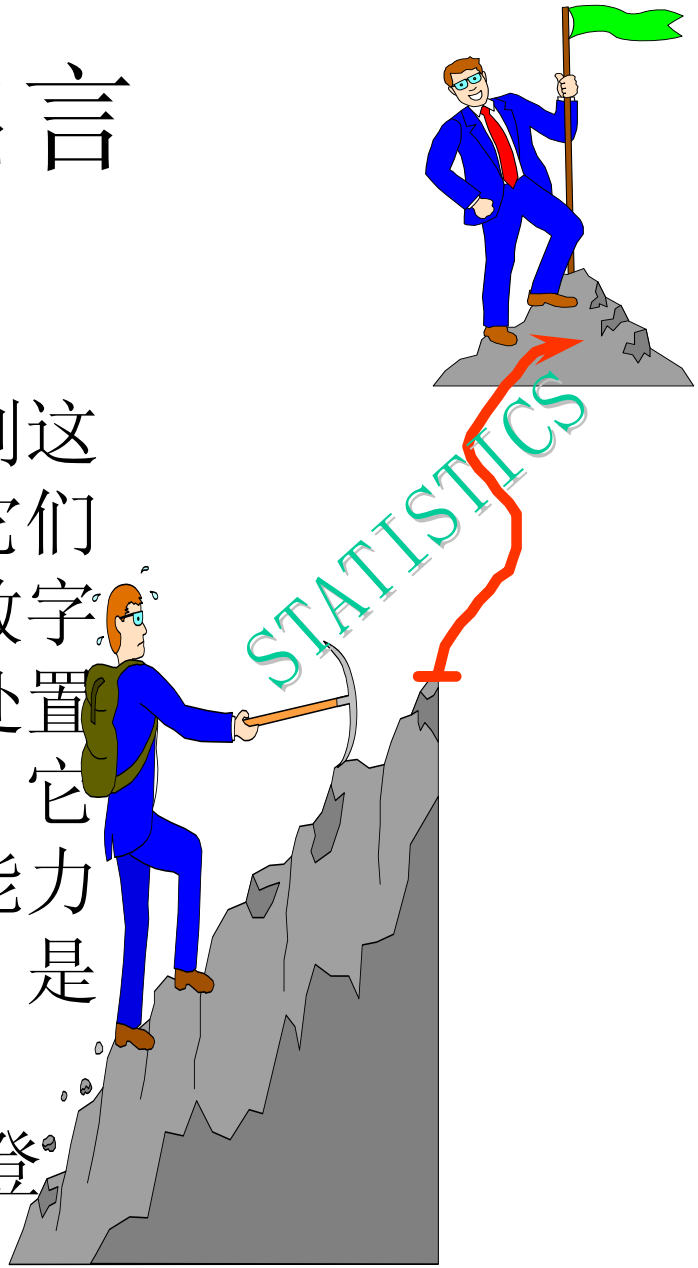
工商管理学院



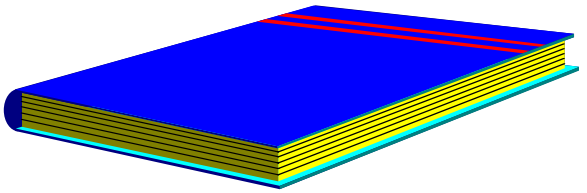
统计信息与数量分析学系

统计学家名言

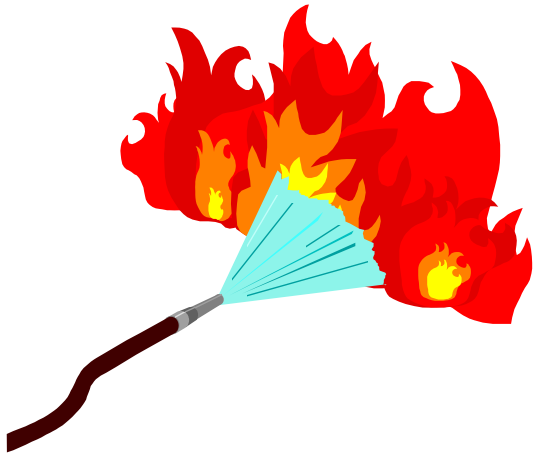
- 一些人厌恶统计数字，甚至听到这个字眼就皱眉头，而我却发现它们妙趣横生。当人们不是将这些数字胡堆乱放，而是用精明手段去处置它们，小心翼翼地作出解释时，它们就显出应付复杂现象的非凡能力。统计对于追求科学的人来说，是披荆斩棘开拓路径的利器。
- ——弗朗西斯.高尔登



正读反思



- 统计学家通常只留恋于平均数的研究，而不去注意更复杂的问题。他们对于千差万别的大千世界的观念是如此平庸，就如同我们英国平原各郡的乡民认为可以把瑞士的山岭填入湖泊一样，除两嫌于一举。
- —— Francis Galton



均值方差均方互辅

万千事物皆有律

教师学子教学相长

浩瀚理海总无涯

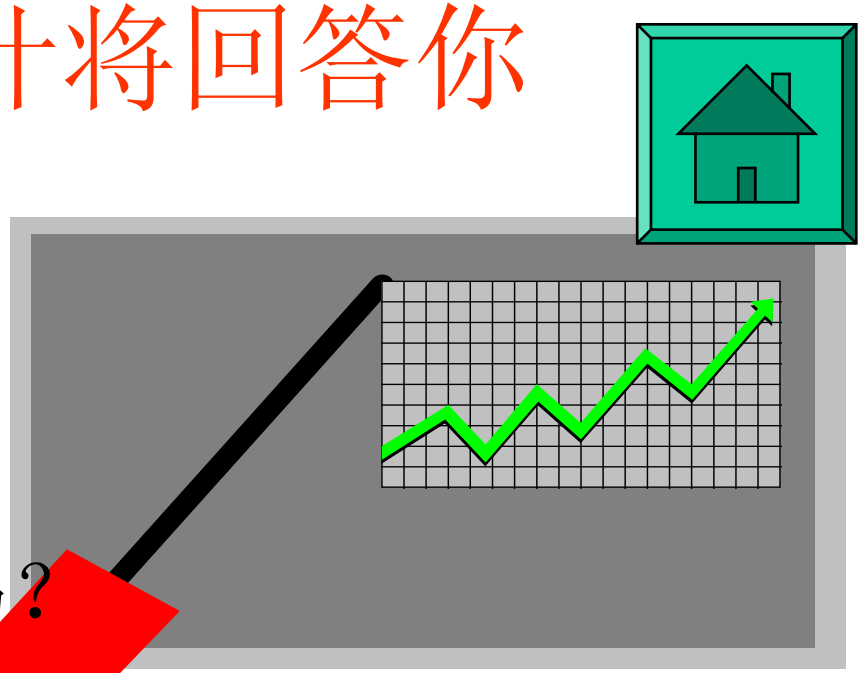
贾怀勤

书赠修统计学的学生

并自勉

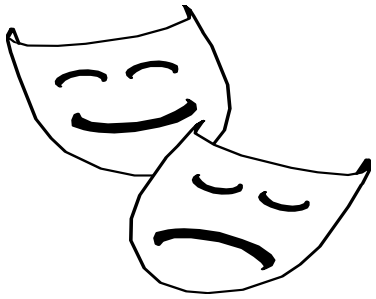
一九八五年元旦

应用统计将回答你



- 统计能做什么？
- 统计是“瞎估计”吗？
- 统计资料哪里来？
- 怎样从数据中找规律？
- 怎样从样本推断总体？
- 统计怎样反映事物间的联系？

企业管理中什麽地方需要统计



- 市场调研的数据怎么来，来了怎么用？
- 生产资源如何优化配置，生产过程和产品质量如何控制？
- 会计数据如何分析？
- 投资效益和风险如何测定？项目怎样优选？
- 绩效怎样与报酬挂钩？职工情态怎样计量？

对外经贸大学统计学系：

应用统计课程的创设者（1982）

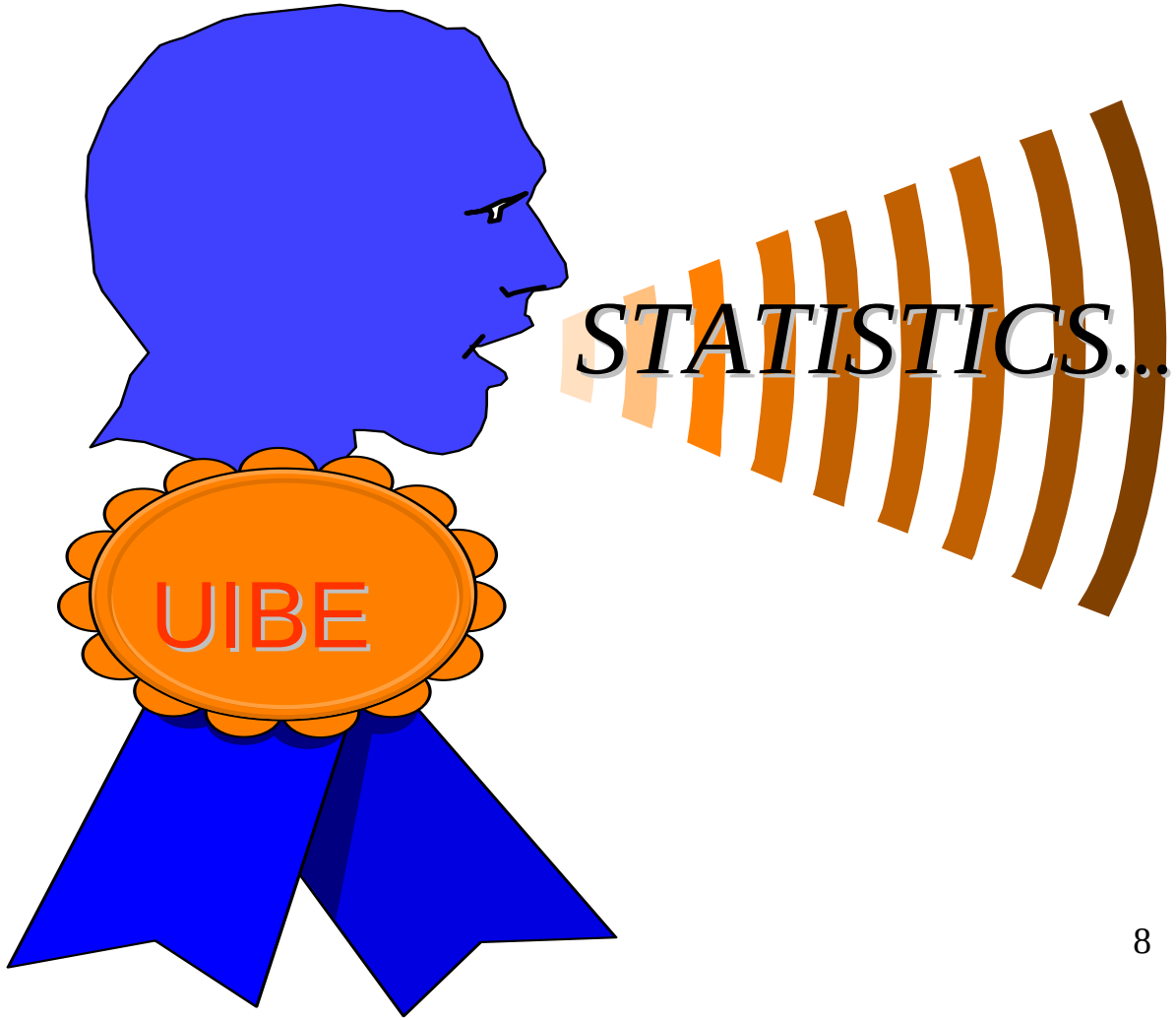
1992年以“创设应用统计新型课程”获北京市高校优秀教学成果一等奖

本校编著的统计学教材

统计分析概论	李志伟	1984
统计分析概论（修订版）	李志伟	1989
应用统计	贾怀勤	1994
应用统计（第二版）	贾怀勤	1998
应用统计（第三版）	贾怀勤	2002
应用统计（第四版）	贾怀勤	2006
管理统计学	贾怀勤	2001
数据模型决策	贾怀勤	2004

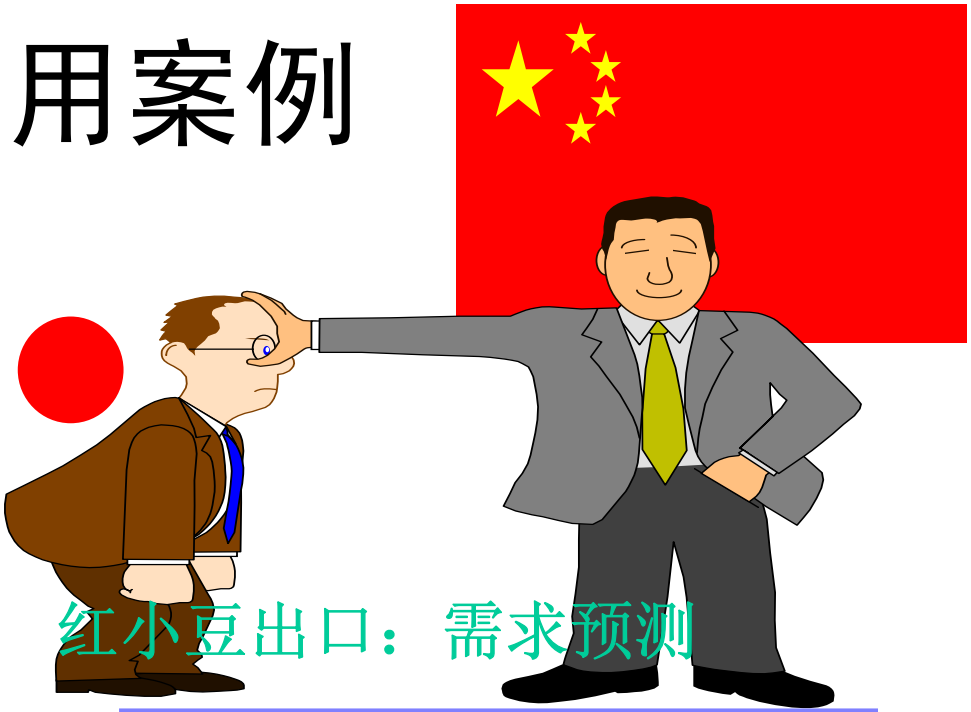
本系统计学教师名录

- 贾怀勤
- 张　　杰
- 杜学孔
- 王玉蓉
- 朱雅华
- 杨宝峰



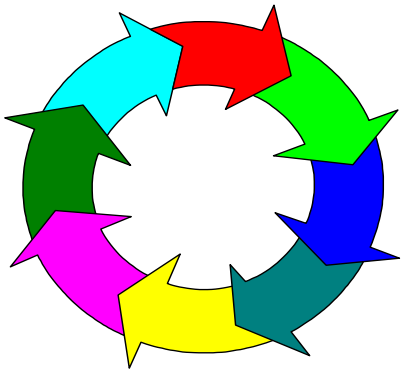
统计应用案例

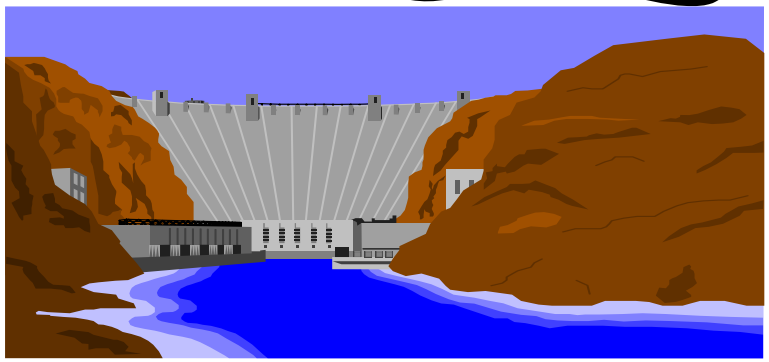
芬兰统计失实
经济陷入混乱



红小豆出口：需求预测

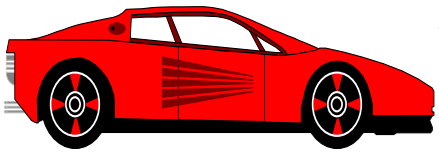
泛世通
防雪轮胎
营销新概念：



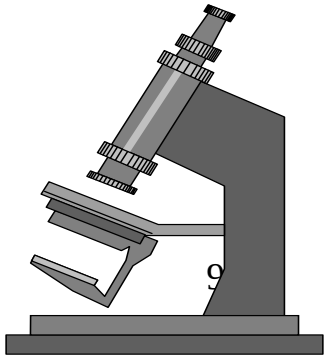


今冬无雪降，

退钱没商量！



天广线
进口电缆
检验和索赔



资料整理和数列类别

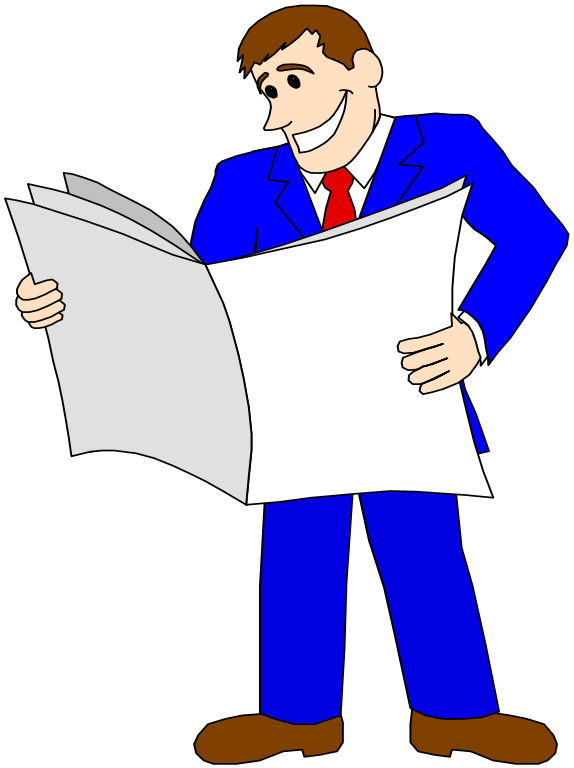
- 总体与单位
- 标志
  - 品质标志
  - 数量标志
- 指标
- 汇总与分组
  - 质别分组
  - 量别分组
    - 等距分组

- 截面数列
- 质别分组数列
- 量别分指数列
- 时间数列
- 绝对数时间数列
- 时期数列
- 时点数列

统计资料的搜集（一）

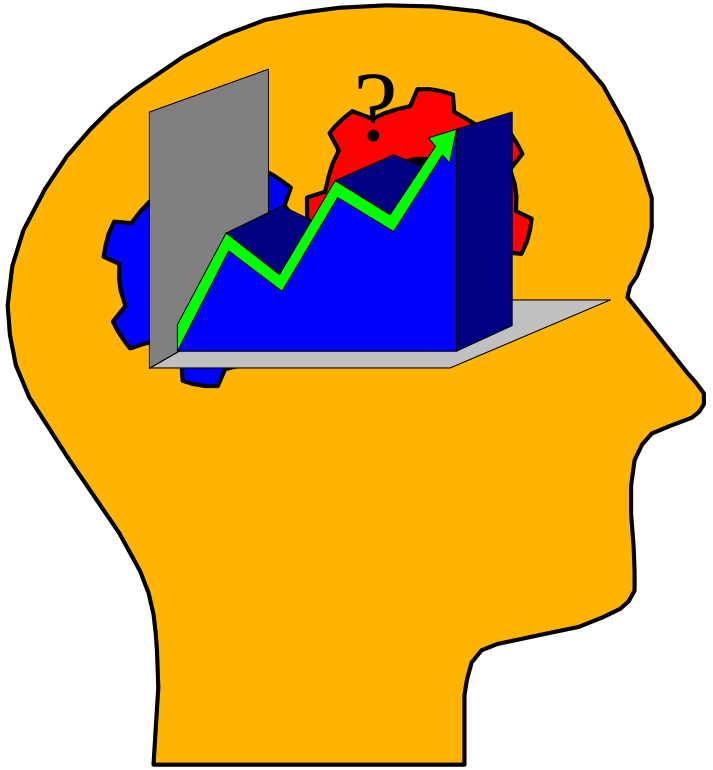
- 初级资料
  - 行政记录
  - 直接调查
    - 询问
    - 观察

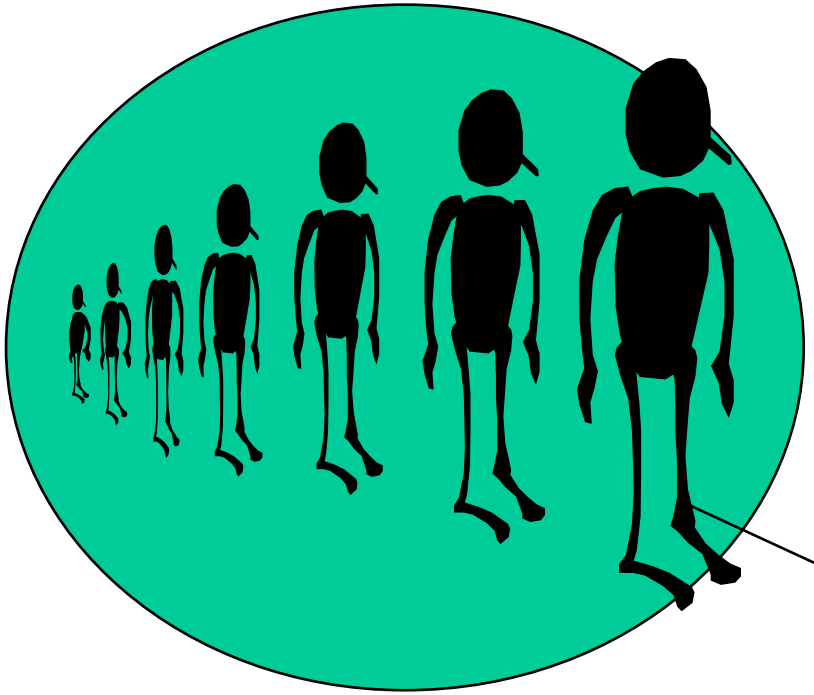
统计资料的搜集（二）



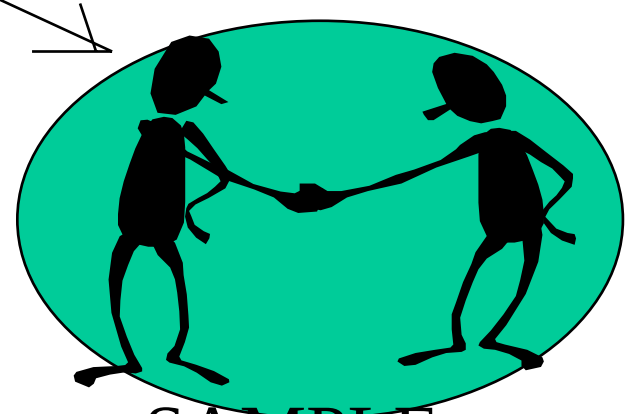
- 次级资料
  - 内部资料
  - 外部资料
    - 政府机构
    - 其他公共机构
    - 新闻媒体
    - 行业
    - 数据供应商
    - 国外资料

抽样调查?



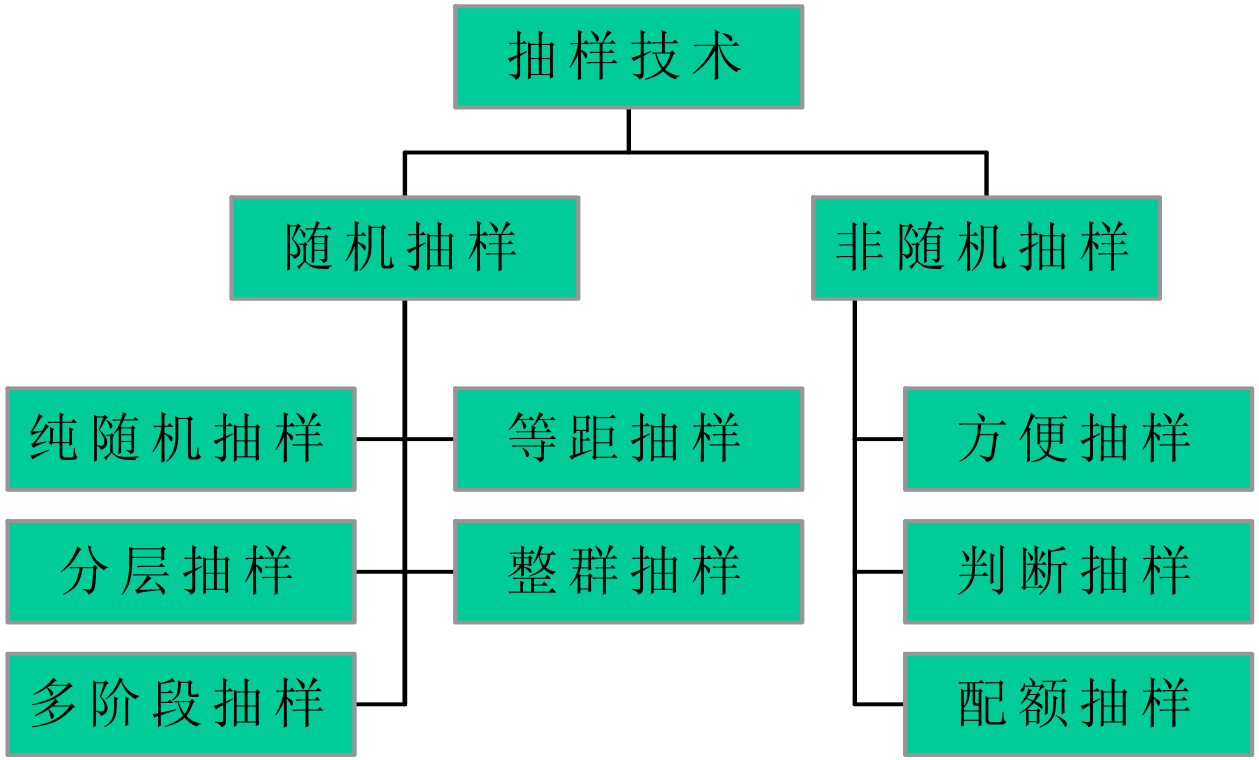


POPULATION



SAMPLE

抽样技术



计算机数据处理与统计分析

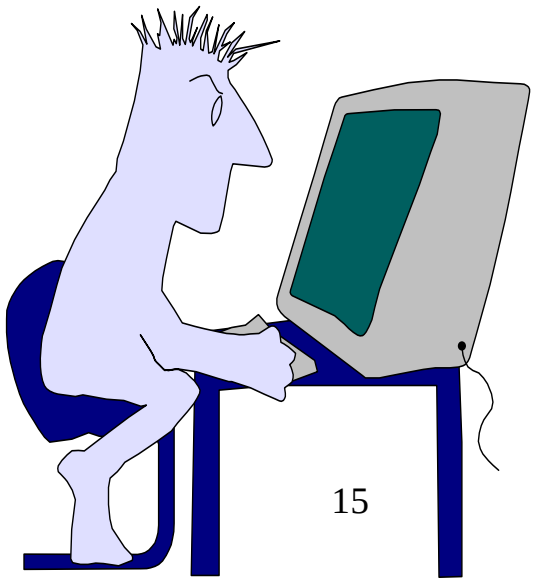
集中趋势量数和离散趋势量数：
算术平均数；方差和标准差；
五大位势量数。

常用概率分布函数与反函数：
二项分布；正态分布；
T分布；F分布；卡方分布。

参数估计
与
假设检验

回归分析与建模

时间数列解析
和
预测



计算机数据处理与统计分析

软件：

统计软件——SPSS，SAS，Minitab

办公自动化软件——Excel

计算机软件

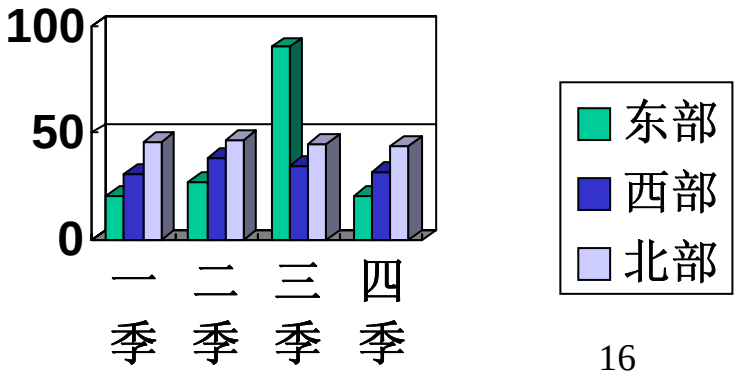
集数据汇总—分组与数据分析于一身，

灵活实现探索性分析，描述性分析和推断性分析

编制和修改各种统计表

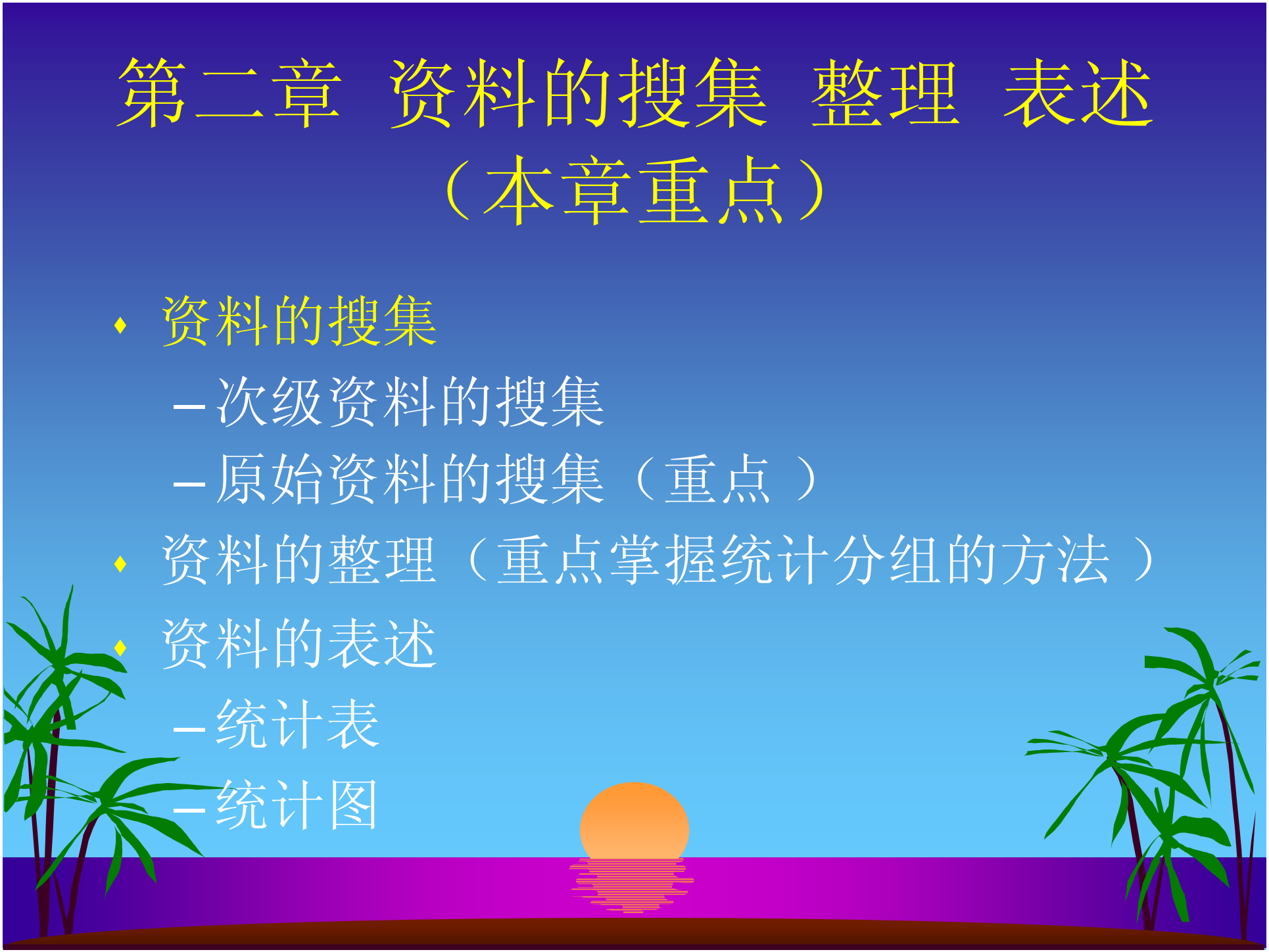
生成多种统计图。

	一季	二季	三季	四季
东部	20.4	27.4	90	20.4
西部	30.6	38.6	34.6	31.6
北部	45.9	46.9	45	3.9



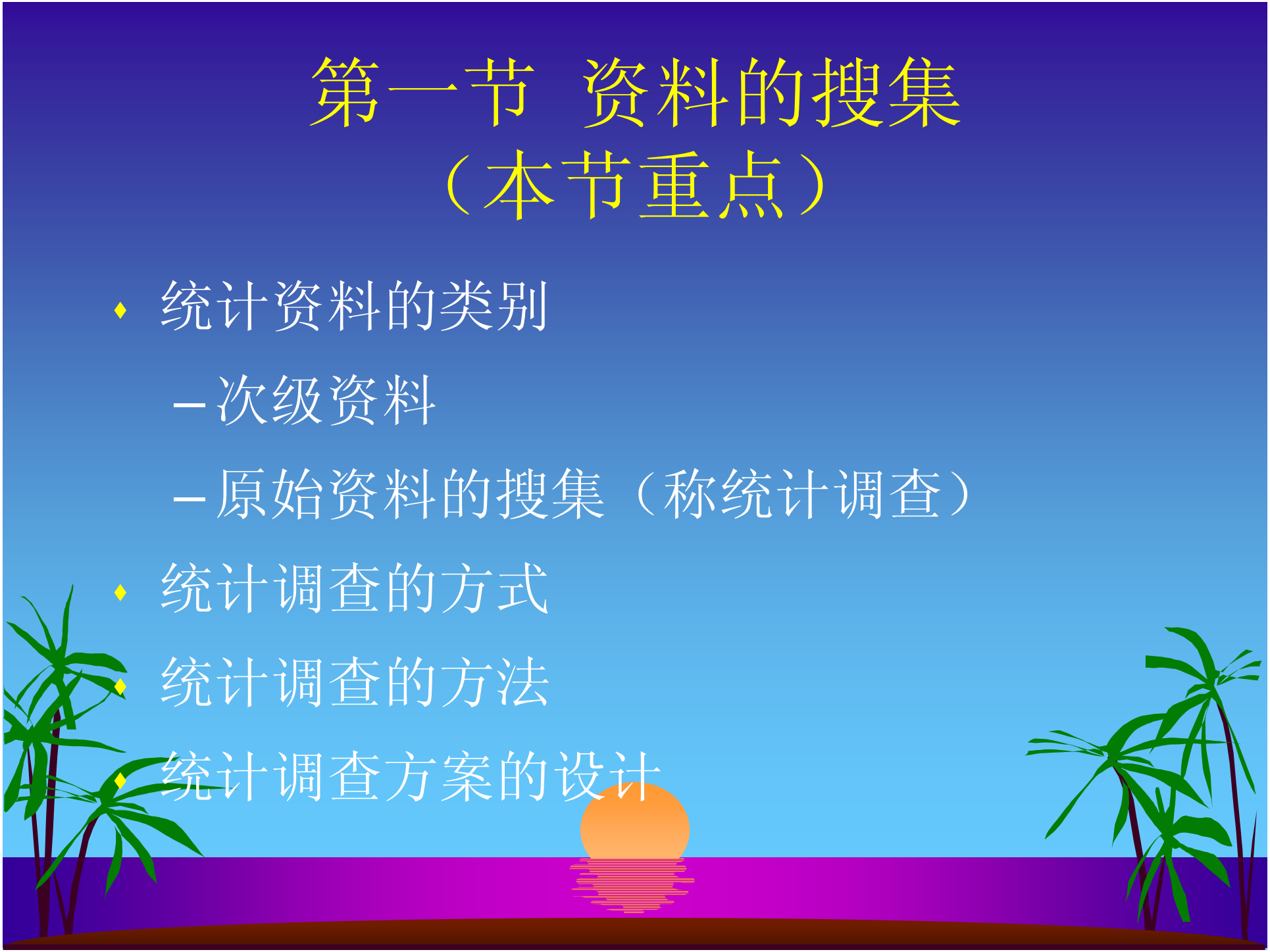
第二章 资料的搜集 整理 表述（本章重点）

- 资料的搜集
  - 次级资料的搜集
  - 原始资料的搜集（重点）
- 资料的整理（重点掌握统计分组的方法）
- 资料的表述
  - 统计表
  - 统计图



第一节 资料的搜集（本节重点）

- 统计资料的类别
  - 次级资料
  - 原始资料的搜集（称统计调查）
- 统计调查的方式
- 统计调查的方法
- 统计调查方案的设计



统计资料的来源

资料来源

原始资料

次级资料

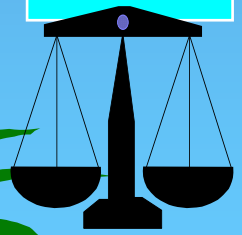
实验

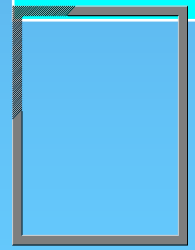
调查

观察

1.出版物

2.网上

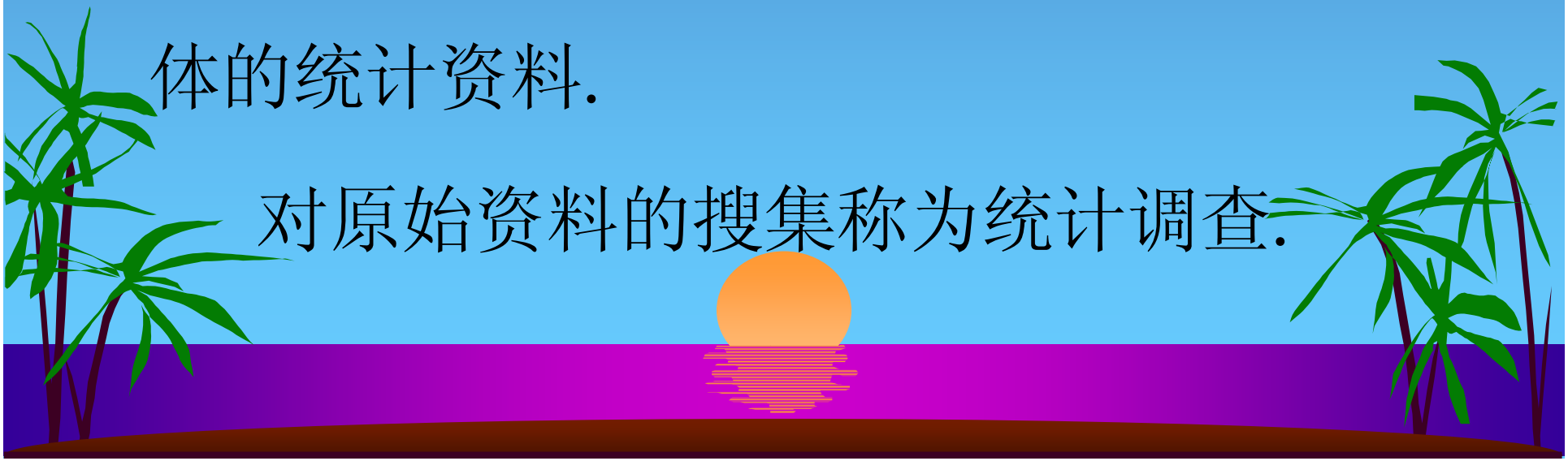




原始资料

原始资料（ primary data ）又称初级资料、第一手资料，是指直接向调查单位搜集的、尚待汇总整理、需要由个体过渡到总体的统计资料.

对原始资料的搜集称为统计调查.



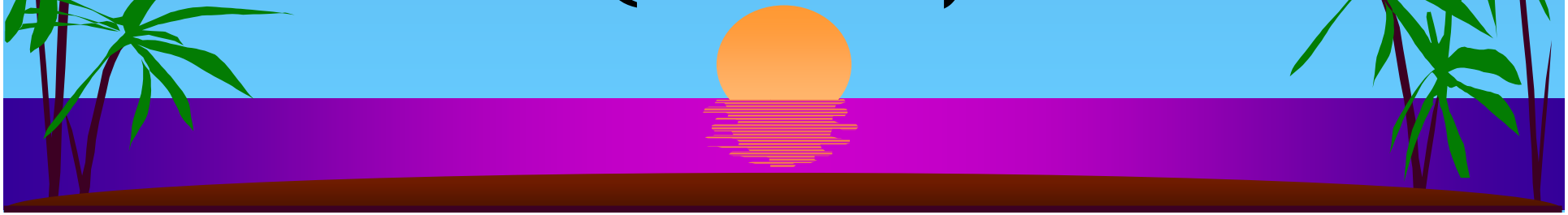
次级资料

次级资料（secondary data）又称间接资料、第二手资料，是指已经经过加工整理、由个体过度到总体,能够在一定程度上说明总体现象的统计资料.

所有的次级资料，都来源于初级资料.初级资料较为形象、生动和可靠，而次级资料的可靠性就差一些.

统计调查方式的种类

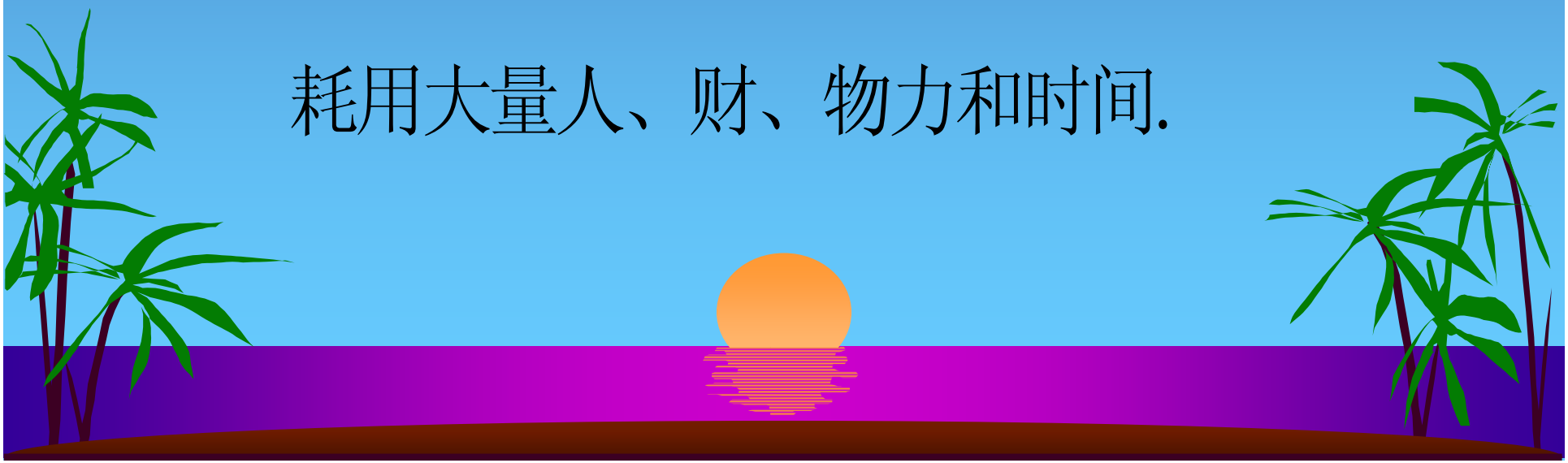
- 全面调查
  - 全面报表
  - 普查
- 非全面调查
  - 专门调查
    - 抽样调查
    - 重点调查
    - 典型调查



普查（Census）

普查是专门组织的、一次性的全面调查.

普查所得资料较为全面和细致，但需耗用大量人、财、物力和时间.



抽样调查

抽样调查的类型

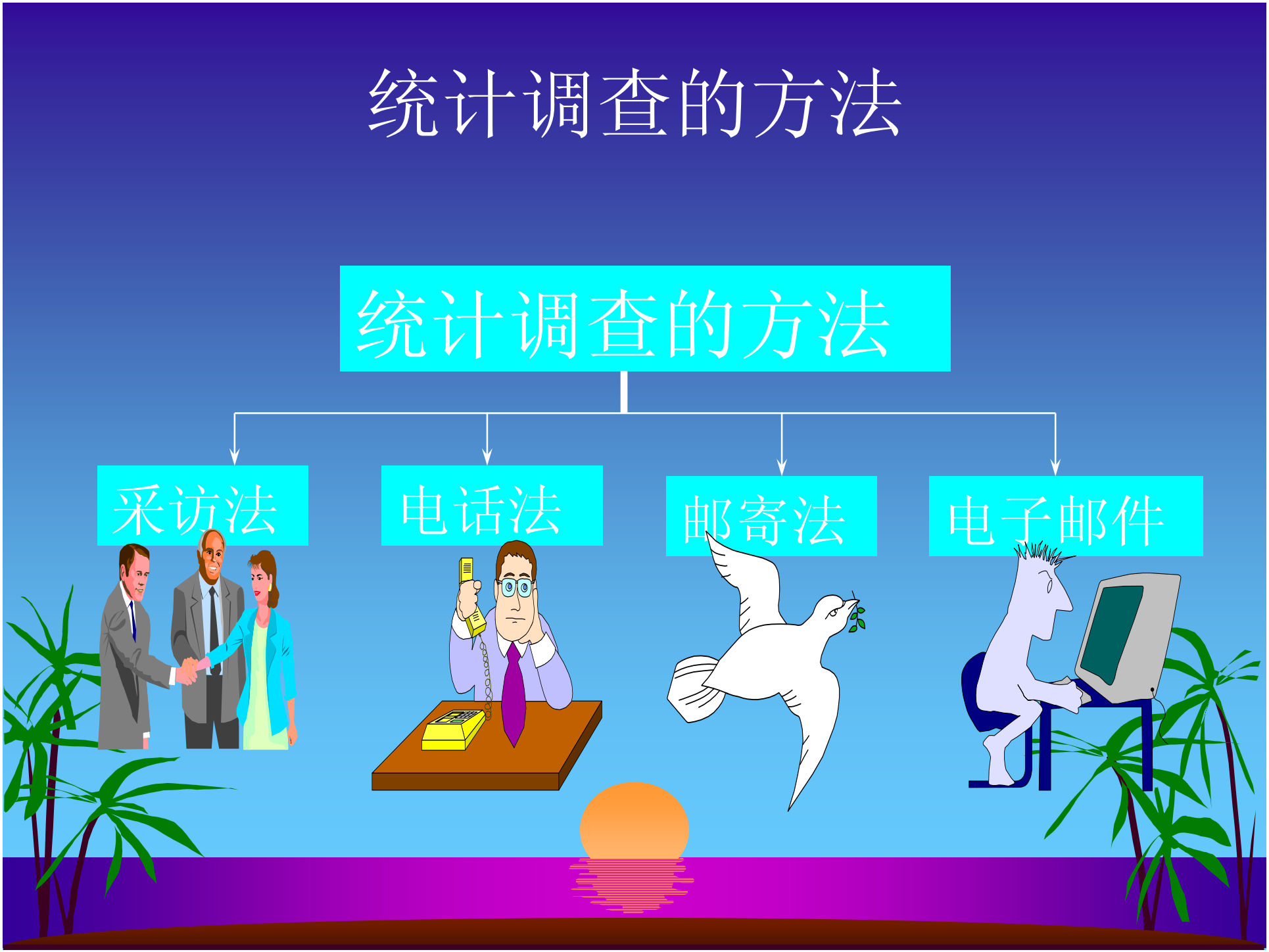
- 非随机抽样
  - 判断抽样
    - 按照一定的标准有意识地在总体中抽取若干合乎标准的样本单位进行调查.
  - 立意抽样
    - 也称方便抽样,抽取样本的标准主要是方便.这样抽出的样本代表性不高.
- 随机抽样 (重点掌握)

随机抽样调查

随机抽样调查----是按照随机的原则，从总体中抽取一部分样本单位进行观察,从而用样本指标估计总体指标的一种调查方法.

随机性----即等可能性，指 1.每个总体单位有同等的被抽取的可能.。 2..每个样本被抽中的可能性相等。

统计调查的方法



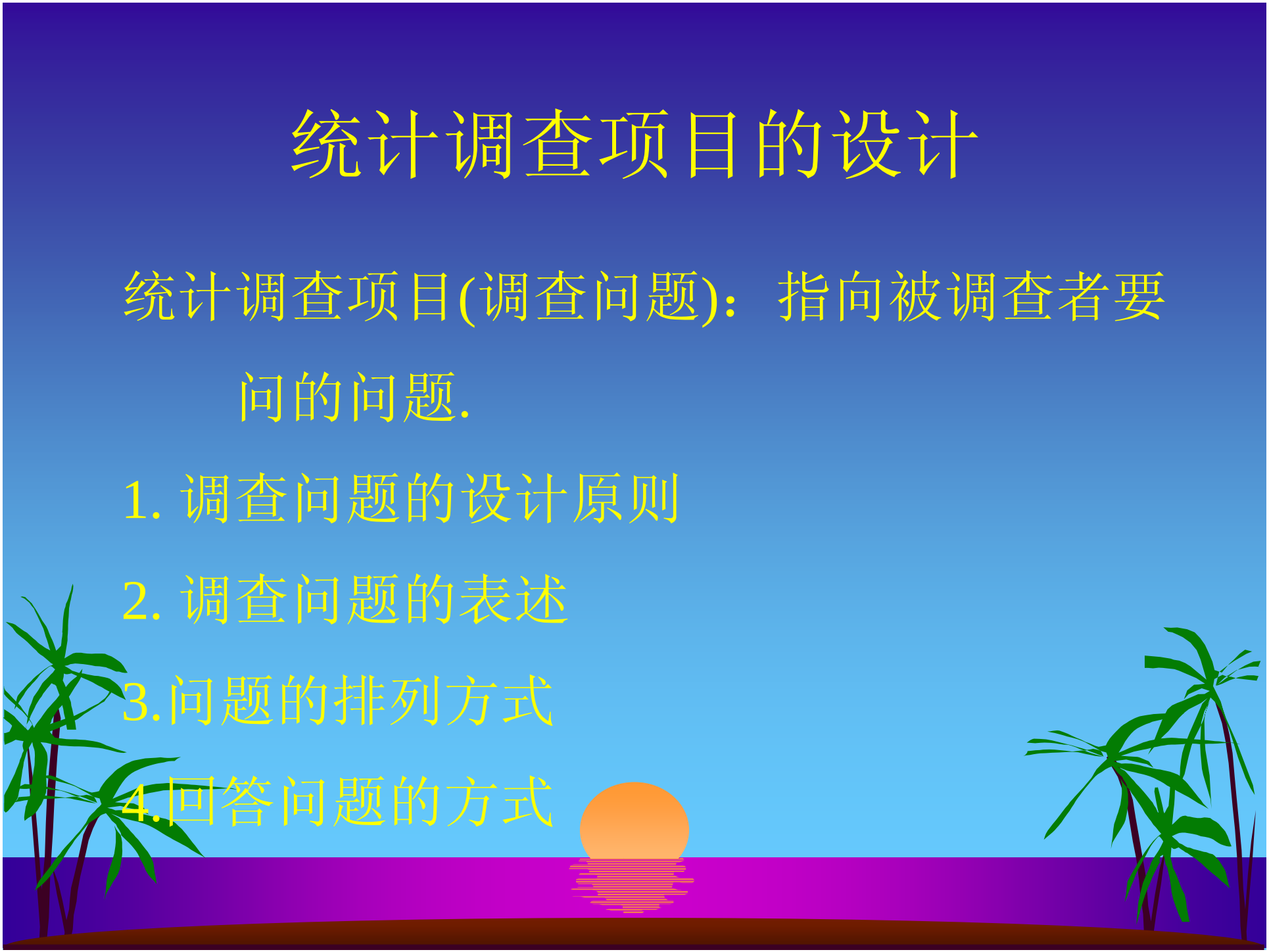
统计调查方案的设计

1. 确定调查目的
2. 确定调查对象和调查单位，报告单位
3. 确定调查项目（标志）
4. 确定调查时间和地点
5. 确定调查方式和方法
6. 确定调查的组织计划

统计调查项目的设计

统计调查项目(调查问题)：指向被调查者要问的问题.

1. 调查问题的设计原则
2. 调查问题的表述
3. 问题的排列方式
4. 回答问题的方式



调查问题的设计的原则

- 紧扣调查目的.
- 必须符合客观实际情况.
- 符合被调查者回答问题的能力.

  如:你的价值观是什么?这类问题很难说清楚.
- 必须考虑真实回答问题的可能性,凡被调查者不可能真实回答的问题,都不应正面提出.

调查问题的表述

- 问题的内容要具体,不要提抽象的,笼统的问题.
  如:你认为青年人应建立什么样的人生观?
  问题太抽象,笼统.
- 问题要单一.
- 问题的用词要通俗,不要使用被调查者陌生的,过分专业化的语言.
- 用词要准确,不要使用模棱两可,含糊不清或容易产生分歧的词和概念.如:也许,好象等词最好不用.

问题的排列方式

- 把同类性质的问题安排在一起
- 先易后难,由浅入深;先事实,行为方面的问题,后观念，感情,态度方面的问题.
- 按时间的顺序排列.过去,现在,将来.

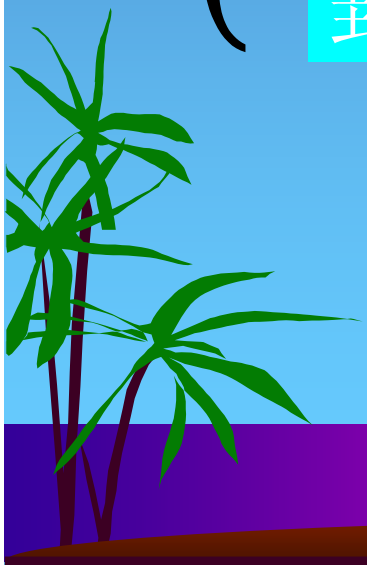
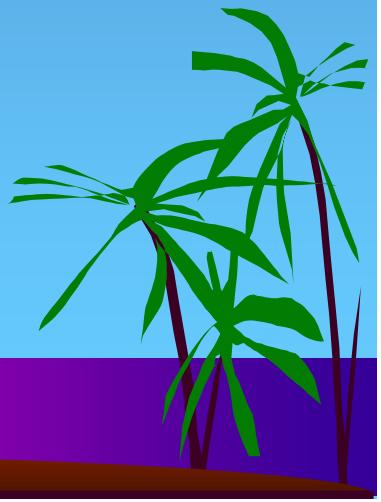
总之问题的排列应便于

1.被调查者顺利回答问题

2.调查后的资料整理和分析.

回答问题的方式

- 开放型回答
- 封闭型回答
  - 填空式
  - 两项式
  - 多项选择
  - 顺序填答式
  - 等级填答式
  - 矩阵式
  - 表格式

顺序填答式

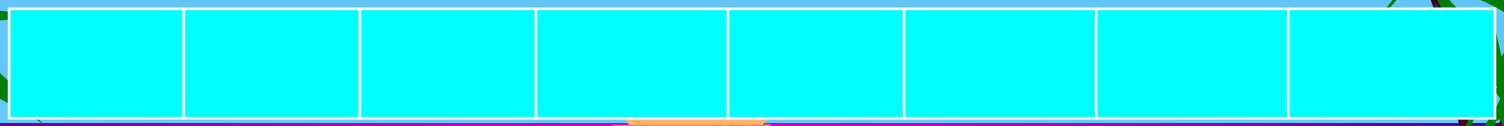
列出若干答案,由被调查者填写各种答案先后顺序的回答方式

如:你认为当前不正之风的突出表现是什么?(请按严重程度把下列问题的编号填写在后面的空格里)

1.大吃大喝　2.行贿受贿　3.乱买小汽车

4.乱盖私房　5.乱发文凭　6.乱发奖金,实物

7.用公款旅游　8.提干走后门



等级填答式

列出不同等级的答案,由被调查者根据自己的意见选择填答的方式.

如: 你是否喜欢吃方便面

1._____ 很喜欢

2._____ 喜欢

3._____ 一般

4._____ 不喜欢

5._____ 很不喜欢

矩阵式

- 将同类的几个问题和答案排列成一个矩阵,由被调查者对比着进行回答的方式.

如:你希望自己的生活在哪些方面得到改善?

	非常迫切	比较迫切	一般	不需改善	无所谓
1.吃的方面					
2.穿的方面					
3.用的方面					
4.住的方面					
5.行的方面					
6.娱乐方面					

第二节 原始资料整理

1. 设计统计整理方案

2. 对原始资料进行审核

？ 逻辑审核和计算审核

3. 对审核后的原始资料进行科学的分组

4. 对各组的资料进行汇总和加工

5. 将汇总整理的结果编制成统计表和统计图

统计分组

概念

统计分组就是根据统计总体的本质特征,按照一定的标志把统计总体区分为若干个组成部分.

对总体单位而言——合

对总体而言——分

统计分组的作用

1. 划分客观现象的类型

2. 反映客观现象的内部结构

3. 分析客观现象间的依存关系

分组标志的选择及统计分组原则

1.分组标志是统计分组的依据,要根据研究的目的、结合具体情况,选择具有本质性的标志或主要标志作为统计分组的依据.

2. 原则

穷尽性

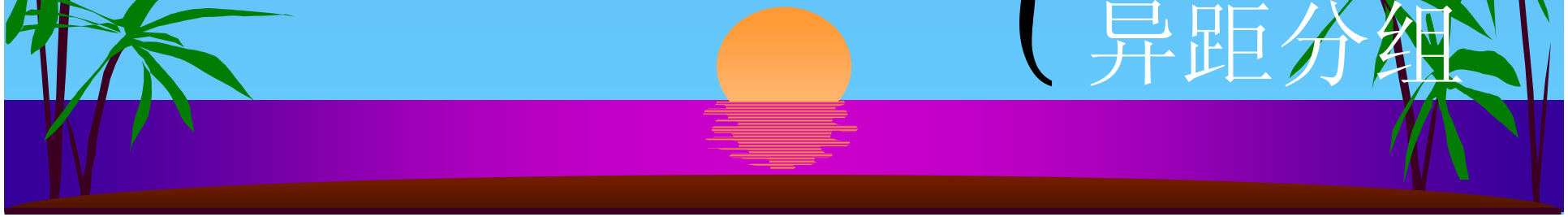
互斥性

一个总体单位能且只能归属到一个组中.

统计分组的种类1

依据标志的性质不同

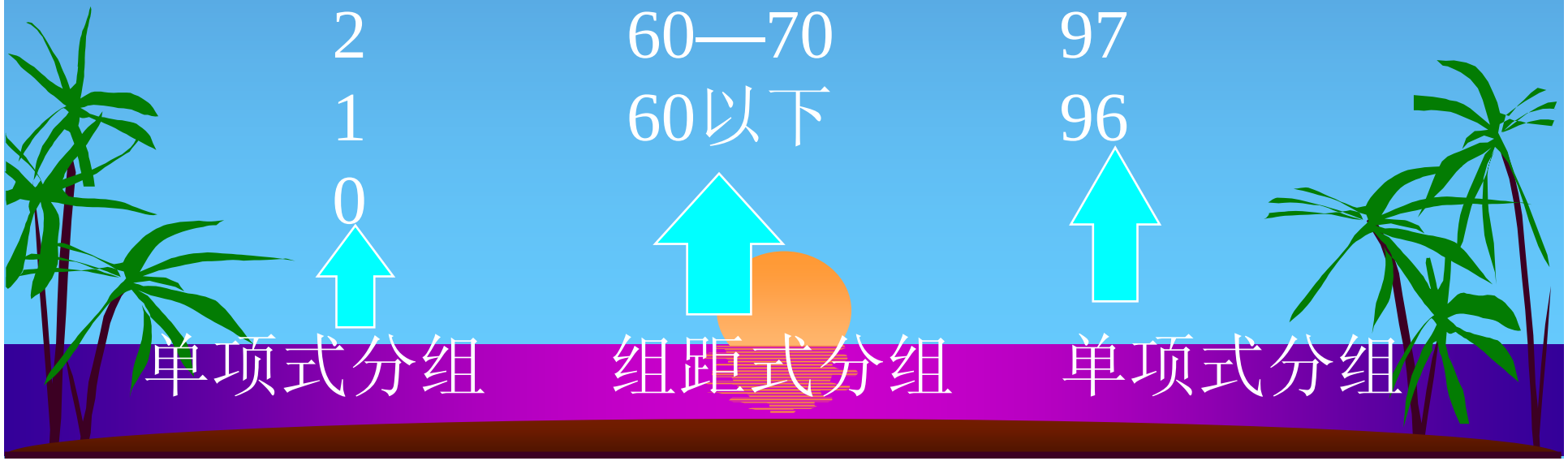
- 质别分组
- 量别分组
  - 单项式分组
  - 组距分组
    - 等距分组
    - 异距分组



分组类型的举例

把学生的成绩分别按五分制和百分制分组

成绩	成绩	成绩
5	90—100	100
4	80—90	99
3	70—80	98
2	60—70	97
1	60以下	96
0		



单项式分组　　组距式分组　　单项式分组

统计分组种类2----据标志的多少分

- 简单分组
  - 如把学生分别按性别、年龄、成绩分组.
- 复合分组
  - 如把学生按性别、年龄、成绩层叠分组.

等距组距式分组方法

1. 组数（number of classes）

组数不宜太多或太少，一般是5~10组

2. 组距（class of interval）

组距是一个组的范围，它等于全距/组数

等距组距式分组方法续

3. 组限（class limit）

A. 上限　　下限

〔a，b）,如〔70，80）,70含在该组中.

B. 组距=相邻两组的上（下）限之差

C. 开口组

D.全距=最大标志值—最小标志值（未分）

全距=最高组上限—最低组下限（已分）

组中值（class middle point)

1. 组中值=（上限+下限）/2
   组中值是一个代表值
2. 缺上限开口组组中值
   =下限+相邻组组距/2
3. 缺下限开口组组中值
   =上限—相邻组组距/2

如 60以下组
组中值= 60—（70 - 60）/2=55

频数分布表

分数	人数（人）	频率（%）
70—80	2	3.22
80—90	7	11.29
90—100	10	16.13
100—110	16	25.81
110—120	14	22.58
120—130	10	16.13
130—140	3	4.84
合计	62	100.00

标志表现　　频数　　频率

有关概念

A. 次数分配—总体单位按组分配的形式.

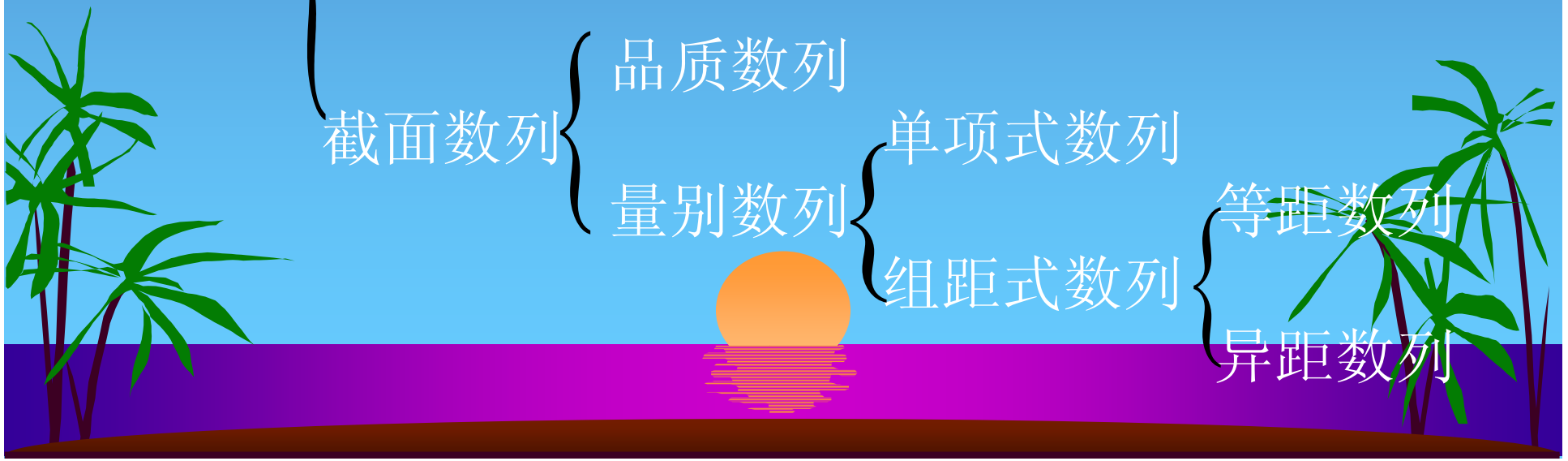
B .频数（次数）—分配在各组的总体单位数目.

C. 频率（相对频数、比重）—分配在各组的总体单位数目占总体单位总量的比重

D. 截面数列—又称频数分布数列、次数分配数列,是将所分各组依一定顺序排列,同时将各组频数（率）相应列出,反映总体中各单位在各组的分配状态和分布特征.

统计数列的种类

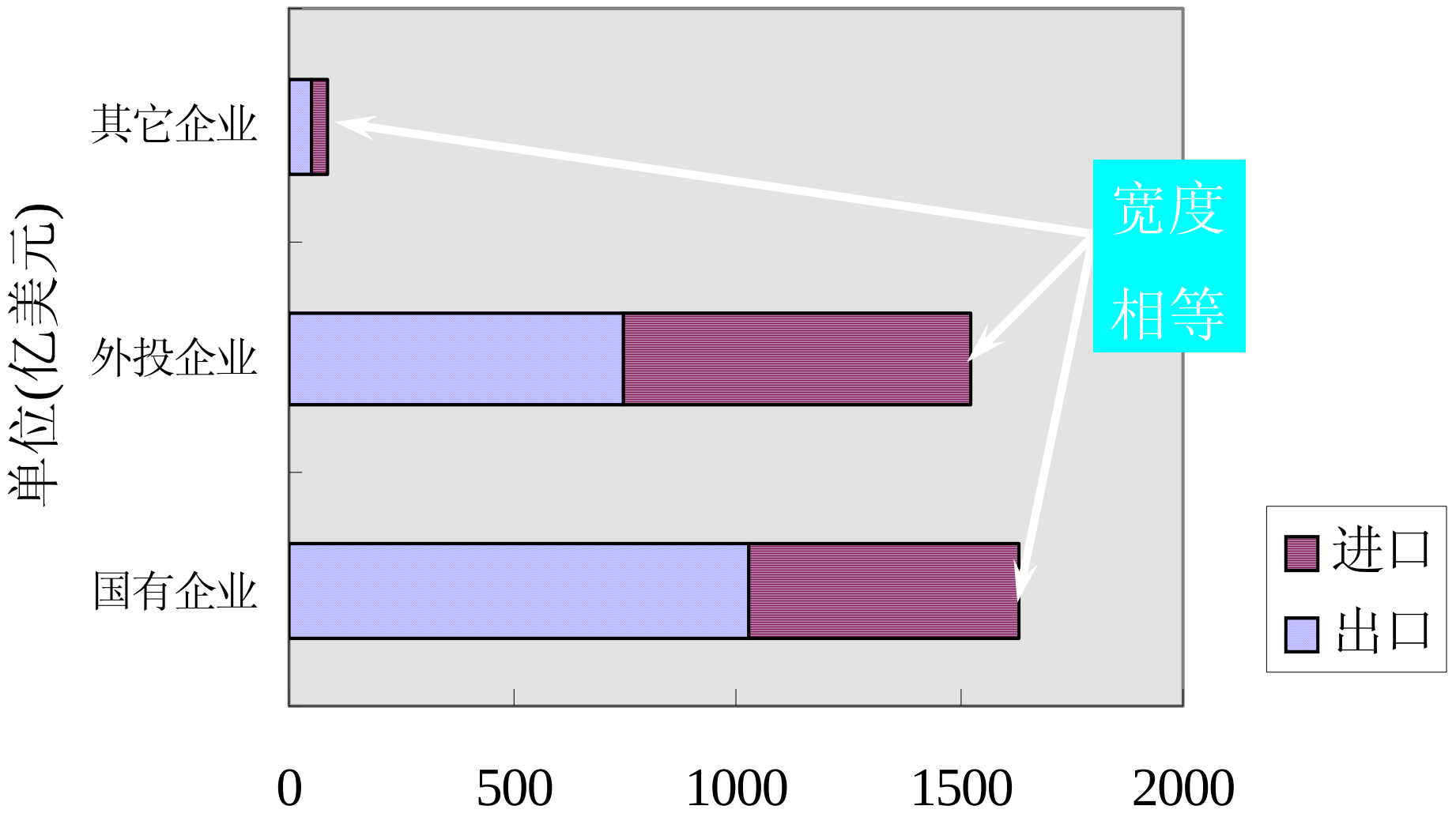
- 统计数列
  - 时间数列
    - 总量指标时间数列
      - 时期数列
      - 时点数列
    - 相对指标时间数列
    - 平均指标时间数列
  - 截面数列
    - 品质数列
    - 量别数列
      - 单项式数列
      - 组距式数列
        - 等距数列
        - 异距数列



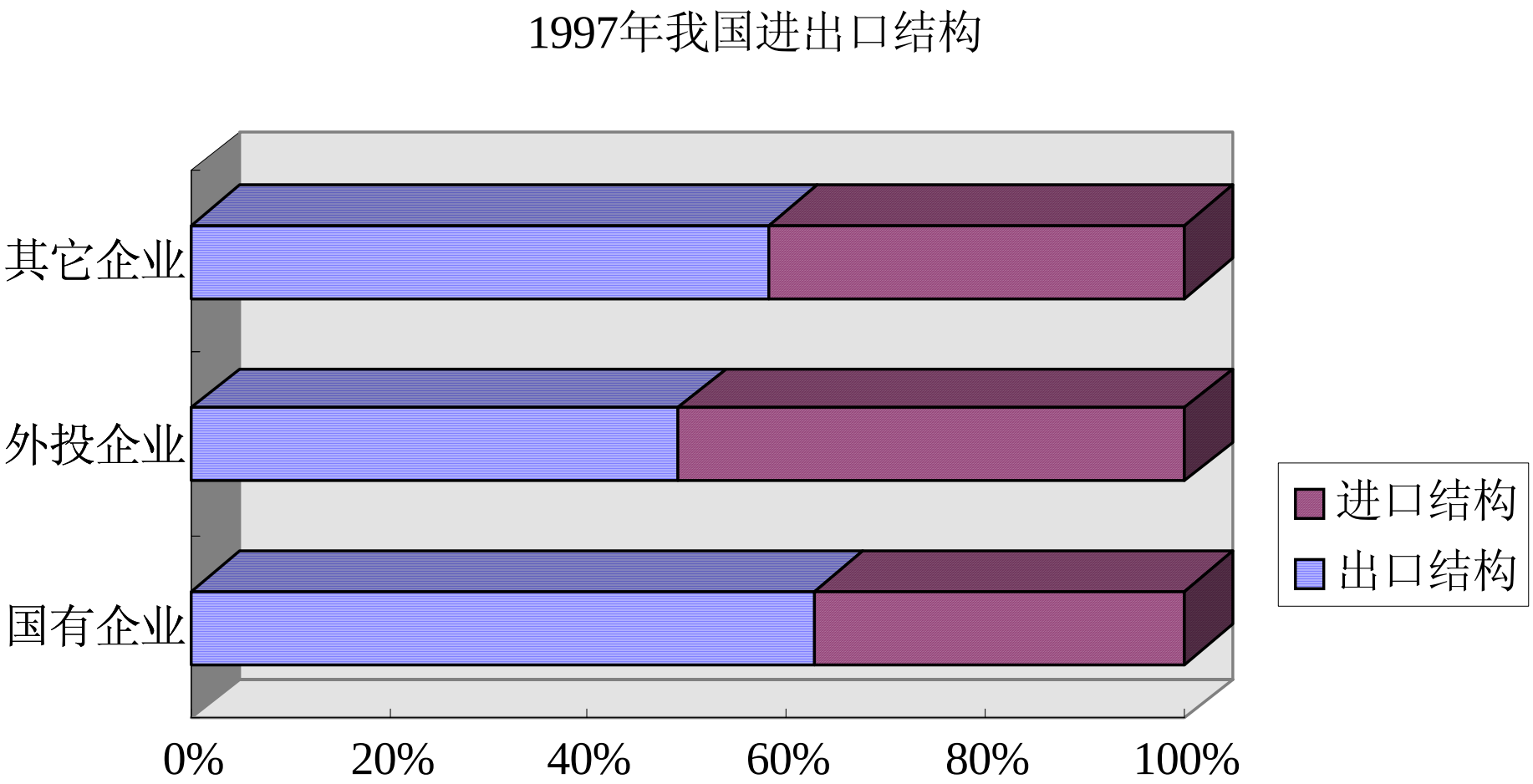
第三节　资料的表述 (统计表和统计图)

- 统计表
  - 表的结构(从表式和内容两个方面)
  - 表的种类
  - 表的设计
- 统计图
  - 条形图
  - 圆形图
  - 直方图,频数多变图,累计频数曲线图
  - 茎叶图,箱索图

条形图



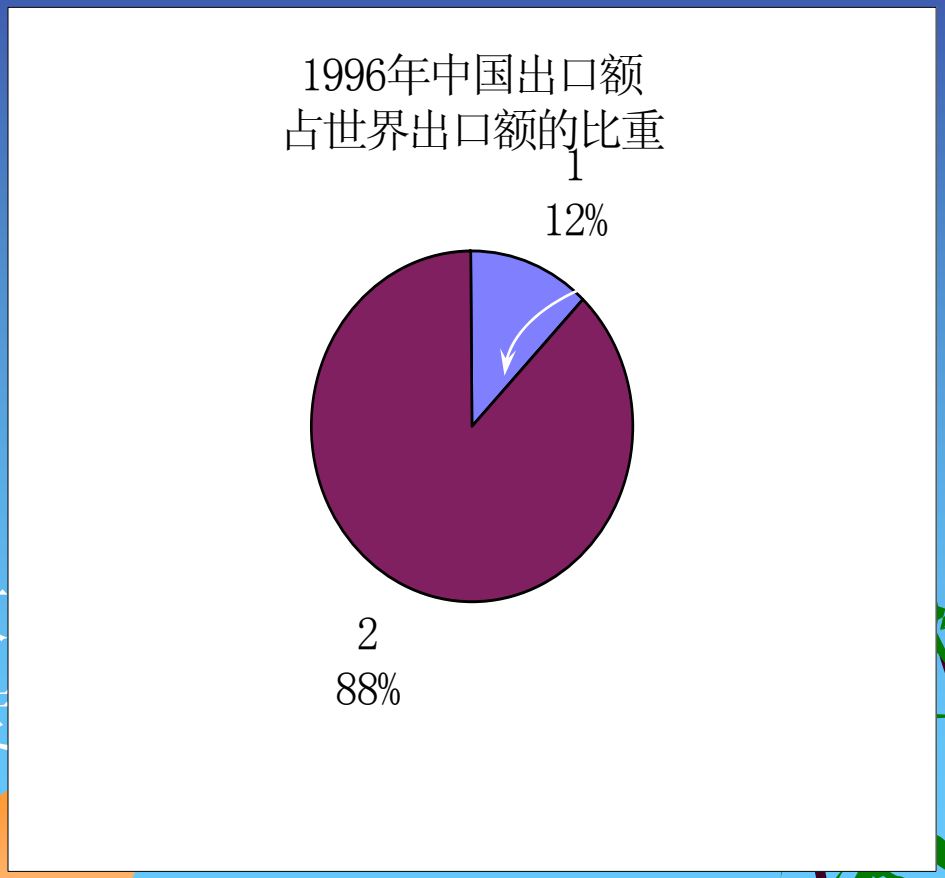
条形图



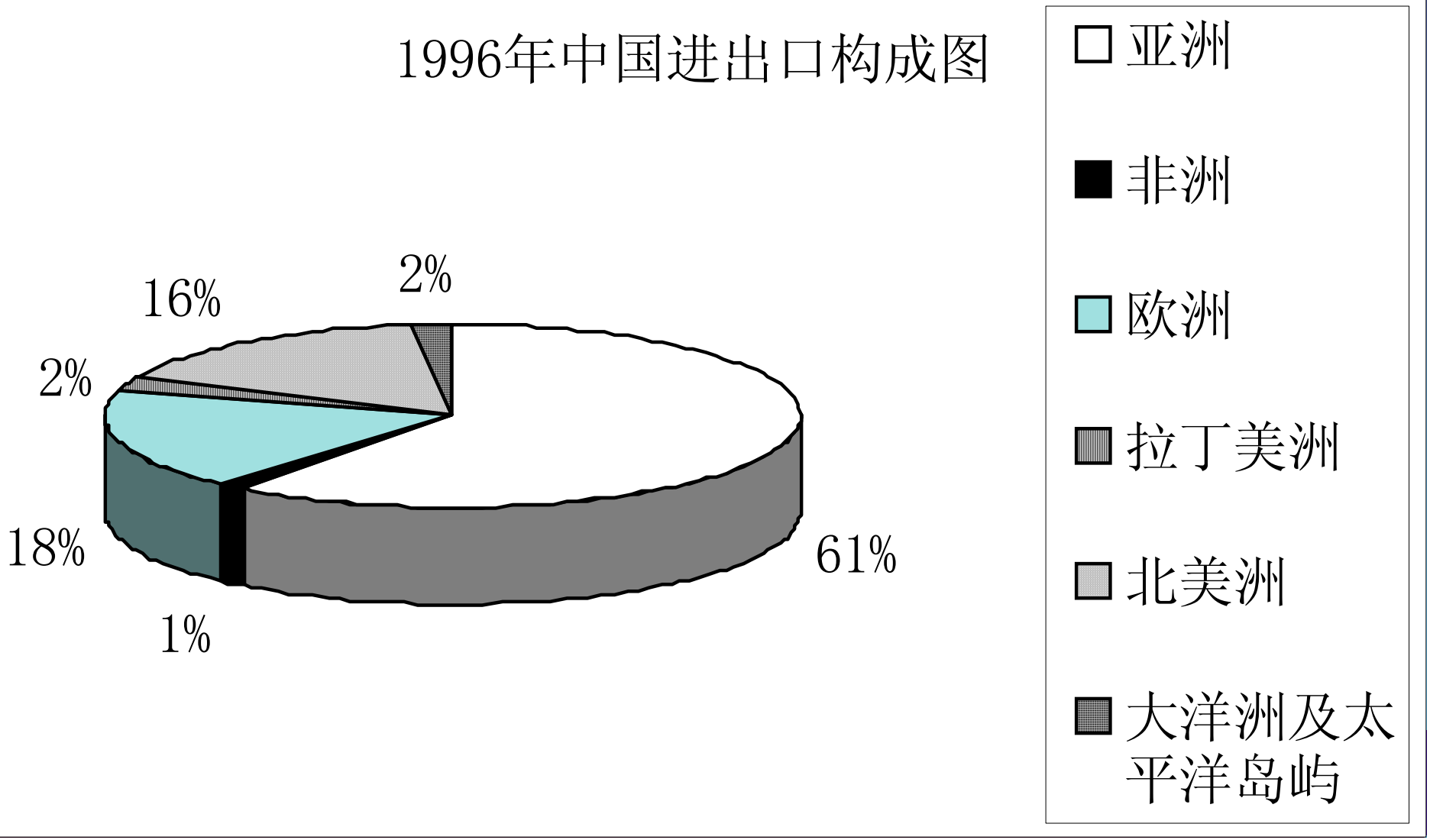
圆形图

- 最适用于反映总体的内部结构。
- 圆的面积等于100%
- 任何一部分的面积对应的圆心角等于（360度）*结构相对数

如：360*12%=43.2度



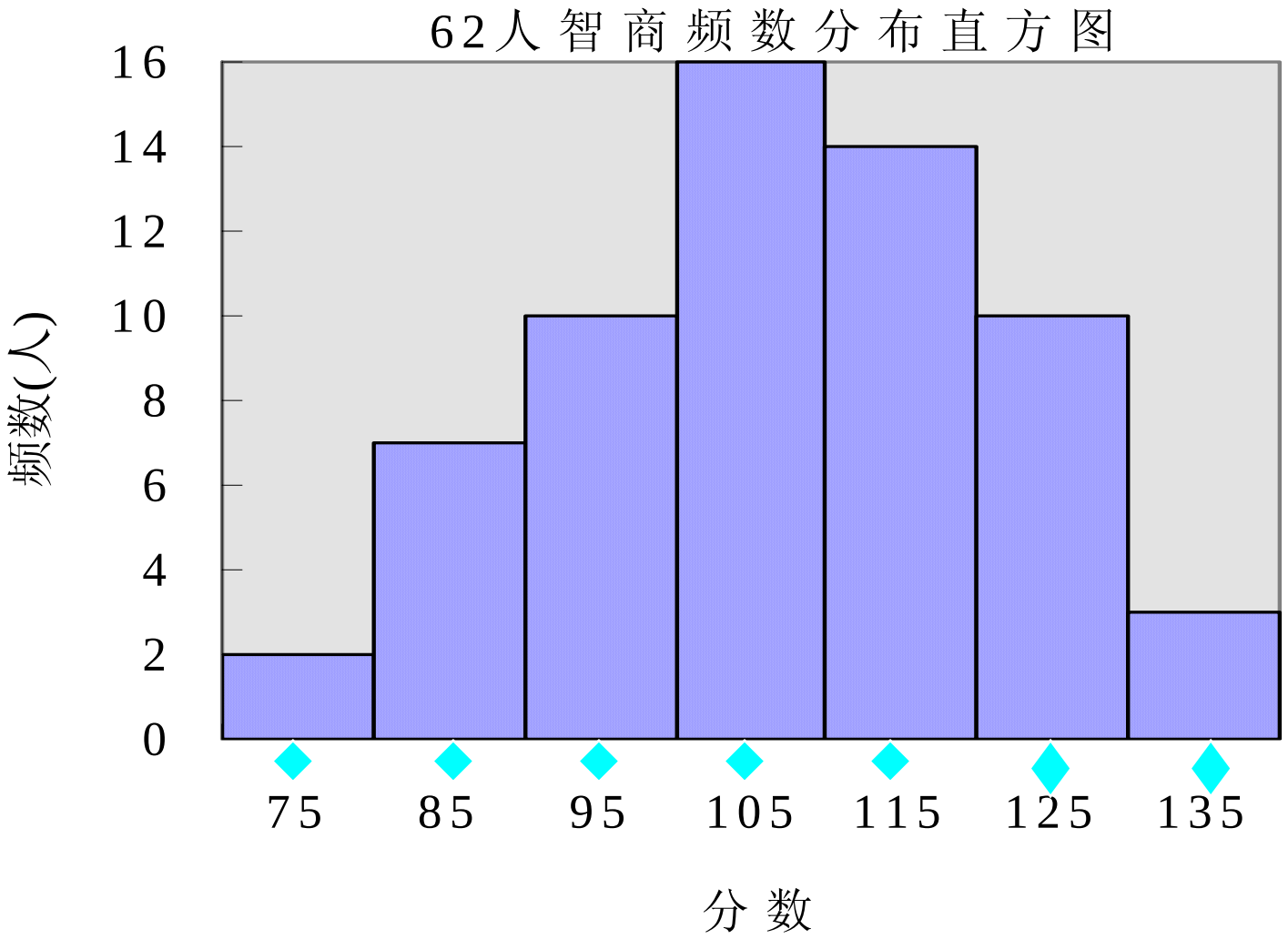
圆形图



62人皮尔逊知商得分频数分布表

分数	频数（人）
70--80	2
80--90	7
90--100	10
100-110	16
110-120	14
120--130	10
130--140	3
合计	62

直方图（反映变量数列的分布状况）

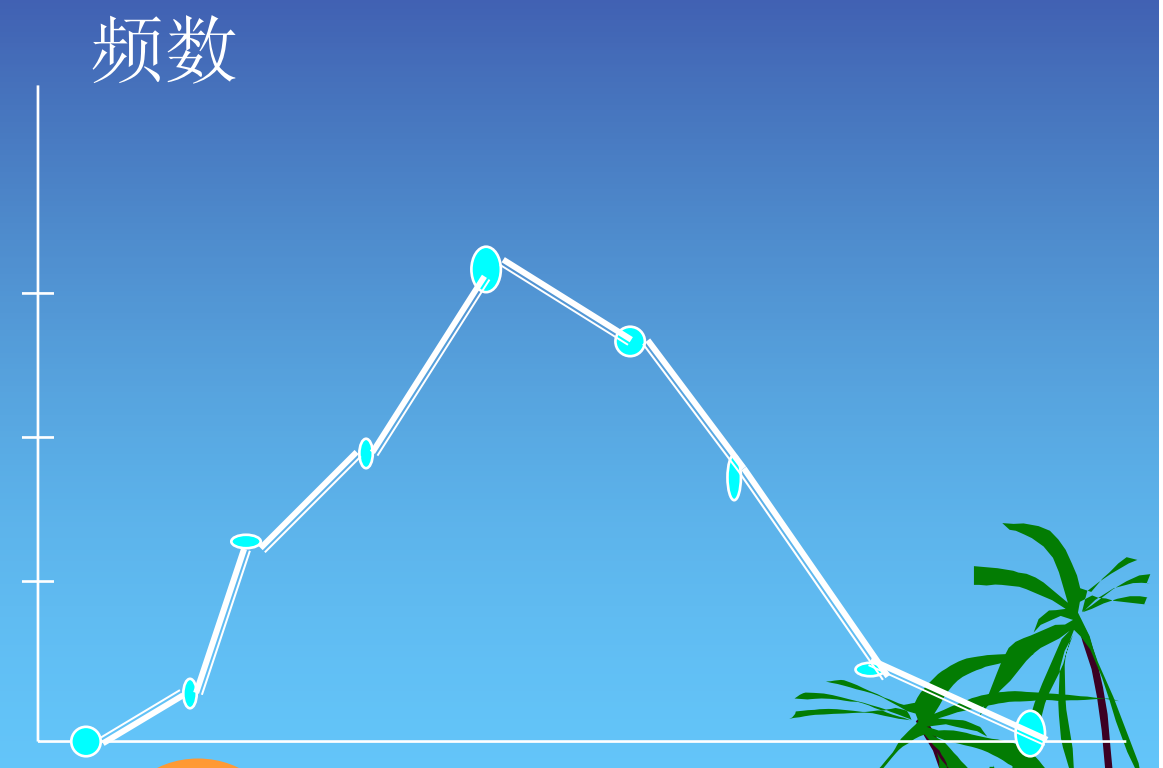


频数多边图

1.适用于频数分布表

2.反映变量数列的分布状况

3.可在同一个坐标系中反映和比较多个变量数列的分布状况.



职工工龄与售房政策态度列联表

工龄	赞成	反对	合计
10 年以下	21	12	33
10—20 年	9	9	18
20 年以上	10	19	29
合计	40	40	80

交叉分析交

第三章 数据的对比分析

第一节 静态对比分析（通过计算以下 相对数）

一，结构相对数

二，比例 相对数

三，类比 相对数

四，强度 相对数

第二节 动态 对比分析

一，发展速度

二，增长速度

第三节 评价 对比分析（计算 评价 相对数）

相对数的概念

相对数（relative），又称相对指标，是两个有联系的指标相比的比值或比率．它表明现象间的数量对比关系．

$$\text{相对数} = \frac{\text{比较对象}}{\text{比较基准}}$$

如我国外贸进出口总额1997年为3251亿美元，1996年为2899亿美元，则1997年为1996年的112.1%．

相对数的作用

1. 可以说明事物之间相互联系的程度和发展的程度. 如产品合格率, 列车正点率

2. 使原来无法比较的两个指标变得可比了. 如比较钢铁厂和纺织厂的发展情况.

相对数的计量形式

计量形式—无名数与有名数

- 无名数
  - 系数
  - 倍数
  - 成数
  - 百分数
  - 千分数

有名数一般是复合计量单位,如 人/平方公里，人/万人， 公里/平方公里

结构相对数

1. 定义：将总体按某一标志分组，以各组的总量指标数值与总体的总量指标数值相对比求得比重或比率，来反映总体内部组成状况。

2. 公式：结构相对数=$\frac{\text{某一组总量指标数值}}{\text{总体的总量指标数值}}$ ×100%

结构相对数的作用

1. 可以认识事物内部构成及发展变化趋势.
2. 在各级管理中，可以说明事物的运作情况.

注： 各组结构相对数之和为 100%.

比例相对数

1.定义：将总体按某一标志分组，以各组的总量指标数值相对比而求得的比值或比例

2.公式:比例相对数= $\dfrac{\text{总体中某一组的总量指标}}{\text{总体中另一组的总量指标}}$

类比相对数

1.定义： 把同一指标在相同时间、不同空间条件下的数值相对比的比值或比率.

2.公式:类比相对数= $\frac{\text{某一空间条件下的某指标}}{\text{另一空间条件下的某指标}}$

强度相对数

1.定义：统一总体或不同总体的两个性质不同但有一定联系的总量指标相比的比值

2.公式:

$$\text{强度相对数} = \frac{\text{某一总量指标}}{\text{另一有联系的总量指标}}$$

3.例如:人均产值,人口密度等 .

评价相对数(计划完成相对数）

$$\text{计划完成相对数} = \frac{\text{实绩完成数}}{\text{计划任务数}} * 100\%$$

计划任务数可以是绝对数，平均数和相对数 .

计划完成相对数的计算1

计划任务数是相对数时

如某厂去年计划提高利润 8%，实际提高了 10%，则

$$\text{计划完成相对数}=\frac{1+10\%}{1+8\%}=101.85\%$$

计划完成相对数的计算2

- 计划数是相对数时

如:某公司去年计划单位成本降低率为4.5%,而实际为5.5%.

则计划完成相对数为:

$$\frac{100\%-5.5\%}{100\%-4.5\%}=\frac{94.5\%}{95.5\%}=98.95\%$$

计划完成相对数结果的解释

是否超额的判断

正指标 >100%为超额，如销售额、利润额

逆指标 <100%为超额，如成本、费用

动态对比分析（本节应掌握的内容）

- 时间数列
- 发展水平
- 增减量（逐期 增减量，累积 增减量）
- 发展速度（ 定基 发展速度， 环比 发展速度）
- 增长速度（定基 增长速度，环比 增长速度）

时间数列，发展水平

时间数列：不同时间条件下的同一指标按照时 间的先 后顺序排列

发展水平
- 最初水平 a_0
- 中间水平 a_i
- 最末水平 a_n

增减量

增减量 = 报告期发展水平 – 基期发展水平

1. 逐期增减量 = $a_i - a_{(i-1)}$

2. 累积增减量 = 报告期发展水平 – 固定基期发展水平

$= a_i - a_0$

发展速度

$$\text{发展速度} = \frac{\text{报告期发展水平}}{\text{基期发展水平}} \times 100\%$$

1. 环比发展速度 $= \dfrac{\text{报告期发展水平}}{\text{上期发展水平}} = \dfrac{a_i}{a_{i-1}}$

如1995年进出口额环比发展速度 $= a_{1995} / a_{1994} = \dfrac{2808.63}{2165.56} = 129.7\%$

2. 定基发展速度 $= \dfrac{\text{报告期发展水平}}{\text{固定基期发展水平}} = \dfrac{a_i}{a_0}$

如1995年进出口额定基发展速度 $= \dfrac{a_{1995}}{a_{1990}} = \dfrac{2808.63}{851.18} = 329.97\%$

定基发展速度与环比发展速度的关系

- 定基发展速度等于 环比发展速度的连成积。

  1995年的 定基发展速度(329.97%)等于

  118.38%*128.77%*130.57%*127.82%*129.7%

- 环比发展速度等于

  本期定基发展速度/上年定基发展速度

1995年的环比发展速度=3.2997/2.5442=129.7%

1990---1995年中国进出口额

单位:亿美元

时间	进出口额	增减量		发展速度	
		逐步	累积	环比	定基
1990	851.18			100	100
1991	1007.65	156.47	156.47	118.38	118.38
1992	1297.53	289.88	446.35	128.77	152.44
1993	1694.23	396.7	843.05	130.57	199.04
1994	2165.56	471.33	1314.38	127.82	254.42
1995	2808.63	643.07	1957.45	129.7	329.97

增长速度

$$= \frac{\text{增减量}}{\text{报告期发展水平}} = \frac{\text{报告期发展水平} - \text{基期发展水平}}{\text{基期发展水平}}$$

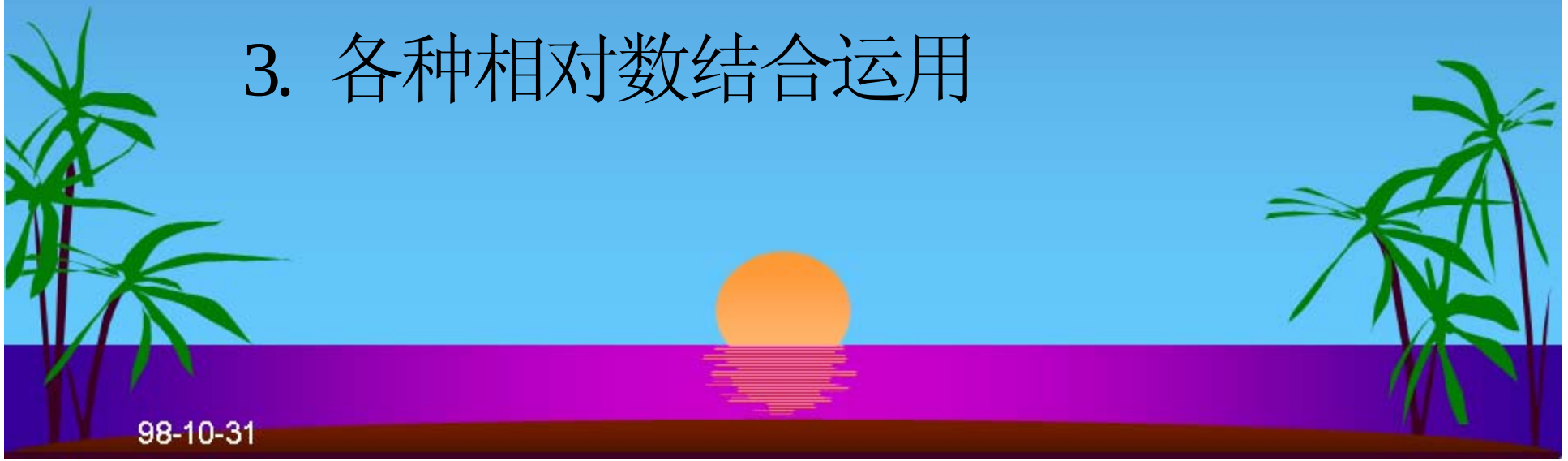
$$= \frac{\text{报告期发展水平}}{\text{基期发展水平}} - 100\% = \text{发展速度} - 100\%$$

定基 增长速度=定基发展速度-100%

环比 增长速度=环比发展速度-100%

对比分析应遵循的原则

1. 保持对比双方的可比性
2. 相对数要和绝对数结合起来使用
3. 各种相对数结合运用



对比分析应注意的问题

1. 除结构相对数外，其它相对数不能简单相加.

2. 比例、比较和强度相对数的分子和分母可以互换.

3. 当相对比的数值太小时，不宜用相对数.

4. 计算强度相对数要注意二者是否有联系.

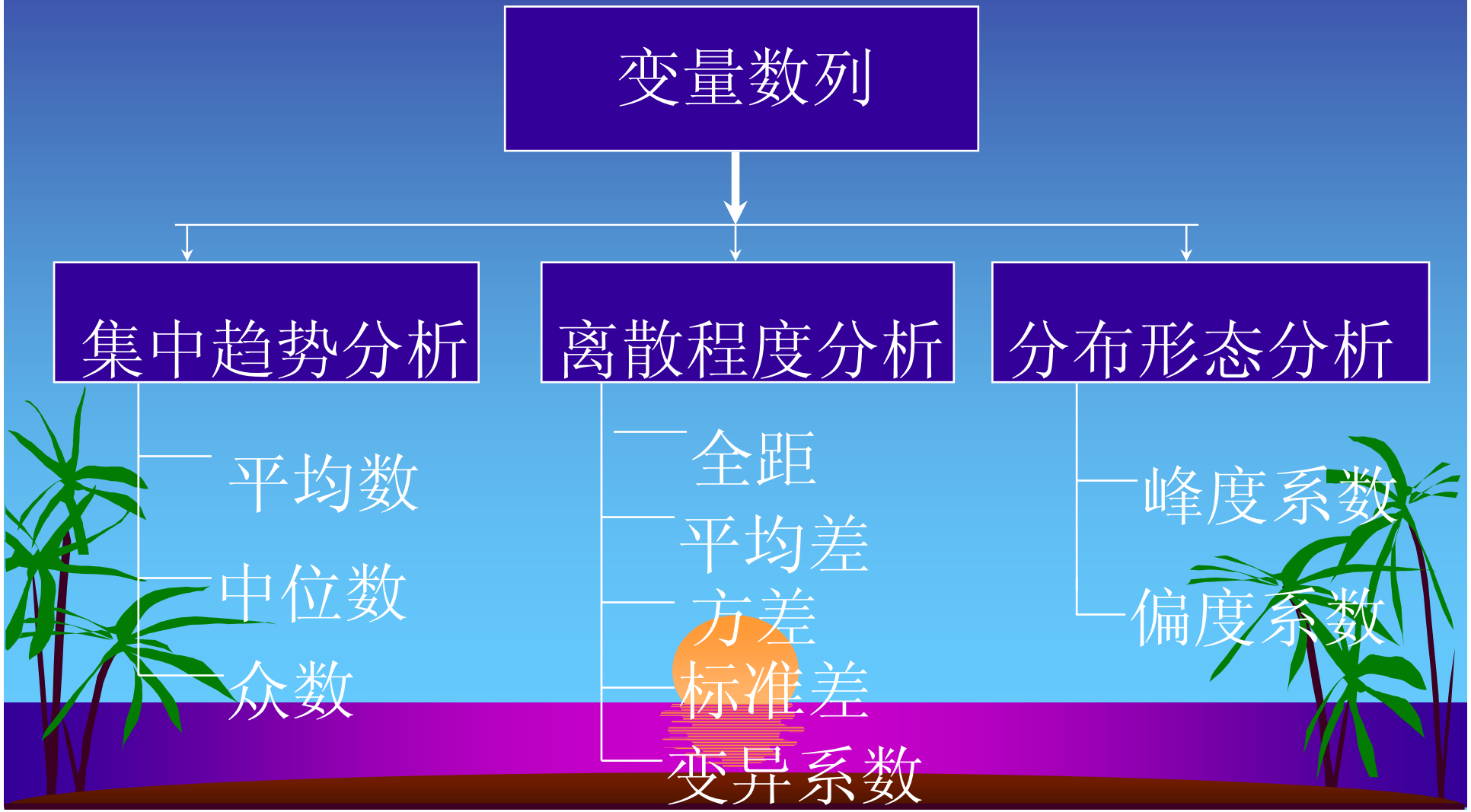
本章重点

- 静态对比分析（通过计算以下 相对数）
  - 一，结构相对数
  - 二，比例 相对数
  - 三，类比 相对数
  - 四，强度 相对数
- 动态 对比分析
  - 一，发展速度
  - 二，增长速度
- 评价 对比分析（计算 评价 相对数）

第四章 单变量截面数据的描述性分析（本章重点）

- 了解描述 截面数据分布中心的指标
  - 平均数
  - 中位数
  - 众数
- 了解描述 截面数据分散程度的指标
  - 全距，平均差
  - 方差，标准差，变异系数
- 位次指标

分析单变量截面数据特征的指标

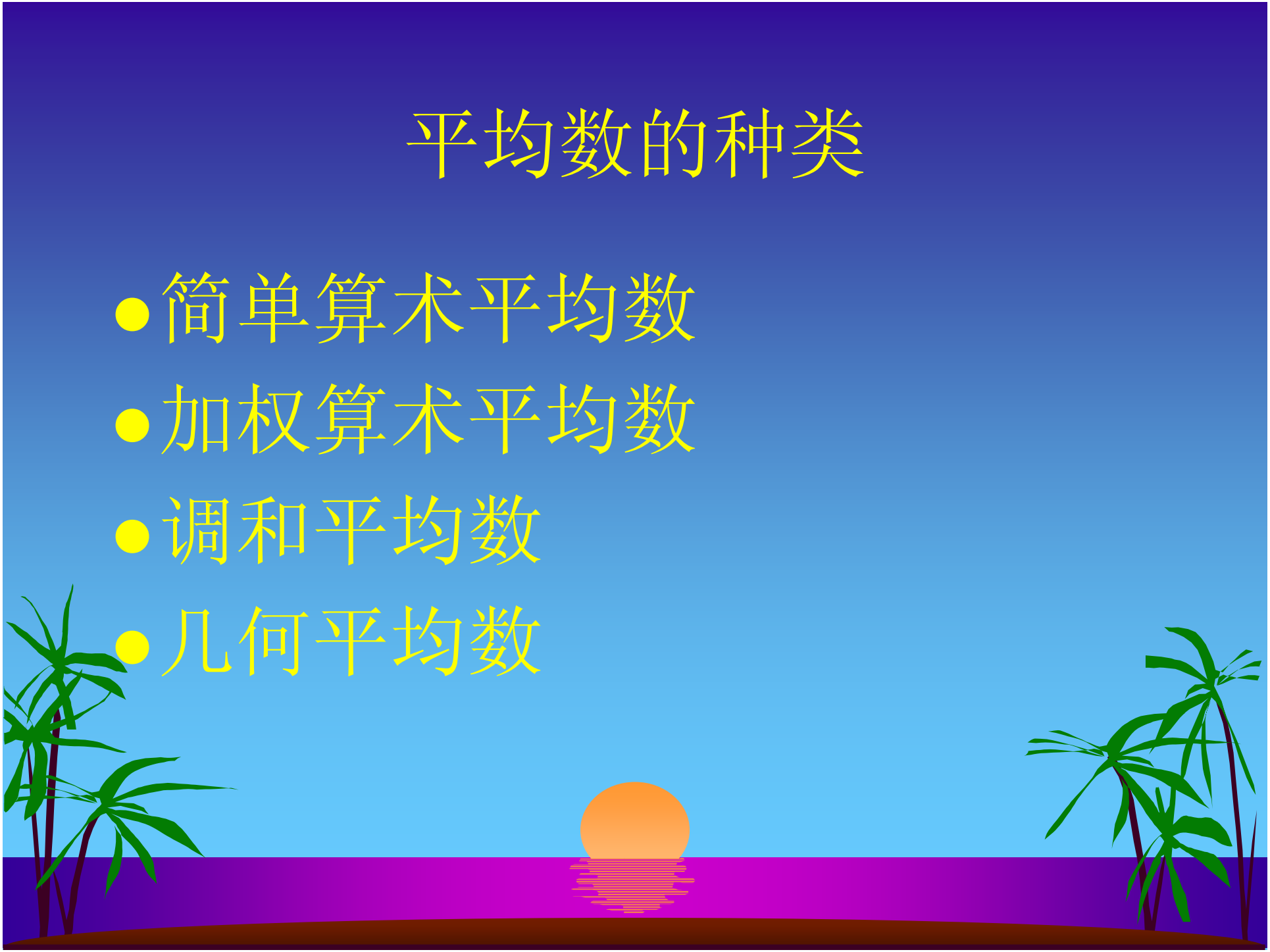


集中趋势指标----平均数

- 衡量变量数列分布中心的指标
- 最常用的 集中趋势指标
- 容易受极端值的影响
  - 极端值：远离分布中心的数值

平均数的种类

- 简单算术平均数
- 加权算术平均数
- 调和平均数
- 几何平均数



简单算术平均数

- 公式：$\overline{X}=\frac{X_1+X_2+\Lambda+X_n}{n}=\frac{\sum_{i=1}^{n}X_i}{n}=\frac{\sum X}{n}$

$$=X_1\frac{1}{n}+X_2\frac{1}{n}+\Lambda+X_n\frac{1}{n}$$

- 适用情况
  1. 资料未分组
  2. 每一个标志值的作用相同
- 影响平均数大小的因素只有标志值

加权算术平均数

- 定义:将各变量值分别乘以代表该变量值重要程度的权数,然后用此乘积之和除以权数之和,所得的商为 加权算术平均数.
- 公式:

$$\overline{X}=\frac{X_1W_1+X_2W_2+\Lambda+X_kW_k}{W_1+W_2+\Lambda+W_k}=\frac{\sum_{i=1}^{k}X_iW_i}{\sum_{i=1}^{k}W_i}=\frac{\sum XW}{\sum W}$$

$$=X_1\frac{W_1}{\sum W}+X_2\frac{W_2}{\sum W}+\Lambda+X_k\frac{W_k}{\sum W}$$

加权算术平均数

- 适用情况
  1. 资料已分组
  2. 每一个标志值的作用不同
- 权数的确定方法
  1.主观确定法(专家确定)
  2.客观存在(频数分布表中的相对频数)
- 影响平均数大小的因素有
  1.标志值的大小
  2.权数的大小

求加权算术平均数的例题1

- 原始数据:某企业用两个指标考核职工,他们的成绩如下:

职工	考勤	产品质量
A	60	90
B	90	60

- 求两位职工的综合成绩
  1.用简单算术平均数
  2.用加权算术平均数

例题1计算结果

1. 用简单算术平均数

$$\bar{X}_A = \frac{60+90}{2} = 60\frac{1}{2} + 90\frac{1}{2} = 75$$

$$\bar{X}_B = (90+60)/2 = 75$$

2. 用加权算术平均数

   1. W_1(出勤的权数)为40%,W_2为60%

$$\bar{X}_A = 78 \qquad \bar{X}_B = 72$$

   2. W_1=60%,W_2=40% $\bar{X}_A = 72, \quad \bar{X}_B = 78$

求加权算术平均数的例题2

如150名工人中，10人每天生产15件，20人每天生产16件，40人每天生产17件，50人每天生产18件，30人每天生产19件.

$$\bar{X}=\frac{\sum Xf}{\sum f}=\frac{15*10+16*20+17*40+18*50+19*30}{150}$$

$$=15*\frac{10}{150}+16*\frac{20}{150}+17*\frac{40}{150}+18*\frac{50}{150}+19*\frac{30}{150}$$

$$=17.47$$

求加权算术平均数的例题3

62人皮尔逊智商分数平均数计算表

分数	人数(人)f	组中值X	Xf	f/62	X*(f/62)
70-80	2	75	150	0.0323	2.419
80-90	7	85	595	0.1129	9.597
90-100	10	95	950	0.1613	15.323
100-110	16	105	1680	0.2581	27.097
110-120	14	115	1610	0.2258	25.968
120-130	10	125	1250	0.1613	20.161
130-140	3	135	405	0.0484	6.532
合计	62	-	6640	1	107.097

例题3的计算结果

$$\overline{X}=\frac{\sum Xf}{\sum f}=\frac{75*2+85*7+\Lambda+135*3}{62}$$

$$=75*0.0323+85*0.1129+\Lambda$$

$$+135*0.0484$$

$$=107.1$$

调和平均数(倒数平均数)

1. 定义:标志值倒数的平均数的倒数
2. 公式:(简单调和平均数)

$$\overline{X}_H = \frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\Lambda+\frac{1}{x_n}}{n}} = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\Lambda+\frac{1}{x_n}} = \frac{n}{\sum\frac{1}{x}}$$

简单调和平均数例题

例如，某地三个农贸市场芹菜的市场价分别为 0.5 元、 0.6 元和 0.8 元，那么

（1）若在三市场各买 1 斤芹菜，平均每斤多少钱？

（2）若在三市场各买一元钱的芹菜，平均每斤多少钱？

简单调和平均数例题的计算结果

$$\text{平均单价} = \frac{\text{总金额}}{\text{总数量}}$$

$$1.\ \text{平均单价} = \frac{1\times 0.5+1\times 0.6+1\times 0.8}{1+1+1} = 0.61(\text{元})$$

$$= \frac{\sum x}{n}(\text{算术平均数})$$

$$2.\ \text{平均单价} = \frac{1+1+1}{\frac{1}{0.5}+\frac{1}{0.6}+\frac{1}{0.8}} = 0.61(\text{元})$$

$$= \frac{n}{\sum \frac{1}{x}}(\text{调和平均数})$$

加权调和平均数公式

$$\overline{X}_h = \frac{1}{\frac{1}{x_1}w_1 + \frac{1}{x_2}w_2 + \Lambda + \frac{1}{x_i}w_i} = \frac{\sum w}{\sum \frac{w}{x}}$$

加权调和平均数例题

某企业集团有三个子公司,她们的销售净利润率和净利润资料如下:

公 司	销 售 净 利 润 率 （ % ） X	净 利 润 额 （ 亿 元 ） W
A	7	7
B	8	8
C	9	1 8

求集团的平均利润率

加权调和平均数例题计算结果

x: 利润率; w: 利润额

$$\text{平均利润率} = \frac{\text{利润总额}}{\text{销售总额}}$$

$$\text{平均利润率} = \frac{\sum w}{\sum \frac{w}{x}} = \frac{7+8+18}{\frac{7}{0.07}+\frac{8}{0.08}+\frac{18}{0.09}} = 8.25\%$$

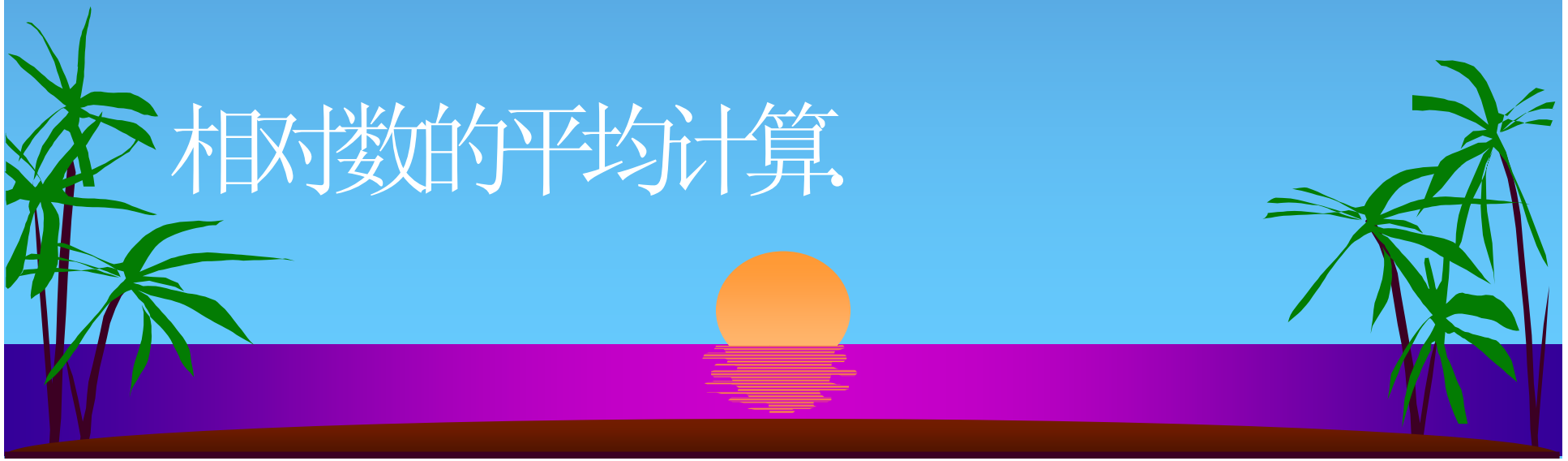
注:调和平均数适用于分母资料未知时计算平均数

几何平均数

公式: $\overline{X}_g = \sqrt{x_1 \times x_2 \times \Lambda \times x_i} = \sqrt{\Pi x}$

几何平均数适用于比例和速度等

相对数的平均计算.



几何平均数的应用1

如： 某流水线有四个工序，第 1 、 2 、 3 和 4 工序的产品合格率分别为 98%、92%、90%和 93%，求平均各工序的合格率.

$$\bar{X}_g = \sqrt[4]{0.98 \times 0.92 \times 0.90 \times 0.93} = 93.2\%$$

几何平均数的应用2

1990-1995年中国进出口额环比发展速度如下

时间:	1990	1991	1992	1993	1994	1995
环比速度:	—	118.38	128.77	130.57	127.82	129.7

求1990----1995年 中国进出口额平均发展速度

$$\text{发展速度} = \sqrt[5]{118.38*128.77*130.57*127.82*129.7} = 126.97\%$$

发展速度

发展速度 $= \sqrt[5]{118.38*128.77*130.57*127.82*129.7} = 126.97\%$

$$= \sqrt[5]{\frac{a_{1991}}{a_{1990}} \times \frac{a_{1992}}{a_{1991}} \times \frac{a_{1993}}{a_{1992}} \times \frac{a_{1994}}{a_{1993}} \times \frac{a_{1995}}{a_{1994}}}$$

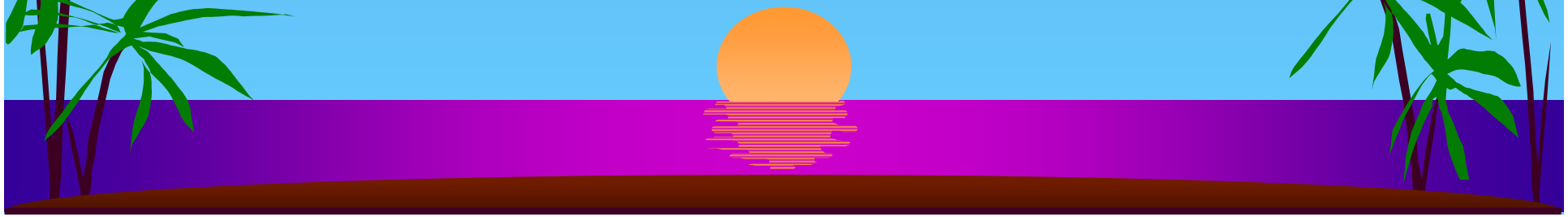
$$= \sqrt[5]{\frac{a_{1995}}{a_{1990}}} = \sqrt[5]{\frac{2808.63}{85118}} = 126.97\%$$

中位数(Median)

将变量数列的各观察值按自小到大的顺序排列，处于中间位置的数值即为中位数.

中位数所在的位置项数=（ｎ＋１）/2

当数列中有极端值存在时，采用中位数求变量值的一般水平比用算术平均数好.



中位数计算举例

原始资料: 10.3　4.9　8.9　11.7　6.3　7.7

按顺序排列: 4.9　6.3　7.7　8.9　10.3　11.7

位置:　　1　　2　　3　　4　　5　　6

中位数所在的位置为: $\frac{n+1}{2} = \frac{5+1}{2} = 3.5$

中位数=(7.7+8.9)/2=8.3

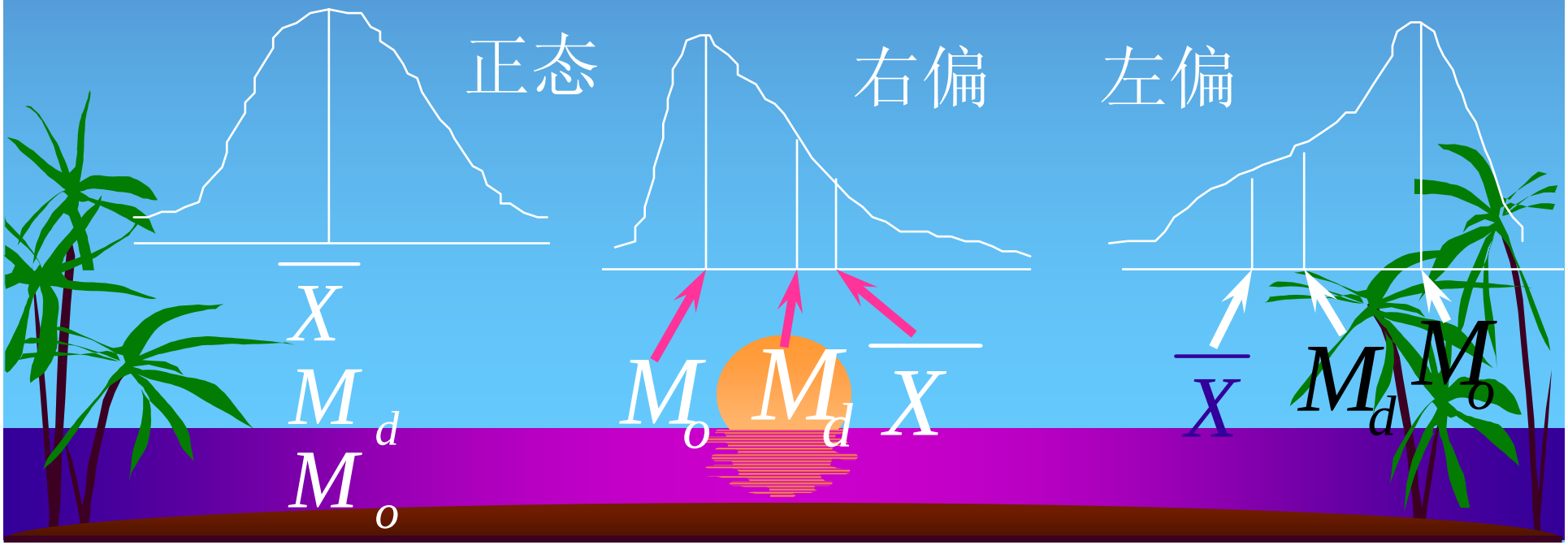
众数(Mode)

- 出现次数最多的那个变量值
- 是一个常用的集中趋势指标
- 它不受极端值的影响
- 并非所有的数列都存在众数

平均数 中位数 众数的关系

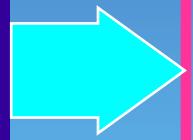
1.正态分布时 $\overline{X} = M_d = M_o$

2.偏态时 $\overline{X} - M_o = 3(\overline{X} - M_d)$



离散趋势指标

反映变量数列分散程度的指标



- 全距
- 平均差
- 方差
- 标准差
- 变异系数

全距

- 全距=最大值-最小值
- 原始资料：17 16 21 18 13 16 12 11
- 顺序排列：11 12 13 16 16 17 18 21
- 全距=21-11=10

注意：全距只考虑了两个值的距离,如果数列中存在极端值,它会片面地夸大数列的分散程度.

平均差（平均绝对差）

$$M.A.D = \frac{\sum |X - \overline{X}|}{n}$$

优点：考虑了每个标志值对平均数的离差.

缺点:公式中带绝对值不便进行数学运算

方差

1. 总体方差 $\sigma^2 = \dfrac{\sum (X-\mu)^2}{N}$ (资料未分组)

$\sigma^2 = \dfrac{\sum f(X-\mu)^2}{\sum f}$ (资料已分组)

2. 样本方差 $S^2 = \dfrac{\sum (X-\overline{X})^2}{n-1}$ (资料未分组)

$S^2 = \dfrac{\sum f(X-\overline{X})^2}{\sum f - 1}$ (资料已分组)

样本方差计算1(未分组)

- 原始数据: 17 16 21 18 13 16 12 11

$$S^2 = \frac{\sum (X - \overline{X})^2}{n-1} \qquad \overline{X} = \frac{\sum X}{n} = 15.5$$

$$S^2 = \frac{(17-15.5)^2 + (16-15.5)^2 + \Lambda + (11-15.5)^2}{8-1}$$

$$= 11.14$$

样本方差计算2(已分组)

分组	组中值X	频数f	Xf	X-X	$(X-X)^2f$
70-80	75	2	150	-32.1	2060.82
80-90	85	7	595	-22.1	3418.87
90-100	95	10	950	-12.1	1464.1
100-110	105	16	1680	-2.1	70.56
110-120	115	14	1610	7.9	873.74
120-130	125	10	1250	17.9	3204.1
130-140	135	3	405	27.9	2335.23
合计		62	6640		13427.42

样本方差计算2续(已分组)

$$S^2 = \frac{\sum f\left(X - \overline{X}\right)^2}{\sum f - 1} = \frac{13427}{62 - 1} = 220.11$$

优点:克服了平均差和全距的缺点,保存了它们的优点.

缺点:方差的单位带平方,并非所有的单位带平方都有意义.

标准差(方差的平方根)

1. 总体标准差 $\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum f\left(X-\mu\right)^2}{\sum f}}$

2. 样本标准差 $S = \sqrt{S^2} = \sqrt{\dfrac{\sum f\left(X-\overline{X}\right)^2}{\sum f - 1}}$

1. 克服了平均差,全距和方差的缺点,保存了它们的优点； 2. σ 和 μ 之间存在数理关系

变异系数

- 用标准差比较两个总体分散程度时必须具备以下条件
  - 单位相同
  - 数据总体水平相同
- 否则必须用变异系数

$$\text{变异系数} C.V = \frac{S}{\bar{X}} \quad \text{或} C.V = \frac{\sigma}{\mu}$$

变异系数应用举例

$n_1 = 100$头羊　$\bar{X} = 50$公斤　$S = 10$公斤

$n_2 = 100$头牛　$\bar{X} = 300$公斤　$S = 10$公斤

问: 哪个样本更集中?

$$C.V_1 = \frac{10}{50} = 0.2 \qquad C.V_2 = \frac{10}{300} = 0.033$$

$\because C.V_1 > C.V_2$　∴ 羊的样本比牛的样本分散

集中趋势指标与离散程度指标的关系

总体中各标志值离集中趋势指标远，那么集中趋势指标代表性就小。

- 离散程度指标大，说明总体分散或者说

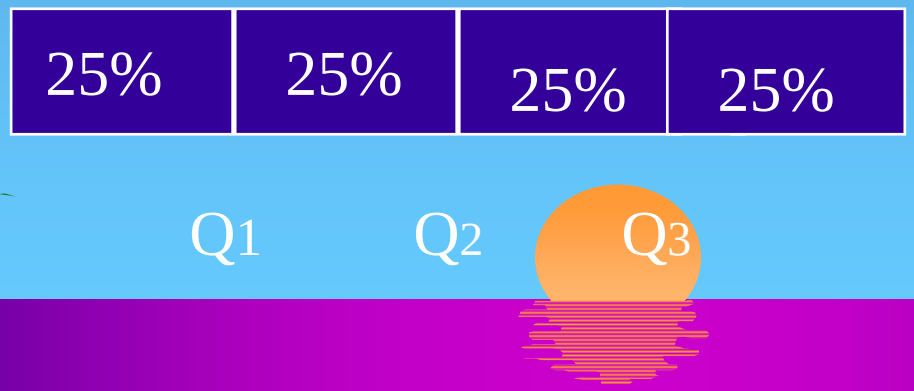
- 离散程度指标小，说明总体集中或者说总体中各标志值离集中趋势指标近，那么集中趋势指标代表性就大。

位次指标

位次指标： 根据观察值在变量数列中的位置而确定的指标

1. 它不是一个集中趋势指标

2.把变量数列(从小到大排列)分成四等份



四分位数的确定

原始数据：10.3 4.9 8.9 11.7 6.3 7.7

按顺序排列： 4.9 6.3 7.7 8.9 10.3 11.7

位置： 1 2 3 4 5 6

第i分位数的位置公式为： Q_i位置= I(n+1)/4

Q_1 位置 =1(n+1)/4=1(6+1)/4=1.75

Q_1=6.3

Q_2位置=2(n+1)/4=2(6+1)/4=3.5

Q_2=(7.7+8.9)/2=8.3

Q_3位置=3(n+1)/4=3(6+1)/4=5.25

Q_3=10.3

本章重点

- 了解描述 截面数据分布中心的指标
  - 平均数
  - 中位数
  - 众数
- 了解描述 截面数据分散程度的指标
  - 全距，平均差
  - 方差，标准差，变异系数
- 位次指标

第六章 时间数列

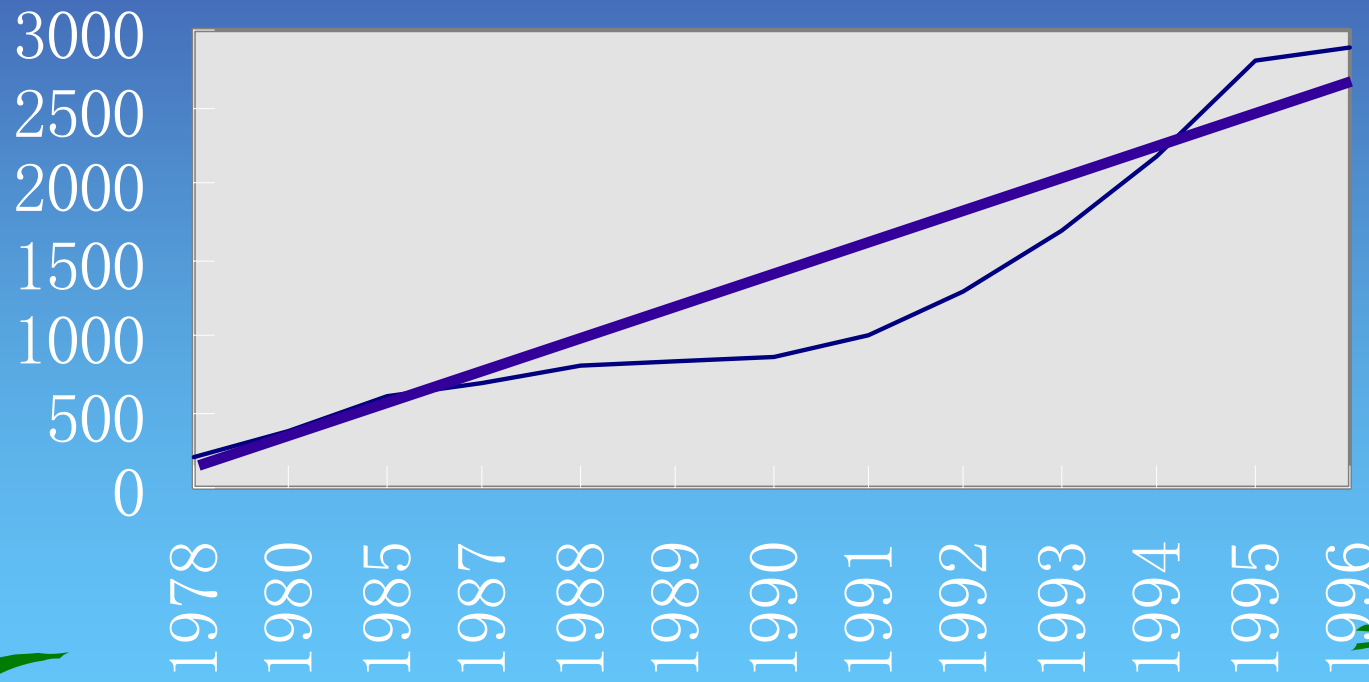
分析方法
- 指标法
  - 绝对比较法
  - 相对比较法
- 数学模型法 （本章内容）

本章重点

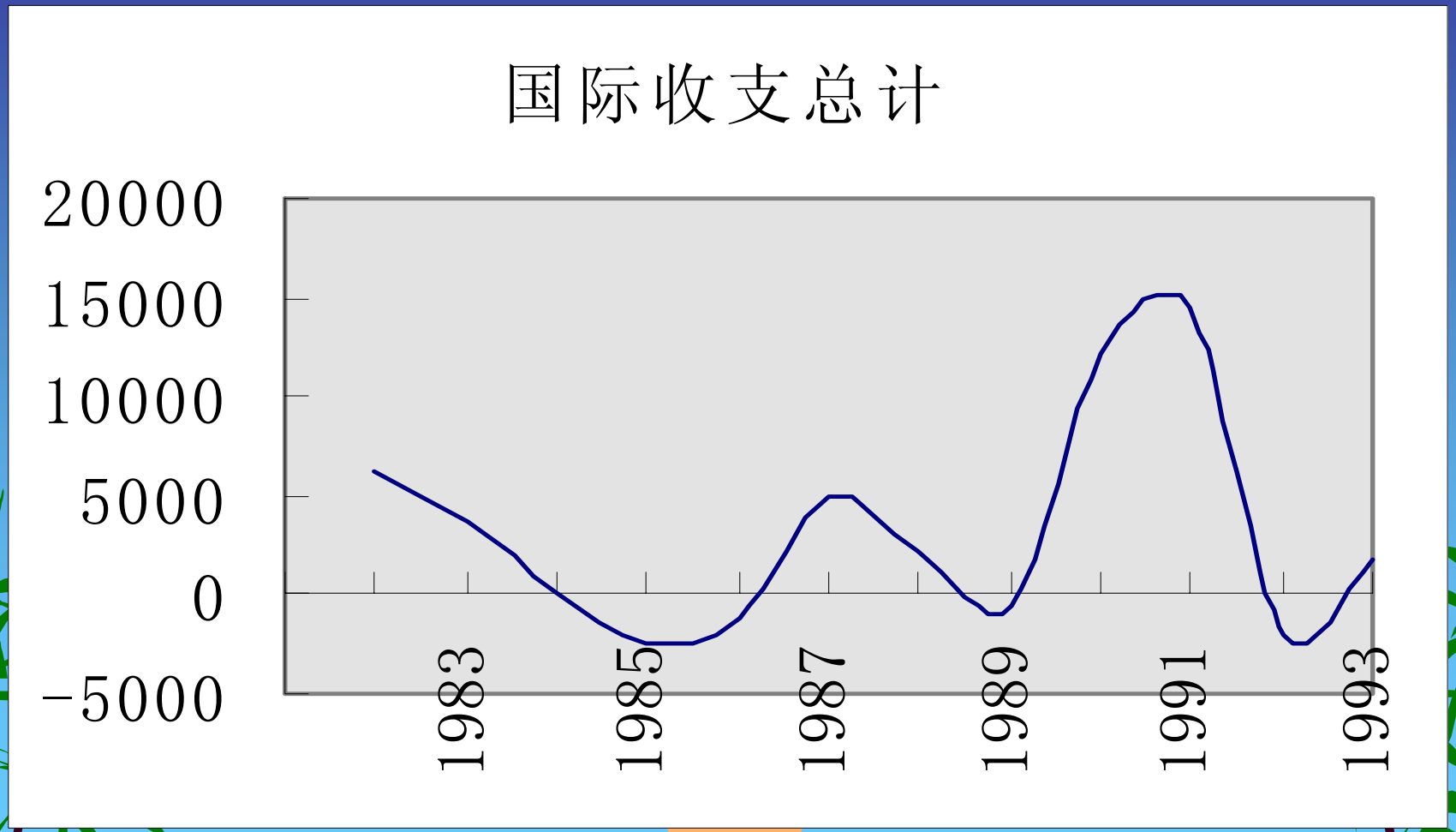
- 了解影响 时间数列变动的因素
  - 长期趋势
  - 季节性变动
  - 循环变动
  - 不规则变动
- 如何根据时间数列测定长期趋势值
  - 建立长期趋势方程
- 如何根据时间数列测定季节性变动
  - 计算季节指数

长期趋势

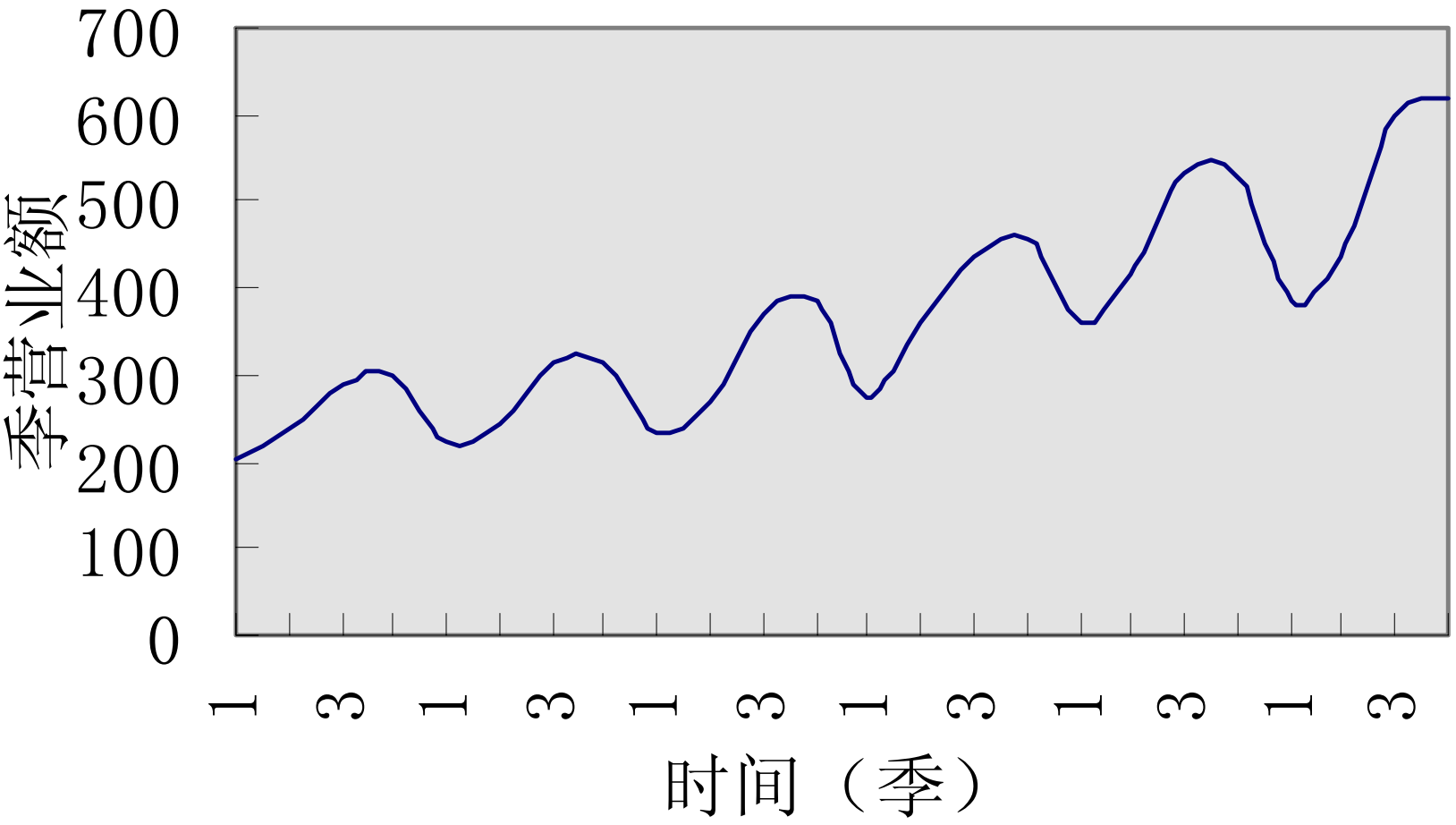
进出口额


3000
2500
2000
1500
1000
500
0
1978
1980
1985
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996

循环变动


国际收支总计
20000
15000
10000
5000
0
-5000
1983
1985
1987
1989
1991
1993

美国迪斯尼公司1983--88的季营业额



时间数列的解释模式

- 加法模式　Y=T+S+C+I
  - Y：发展水平
  - T：　长期趋势
  - S ：季节性变动
  - C：循环变动
  - I：不规则变动
- 乘法模式：Y=T*S*C*I

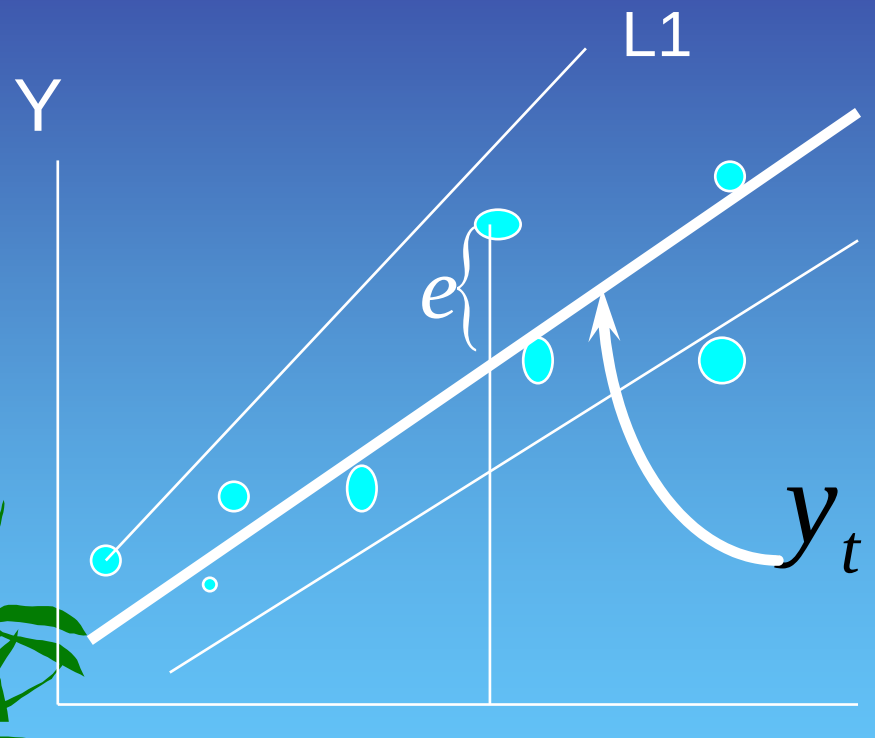
Y为绝对数

S
C　为相对数
I

长期趋势的测定

- 移动平均法
- 建立长期趋势方程（重点）
  - 线性长期趋势方程（ 重点）
    - 方程的建立
    - 方程的应用
      - 能测定各年的长期趋势值
      - 能根据方程进行预测
  - 非线性长期趋势方程
    - 指数曲线
    - 抛物线

线性长期趋势方程的建立 最小平方法



L2直线最好

$\sum e^2$ 最 小

$$y_t(T) = a + bX_t$$

线性长期趋势方程的建立的步骤

- 设方程为：$y_t = a + bX_t$
- 计算a和b

$$b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-\left(\sum X\right)^2}$$

$$a=\bar{Y}-b\bar{X}$$

- 方程的应用

长期趋势方程建立的例题1

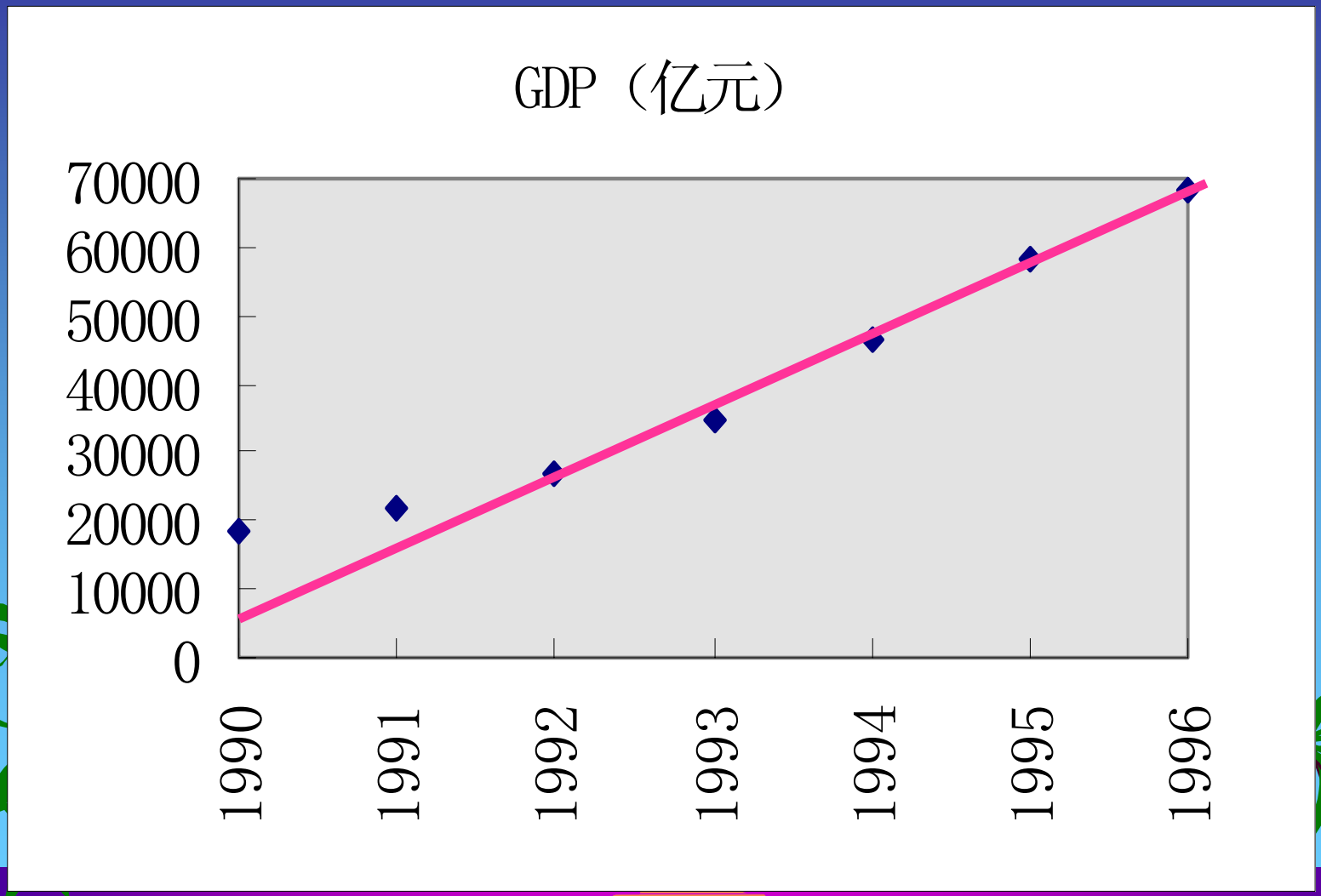
时 间	GDP（亿元）
1990	18547.9
1991	21617.8
1992	26638.1
1993	34634.4
1994	46759.4
1995	58478.1
1996	68593.8

1.建立GDP长期趋势方程

2。解释b的经济含义

3。预测2000年的GDP

例题1的趋势图


GDP（亿元）
70000
60000
50000
40000
30000
20000
10000
0
1990
1991
1992
1993
1994
1995
1996

例题1的计算过程

时 间	GDP（亿元）	Xt	XY	X_t^2
1990	18547.9	1	18547.9	1
1991	21617.8	2	43235.6	4
1992	26638.1	3	79914.3	9
1993	34634.4	4	138537.6	16
1994	46759.4	5	233797	25
1995	58478.1	6	350868.6	36
1996	68593.8	7	480156.6	49
合 计	275269.5	28	1345058	140

例题1的计算结果1

1. $Y_t = a + bX_t$

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{7 * 1345058 - 28 * 275269.5}{7 * 140 - 28^2}$$

$$= 8713.557$$

$$a = \bar{Y} - b\bar{X} = \frac{275269.5}{7} - 8713.557 * \frac{28}{7} = 4469.986$$

长期趋势方程为 $Y_t = 4469.986 + 8713.557X_t$

(原点为1989年中期，单位为一年)

2. $b = 8713.557$，即 X_t 每变动一个单位长期趋势值的增减量.

3. $Y_{t(2000)} = 4469.986 + 8713.557 * (2000 - 1989)$

$= 100319.113$

确定原点的方法

1.原点在1989年年中,即1989年年中记为0

Xt=1,2,3,4,5,6,7

方程为Yt=4469.986+8713.557Xt

Yt(2000)=4469.986+8713.557*11=100319.113

2. 原点在1993年年中,即1993年年中记为0

Xt=-3,-2,-1,0,1,2,3

方程为Yt=39324.21+8713.557Xt

Yt(2000)= 39324.21 +8713.557*7=100319.109

确定原点的方法续

3.X_t=1990,1991,1992,1993,1994,1995,1996

方程为 Y_t= -17326794.89+8713.557X_t

Y_t(2000)= -17326794.89+8713.557*2000

=100319.11

季节变动的测定-----通过计算季节指数

季节指数的概念

是指用于表示具有典型季节性变动的现象年复一年地在每月(季)的变动方向和的幅度的百分数.

如某商场的销售额的 季节指数为 1.130% 2.93% 3.110% 4.67%

计算季节指数的方法----移动平均比率法（步骤）

1．用移动平均法

消除 S 和 I 的影响，从而得出 TC 值。

移动项数 K=4 或 K=12

1031. 1=203.7+239.7+288.9+298.8

2083.4=1031.1+1052.3

260. 43=2083.4/8

2．用比例法求得 SI （ SI=Y/TC ）

移动平均比率法（步骤续）

3．从SI中消除I，方法有：

- ☆求变通平均数
- 求中位数
- 求平均数

4．计算校正系数（400 或 1200/实际比例和）本例的校正系数为400/400.51=0.9987

5。季节指数=校正系数*变通平均数

计算季节指数的应用案例

	美国	迪斯尼	公司	季营业额		
时间	营业额	四季和	八季和	TC 值	SI值	季节指数
	（1）	（2）	（3）	(4)=(3)/8	(5)=(1)/(4)	（6）
83-1	203.7		—	—	—	83.52
		—				
83-2	239.7		—	—	—	92.26
		1031.1				
83-3	288.9		2083.4	260.43	110.93	112.56
		1052.3				
83-4	298.8		2109.2	263.65	113.33	111.66
		1056.9				
84-1	224.9		2139.5	267.44	84.09	83.52
		1082.6				
84-2	244.3		2180.0	272.50	89.65	92.26
		1097.4				
84-3	314.6		#REF!	#REF!	#REF!	112.56

上表续

		#REF!				
84-4	313.6		#REF!	#REF!	#REF!	111.66
.....	.....	.....	.....	.....	.....	.....
91-1	623.8		#REF!	#REF!	#REF!	83.52
		#REF!				
91-2	671.0		#REF!	#REF!	#REF!	92.26
		2864.6				
91-3	759.0		5767.8	720.98	105.27	112.56
		2903.2				
91-4	810.8		5909.5	738.69	109.76	111.66
		3006.3				
92-1	662.4		6144.1	768.01	86.25	83.52
		3137.8				
92-2	774.1		6461.0	807.63	95.85	92.26
		3323.2				
92-3	890.5		—	—	—	112.56
		—				
92-4	996.2		—	—	—	111.66

计算季节指数计算表

		季	节	指	数	计	算	表				
季节	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	变通平均数	季节指数
1	-	84.09	80.35	<77.11>	83.8	80.49	86.2	85.49	85.01	<86.25>	83.63	83.52
2	-	89.65	88.43	<96.75>	92.23	<87.78>	92.2	94.96	93.35	95.85	92.38	92.26
3	110.93	114.27	115.41	110.83	<115.7>	114	109.9	113.59	<105.27>	-	112.7	112.56
4	113.33	112.21	<114.9>	111.39	112.36	110.48	112.1	110.75	<109.78>	-	111.8	111.66
合计	-	-	-	-	-	-	-	-	-	-	400.51	400

季节指数的应用

- 掌握时间数列的季节变动规律

  从而提高预测的精度

- 可以从时间数列中消除季节变动的影响，从而提高分析问题的科学性。

第七章 指数

本章应掌握的主要内容

1.指数的概念

2. 指数的种类

3.编制总指数的方法

4.指数的应用

指数的概念

1.广义:凡能说明现象变动的相对数

2.狭义:指数是指不能直接相加现象在不同时期比较 的综合相对数。

如：全国零售物价总指数、全国货物出口量指数。指数一般用百分数的分子表示。

广义的指数

- 相对数
  - 静态相对数
    - 结构相对数
    - 比例相对数
    - 类比相对数
    - 强度相对数
  - 动态相对数（发展速度）
  - 评价相对数

指数的种类

1.按照说明对象的范围的不同,指数分为

- 个体指数(反映单一事物的变动情况的相对数.
- 总指数(说明多种事物综合变动的相对数)

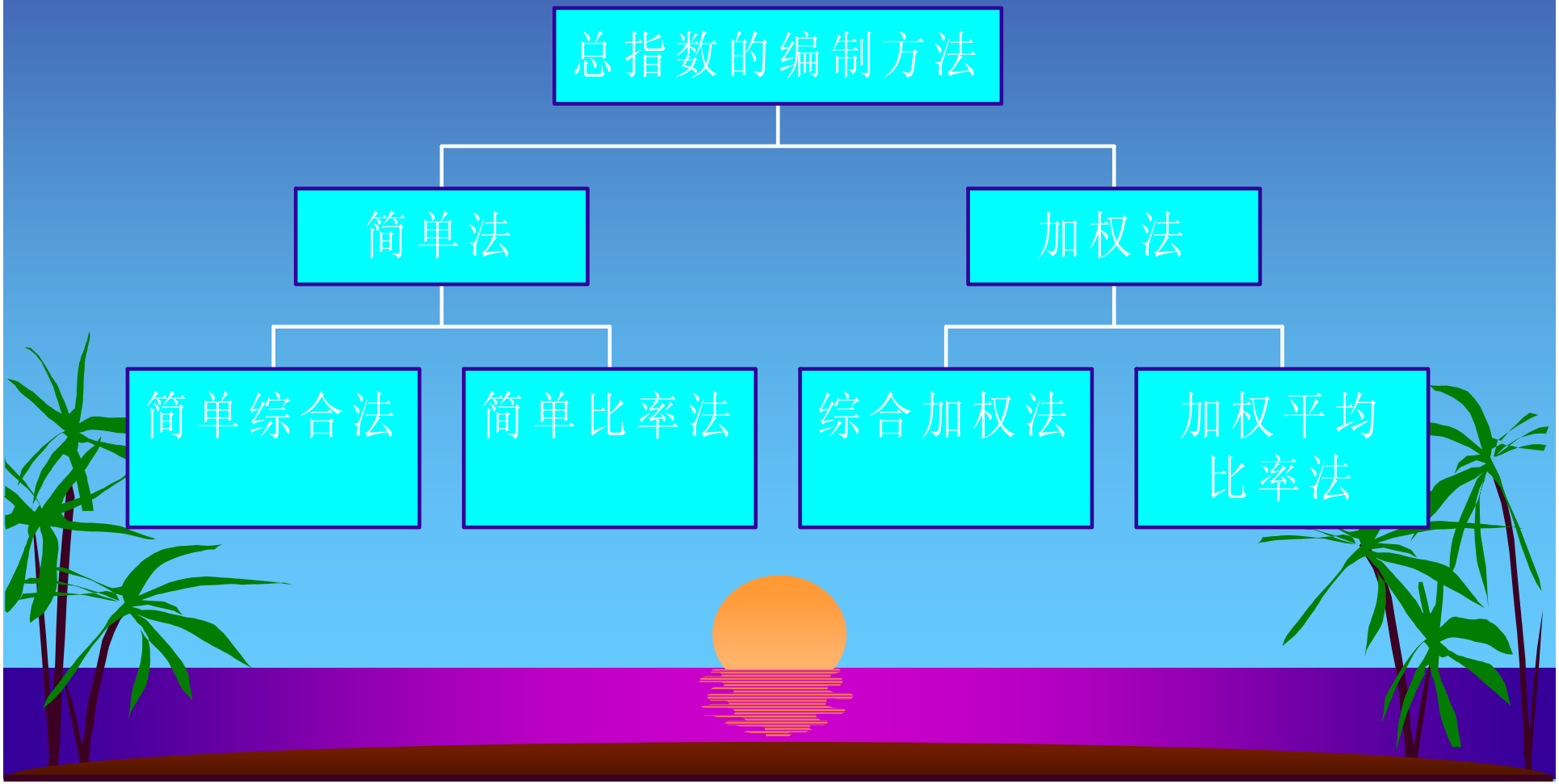
2.指数按所表明现象的数量特征的不同分

- 绝对因子指数:说明总体数量变化的指数
- 相对因子指数:说明总体相对变化的指数

3. 在指数数列中,按所采用的基期不同分为

- 定基指数(基期固定在某时期)
- 环比指数(以上期为基期)

总指数的编制方法



简单综合法

- 定义:将不能直接加总的所研究现象直接加总进行对比.以物价指数为例:
- 公式:

$$P = \frac{\sum_{i=1}^{k} p_{ni}}{\sum_{i=1}^{k} p_{0i}} = \frac{\sum p_{ni}}{\sum p_{0i}} = \frac{\sum p_n}{\sum p_0}$$

总指数计算方法的举例

某商店商品销售情况

商品	计量	销售量		价格（元）	
名称	单位	基期	报告期	基期	报告期
		q_0	q_n	p_0	p_n
A	件	200	250	4.2	4
B	米	750	800	3.6	3
C	台	50	46	9.6	12

简单综合法计算结果

$$P=\frac{\sum p_n}{\sum p_0}=\frac{4+3+12}{4.2+3.6+9.6}=\frac{19}{17.4}=109.20\%$$

缺点：1.把不同商品的价格相加是不科学的.

2.没有加权,既每种商品价格的变动作用相同.

3.价格总指数的大小受计量单位的影响.

如把B产品的计量单位改为元/千米,则有:

$$P=\frac{4+3000+12}{4.2+3600+9.6}=\frac{3016}{3613.8}=83.46\%$$

简单比率平均指数

1.定义:总指数为个体指数的简单算术平均数

2.公式:

$$P = \frac{\sum_{i=1}^{N} \frac{p_{ni}}{p_{0i}}}{N} = \frac{\sum \frac{p_{ni}}{p_{0i}}}{N} = \frac{\sum \frac{p_n}{p_0}}{N}$$

3.优点:不受计量单位的影响.

4.缺点:未加权,既每种商品价格的变动作用相同.

简单比率平均指数计算结果

$$P = \frac{\sum \frac{p_n}{p_0}}{N} = \frac{\frac{4}{4.2} + \frac{3}{3.6} + \frac{12}{9.6}}{3}$$

$$= \frac{0.9524 + 0.8333 + 1.25}{3}$$

$$= \frac{3.0357}{3} = 1.0119 = 101.19\%$$

综合加权指数

1.定义:由两个总量指标对比形成的指数.凡是一个总量指标可以分解为两个或两个以上的因素指标时,将其中一个或一个以上的因素指标固定下来,仅观察剩余的一个因素指标的变动程度.

德国经济学家1864

2.公式:

拉氏公式

$$\begin{cases} P_L\text{（Laspeyres 公式）} = \dfrac{\sum p_n q_0}{\sum p_0 q_0} \\ Q_L = \dfrac{\sum q_n p_o}{\sum q_0 p_o} \end{cases}$$

Q为数量指数

综合加权物价指数(拉氏)

$$P_L = \frac{\sum p_n q_0}{\sum p_0 q_0} = \frac{200\times 4 + 750\times 3 + 50\times 12}{200\times 4.2 + 750\times 3.6 + 50\times 9.6} = \frac{3650}{4020}$$

$$= 0.9080 = 90.8\%$$

$$\sum p_n q_0 - \sum p_0 q_0 = -370$$

结果的解释:1.三种商品综合起来报告期物价比基期的?物价下降了.

2. 三种商品综合起来报告期物价比基期的物价下降了9.2%

3.由于物价下跌使报告期的销售额比基期减少了370元.

综合加权数量指数(拉氏)

$$Q_L = \frac{\sum q_n p_0}{\sum q_0 p_0} = \frac{250*4.2+800*3.6+46*9.6}{200*4.2+750*3.6+50*9.6}$$

$$= \frac{4371.6}{4020} = 1.0875 = 108.75\%$$

$$\sum q_n p_0 - \sum q_0 p_0 = 351.6$$

结果的解释:1.三种商品综合起来报告期销售量比基期的?销售量上涨了.

2. 三种商品综合起来报告期销售量比基期的销售量上涨了8.75%

3.由于销售量的上涨使报告期的销售额比基期增加了351.6元.

综合加权指数(派氏公式)

德国经济学家1874

派氏公式

$$P_{p(H.Passche)} = \frac{\sum p_n q_n}{\sum p_0 q_n} = \frac{3952}{4371.6} = 90.4\%$$

$$\sum p_n q_n - \sum p_0 q_n = -419.6$$

$$Q_p = \frac{\sum q_n p_n}{\sum q_0 p_n} = \frac{3952}{3650} = 108.27\%$$

$$\sum q_n p_n - \sum q_0 p_n = 302$$

拉氏公式和派氏公式的比较

拉氏

- 把同度量因素固定在基期
- 有利于不同时期的指数进行比较
- 取得资料比较容易
- 如基期离报告期太远,则算出的指数现实意义差.

派氏

- 把同度量因素固定在报告期
- 不利于不同时期的指数进行比较
- 需要搜集的资料较多
- 算出的指数更具有现实意义.

物值指数(这里指销售额指数)的计算

$$V=\frac{\sum p_n q_n}{\sum p_0 q_0}=\frac{3952}{4020}=98.31\%$$

$$\sum p_n q_n-\sum p_0 q_0=3952-4020=-68$$

结果解释:三种商品的销售额报告期比基期

1.从相对数看:下降了1.69%

2.从绝对数看:减少了68元

物价指数,物量指数及物值指数的关系

从相对数看:

$$V = Q_L \times P_P = \frac{\sum q_n p_0}{\sum q_0 p_0} \times \frac{\sum p_n q_n}{\sum p_0 q_n}$$

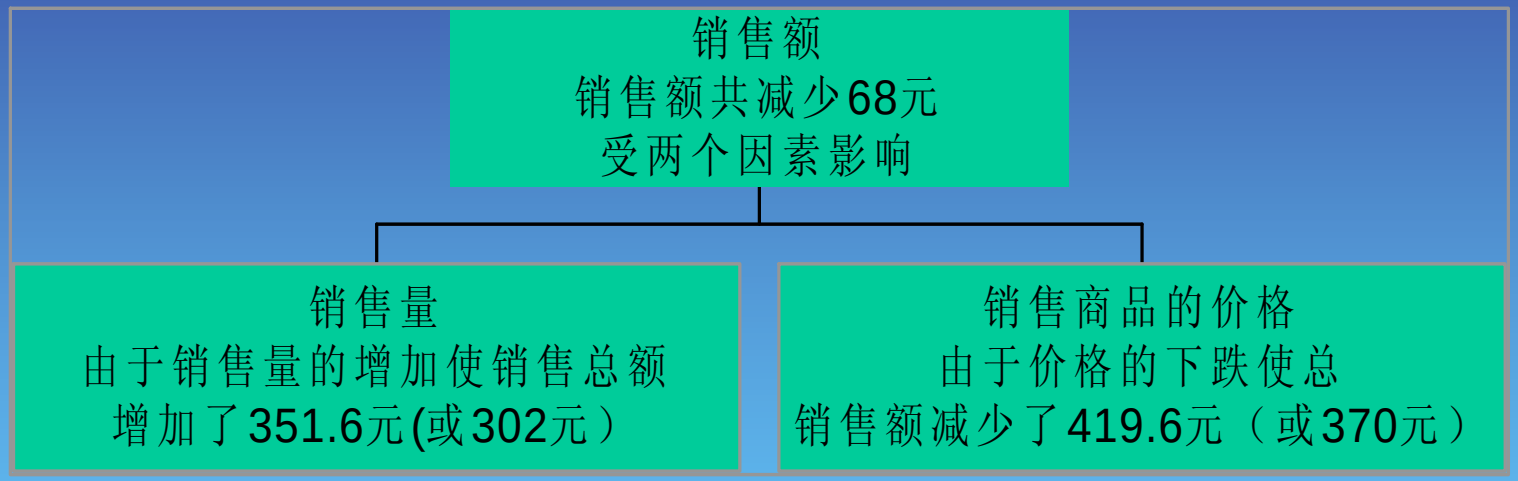
$$V = Q_p \times P_L = \frac{\sum p_n q_n}{\sum p_n q_0} \times \frac{\sum p_n q_0}{\sum q_0 p_0}$$

从绝对数看:

$$\left(\sum p_n q_n - \sum p_0 q_0\right) = \left(\sum q_n p_0 - \sum q_0 p_0\right) + \left(\sum p_n q_n - \sum p_0 q_n\right)$$

或$$= \left(\sum p_n q_n - \sum p_n q_0\right) + \left(\sum p_n q_0 - \sum p_0 q_0\right)$$

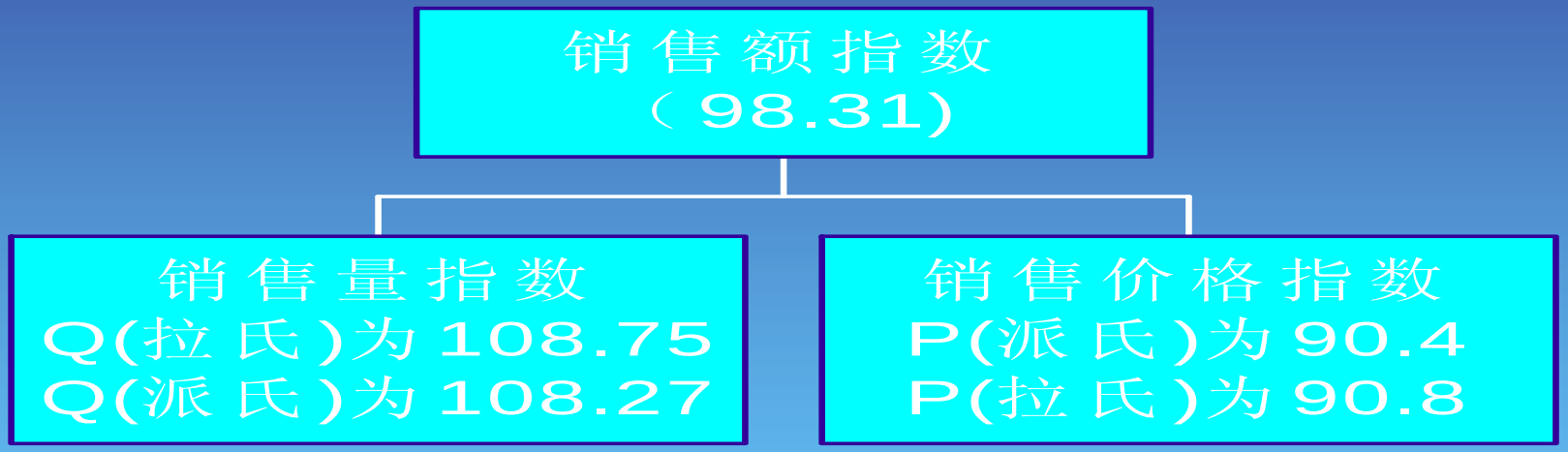
销售额指数、销售量指数和价格指数的关系（从绝对数看）



351.6+(-419.6)=-68

302+(-370)=-68

销售额指数、销售量指数和价格指数的关系（从相对数看）



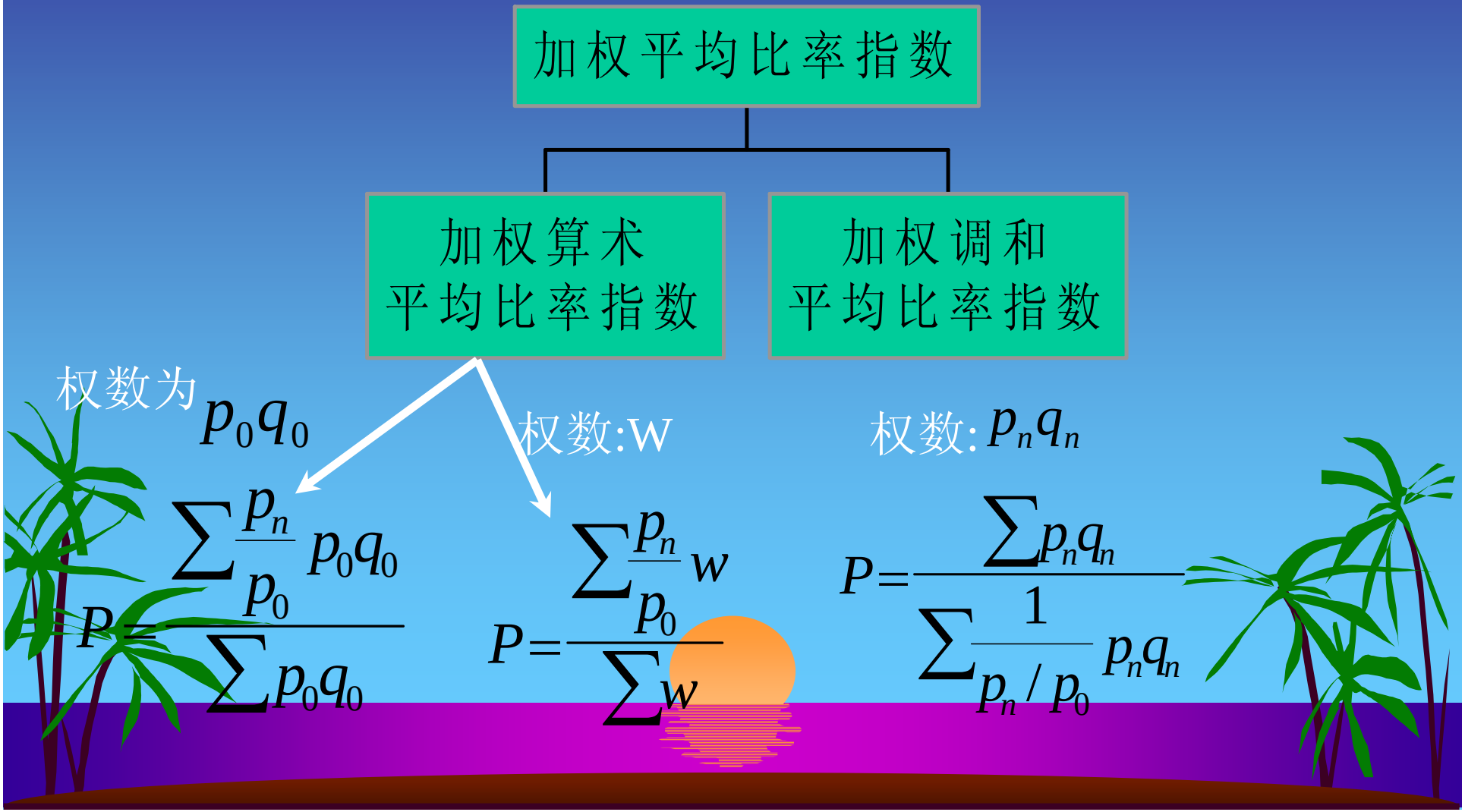
108.75%*90.40%=98.31%

108.27%*90.8%=98.31%

加权平均比率法

- 1.定义:　　它是根据非全面调查资料,先计算个体指数,然后给出权数,对个体指数进行
  加权平均.
- 2.基本形式:

加权平均比率指数的基本形式



指数体系

- 定义：如有指数A、B、C、D、E.?它们之间存在如下关系，即：A=B*C*D*E,则称A、B、C、D、E构成一个指数体系.
- 举例: 因为:$V=Q_p*P_L$ 或 $V=Q_L*P_p$ 所以 V、Q、P之间构成一个指数体系

指数体系的作用

- 根据指数体系各指数之间的关系，可以据已知的两个或两个以上的指数去推算另一个指数。如：据物值指数和数量指数可以算出物价指数。
- 可以分析各因素变动对总变动的影响程度和影响的绝对值。即可以进行因素分析。如上例中，分析销售价格和销售量对销售额的影响程度和绝对值的影响。

指数的应用

- 股票价格指数
- 零售物价总指数
- 工业生产指数
- 消费品价格指数
- 货币购买力指数
- 随价调整

中国零售物价总指数的编制

1.把所有的商品分为 14大类，大类中有分为

- 中类、小类、商品。
- 如食品大类中分为10个中类

在粮食中类中分为2个小类（细粮和粗粮）

在细粮小类中分为：面粉、大米、江米、挂面。每一小类中再选代表品。

共选352种至397种之间。

2.计算各个代表品的个体指数，如大米的个

中国零售物价总指数的编制续1

体指数为：Kp=3.0/2.4=125%

3.把个体指数乘上相应的权数后相加,计算其算术平均数得小类指数,细粮小类指数为:

$$\overline{K}_p = \frac{\sum K_p p_0 q_0}{\sum p_0 q_0} = \sum K_p W$$

$$= 1.25 * 0.9 + 1.3333 * 0.1 = 125.83\%$$

中国零售物价总指数的编制续2

4.把小类指数乘上相应的权数后相加,计算其算术平均数得中类指数.如粮食中类指数为:

$$\overline{K_p} = \Sigma K_p W = 1.2583 * 0.98 + 1.4060 * 0.02 = 126.12\%$$

5.把中类指数乘上相应的权数后相加,计算其算术平均数得大类指数.如食品大类指数为:

中国零售物价总指数的编制续3

$$\overline{K}_p = \sum K_p W$$

$$= 1.2612*0.13 + 1.5026*0.03 + 1.5830*0.26$$

$$+\Lambda + 1.3028*0.18 = 141.93\%$$

6.把大类指数乘上相应的权数后相加,计算其算术平均数得总指数.

$$\overline{K}_p = \sum K_p W$$

$$= 1.4193*0.27 + 1.0825*0.12 + 1.1548*0.09$$

$$+\Lambda + 0.9621*0.01 = 123.25\%$$

货币购买力指数和通货膨胀率

- 货币购买力指数为居民消费价格指数的倒数.

$$\text{通货膨胀率} = \frac{\text{报告期居民消费价格指数} - \text{基期居民消费价格指数}}{\text{基期居民消费价格指数}} \times 100\%$$

编制指数应注意的问题

- 抽选样本
- 选择基期(正常年份)
- 选择权数
- 选择公式

编制总指数公式归纳

总指数

简单法

综合加权法

加权平均比率法

$$P=\frac{\sum p_n}{\sum p_0}$$

$$\frac{\sum \frac{p_n}{p_0}}{N}$$

$$P=\sum p_n q_0 / \sum p_0 q_0$$

$$P=\sum p_n q_n / \sum p_0 q_n$$

$$P=\sum p_n q_a / \sum p_0 q_a$$

$$P=\sum p_n (q_0+q_n) / \sum p_0 (q_0+q_n)$$

$$P=\frac{\sum K_p p_0 q_0}{\sum p_0 q_0}$$

$$P=\sum K_p W$$

$$P=\frac{\sum p_n q_n}{\sum \frac{1}{K_p} p_n q_n}$$

本章重点

1.指数的概念

2. 指数的种类

3.编制总指数的方法

4.指数的应用

全国及京津沪三直辖市居民消费价格分类指数（1996年）

（上年=100）

指数类别	全　国	北　京	天　津	上　海
总指数	**108.3**	111.6	109.0	109.2
商品	**107.6**	107.7	106.7	109.7
粮食	**106.5**	112.2	106.1	109.4
油脂	**92.1**	90.8	90.4	95.6
肉禽及其制品	**104.5**	101.5	100.0	107.4
蛋	**116.5**	118.1	119.7	124.5
水产品	**106.0**	104.5	105.3	111.8
菜	**119.1**	113.2	109.5	122.8
酒和饮料	**106.1**	106.9	104.3	105.5
干鲜瓜果	**104.5**	105.6	109.5	107.2
饮食业	**109.4**	106.2	110.2	111.0
衣着	**107.4**	119.1	104.8	108.8
家用设备及用品	**103.8**	104.6	103.4	99.9
医疗保健用品	**109.3**	110.1	106.6	107.6
交通和通讯工具	**98.8**	102.1	100.8	98.6
娱乐教育文化用品	**110.4**	110.1	107.8	105.3
居住	**111.4**	129.2	137.1	109.7
服务项目	**116.0**	120.0	117.4	118.9

资料来源：《中国统计年鉴·1997》。

第八章　概率

- 本章重点：几个基本概念，
  - 随机现象：即事物发展的结果事先不能确定的现象
  - 随机试验：对随机现象进行试验或观察
  - 随机事件：随机现象出现的不同结果
  - 概率：衡量随机事件结果发生可能性大小的数值

$$0 \leq P(x) \leq 1$$

$p(x) = 0$　不可能事件

$P(x) = 1$　必然事件

种子发芽相对频数的稳定趋势

样本序号	样本容量	发芽数	相对频数
1	2	2	1.000
2	5	4	0.800
3	10	9	0.900
4	70	60	0.857
5	130	116	0.892
6	310	282	0.910
7	700	639	0.913
8	1500	1364	0.909
9	2000	1806	0.903
10	3000	2715	0.905

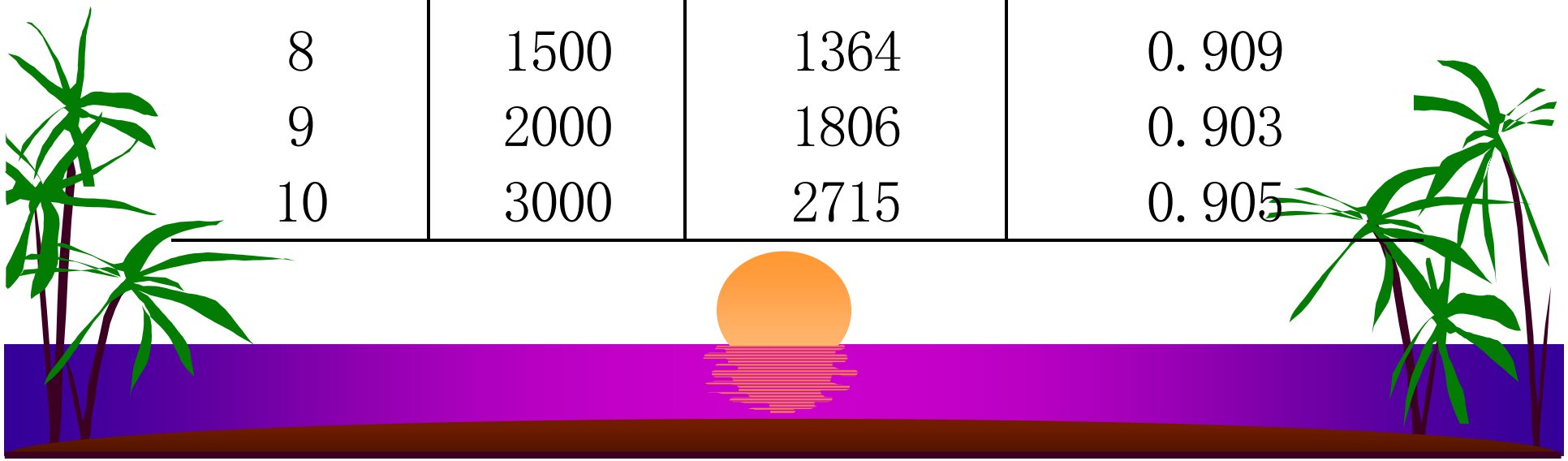
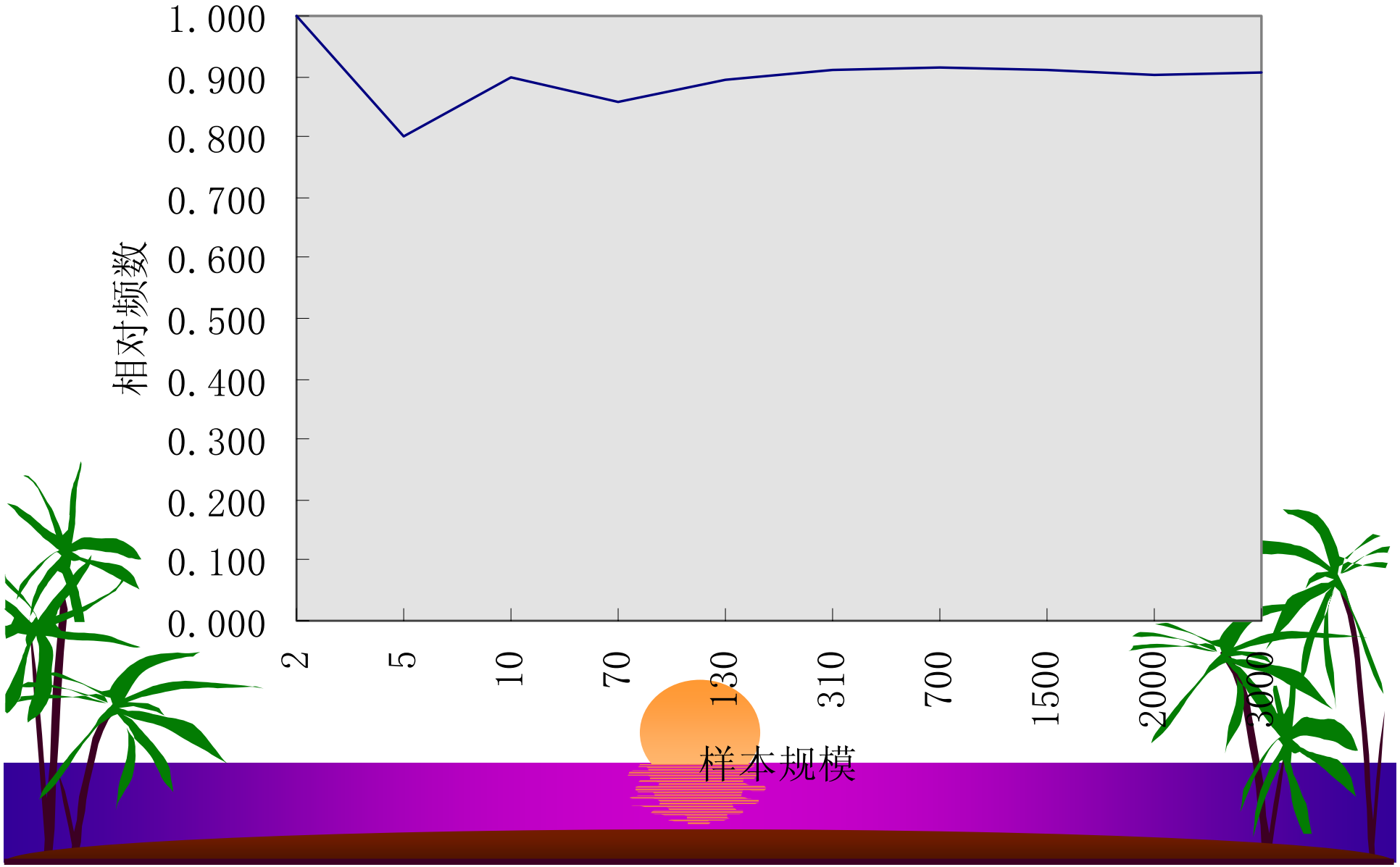
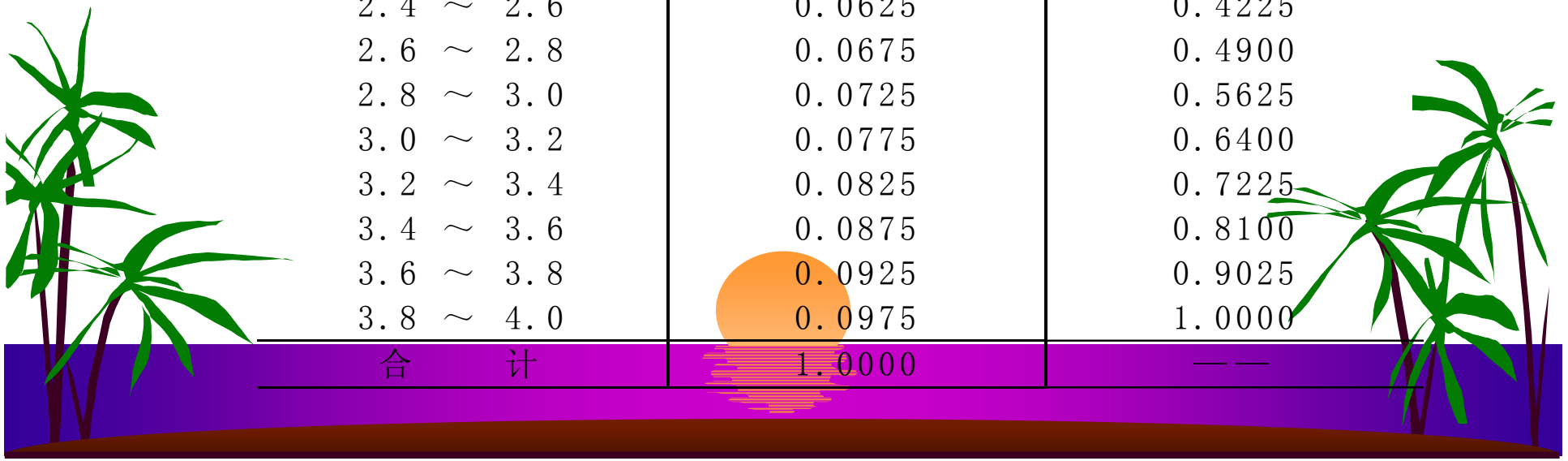


图4.1　种子发芽相对频数的趋势

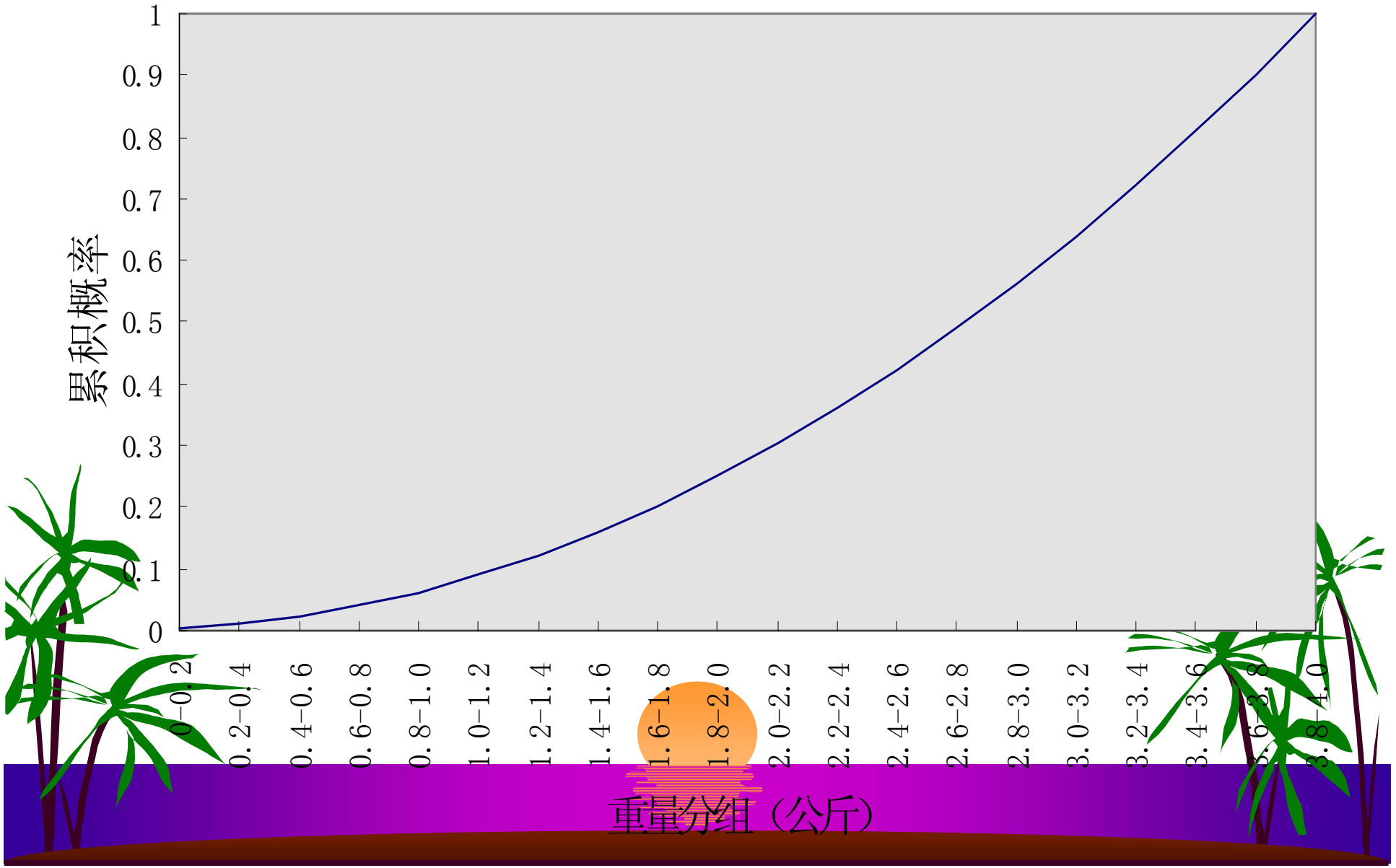


邮包重量（连续型变量）的概率分布

邮包重量分组（公斤） $x_1 \leqslant X \leqslant x_2$	概　　率 $P(x_1 \leqslant X \leqslant x_2)$	累计概率 $P(X \leqslant x_2)$
0 ～ 0.2	0.0025	0.0025
0.2 ～ 0.4	0.0075	0.0100
0.4 ～ 0.6	0.0125	0.0225
0.6 ～ 0.8	0.0175	0.0400
0.8 ～ 1.0	0.0225	0.0625
1.0 ～ 1.2	0.0275	0.0900
1.2 ～ 1.4	0.0325	0.1225
1.4 ～ 1.6	0.0375	0.1600
1.6 ～ 1.8	0.0425	0.2025
1.8 ～ 2.0	0.0475	0.2500
2.0 ～ 2.2	0.0525	0.3025
2.2 ～ 2.4	0.0575	0.3600
2.4 ～ 2.6	0.0625	0.4225
2.6 ～ 2.8	0.0675	0.4900
2.8 ～ 3.0	0.0725	0.5625
3.0 ～ 3.2	0.0775	0.6400
3.2 ～ 3.4	0.0825	0.7225
3.4 ～ 3.6	0.0875	0.8100
3.6 ～ 3.8	0.0925	0.9025
3.8 ～ 4.0	0.0975	1.0000
合　　计	1.0000	——



邮包重量累积概率


累积概率
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
1.0-1.2
1.2-1.4
1.4-1.6
1.6-1.8
1.8-2.0
2.0-2.2
2.2-2.4
2.4-2.6
2.6-2.8
2.8-3.0
3.0-3.2
3.2-3.4
3.4-3.6
3.6-3.8
3.8-4.0
重量分组（公斤）

第九章　概率分布

- 随机变量：一个表述随机事件结果取值的变量
- 概率分布：与随机变量取值联系的一系列概率值
- 概率分布的表述：
  - 表格
  - 图形
  - 公式
- 期望值：随机变量的平均数

$$E(X) = \sum x_i p(x_i)$$

方差（刻画 随机变量的离散程度）

- 公式 $\sigma^2 = \sum [X - E(X)]^2 P(x_i)$

X：抛一枚硬币出现正面的次数

1. P（X）的概率分布

X	0	1
P（x)	0.5	0.5

2.E(X)=0*0.5+1*0.5=0.5

3.方差=$(0-0.5)^2 \times 0.5 + (1-0.5)^2 \times 0.5 = 0.25$

离散型随机变量的典型概率分布----二项分布

- 二项试验以及其特点
- 二项变量
- 二项分布的概率密度函数
- 二项分布的应用
- 二项分布的分布形态

二项试验和二项变量

- 二项试验(贝努里试验)的特点:
  - 每次试验只有两种可能结果,即成功或失败
  - 每一次试验成功或失败的概率已知并不变

	成功	失败
概率	π	$1-\pi$

  - 每次试验是相互独立的
- 二项变量:在n次 二项试验中成功的次数

二项分布的概率密度函数

X 服从二项分布,成功经验的概率为 π

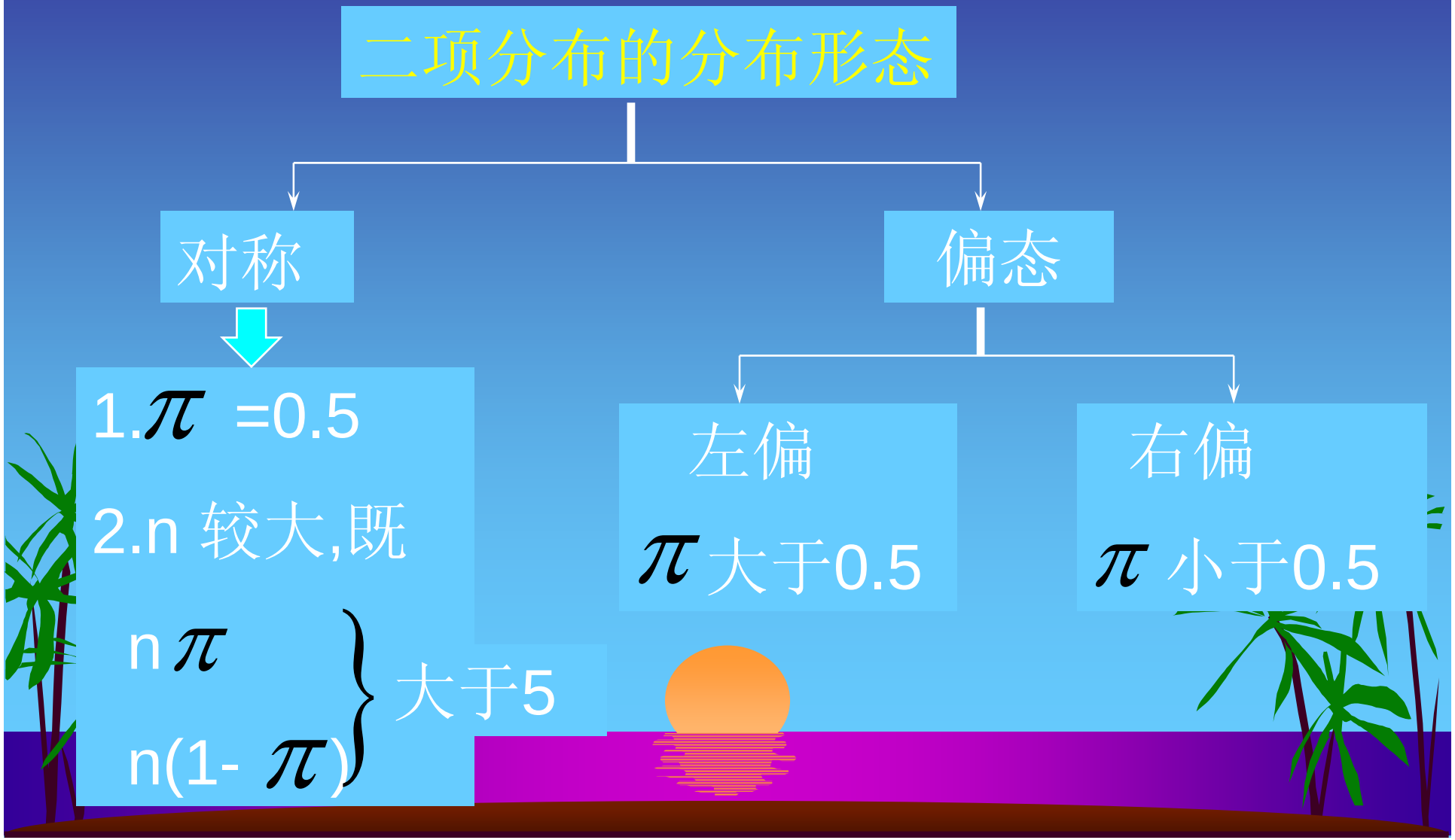
则计为 X～ (n π)

$$p(X = x) = C_n^x \pi^x (1-\pi)^{n-x}$$

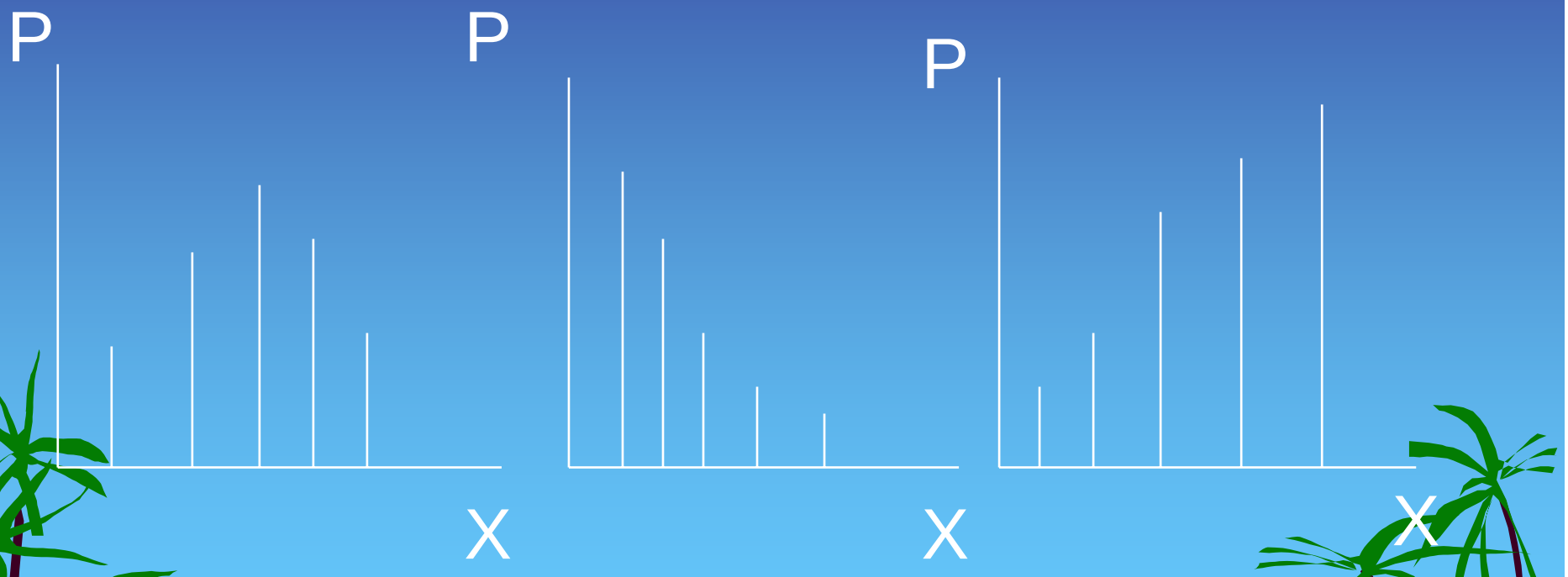
大写的X表示随机变量

小写的x表示随机变量的具体取值

二项分布的分布形态



对称 左偏 右偏的图示

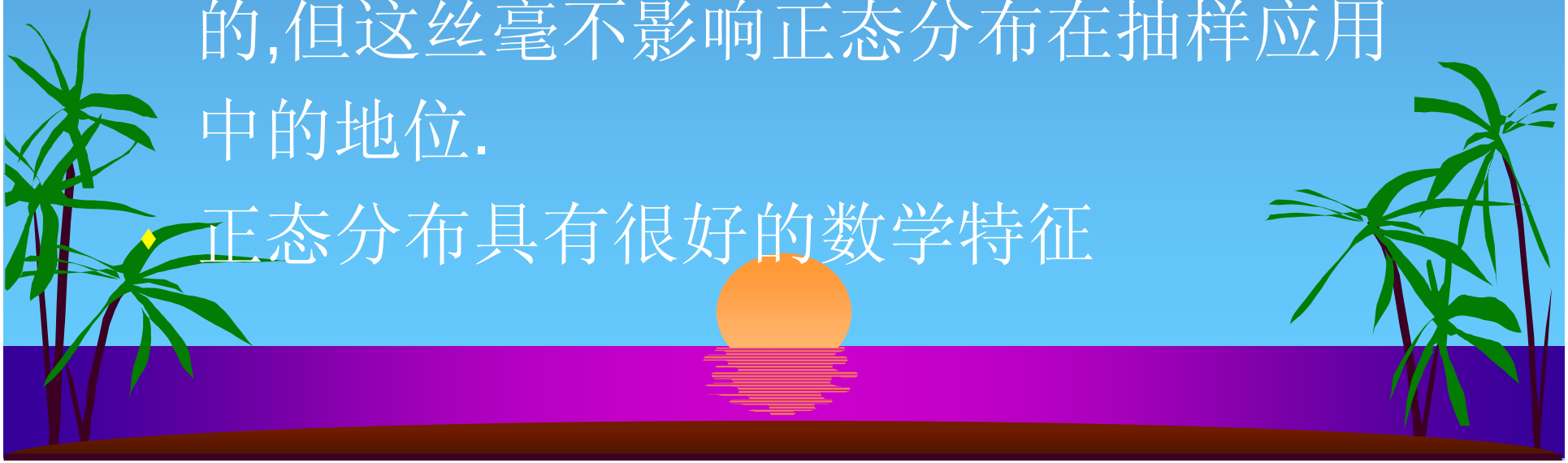


连续型随机变量的典型概率分布----正态分布

- 正态分布在统计学中的地位
- 正态分布的特点
- 正态分布的概率密度函数
- 标准 正态分布
- 标准 正态分布表
- 正态分布的的应用

正态分布在统计学中的地位
是统计和抽样的理论基础

- 客观世界中许多随机现象都服从或近似服从正态分布.
- .尽管经济管理活动中的有些变量是偏斜的,但这丝毫不影响正态分布在抽样应用中的地位.
- 正态分布具有很好的数学特征



正态分布的特点

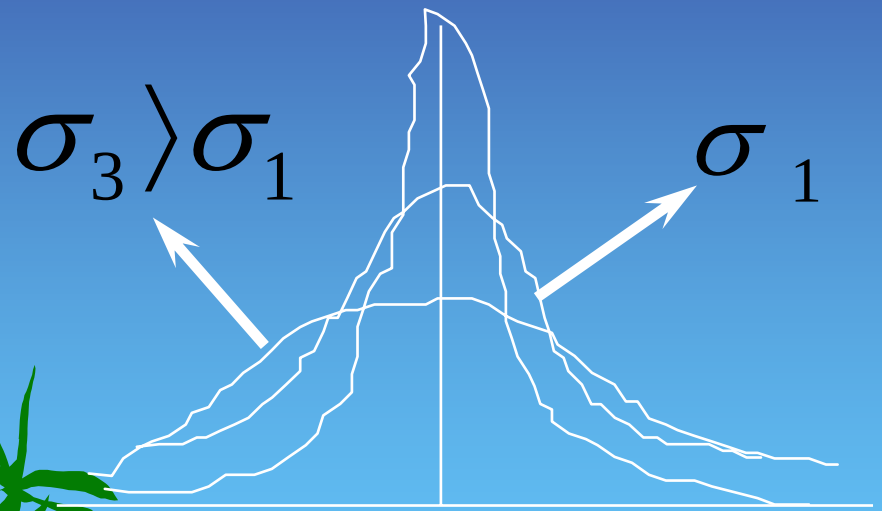
- 正态随机变量X 的取值域为整个X轴,X轴为正态曲线的渐近线.
- 正态曲线与X轴围成的面积为1.
- 正态曲线的分布中心为 μ
- 正态曲线的形状由随机变量的标准差决定. 标准差越大, 正态曲线矮胖; 标准差越小, 正态曲线高瘦.

正态分布图形的特点

$\sigma_2 < \sigma_1$

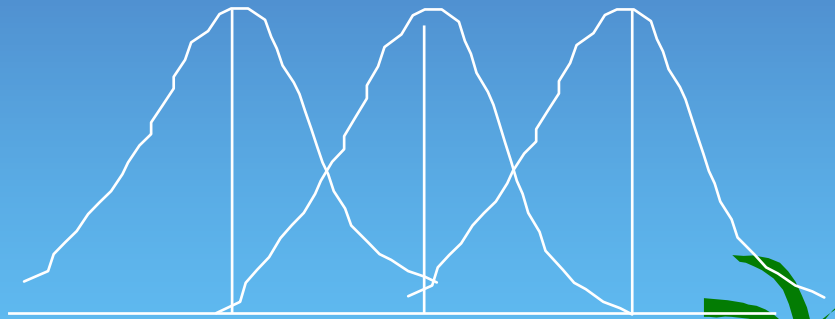
$\sigma_3 > \sigma_1$

σ_1



μ_1

$\sigma_1 = \sigma_2 = \sigma_3$



μ_1 μ_2 μ_3

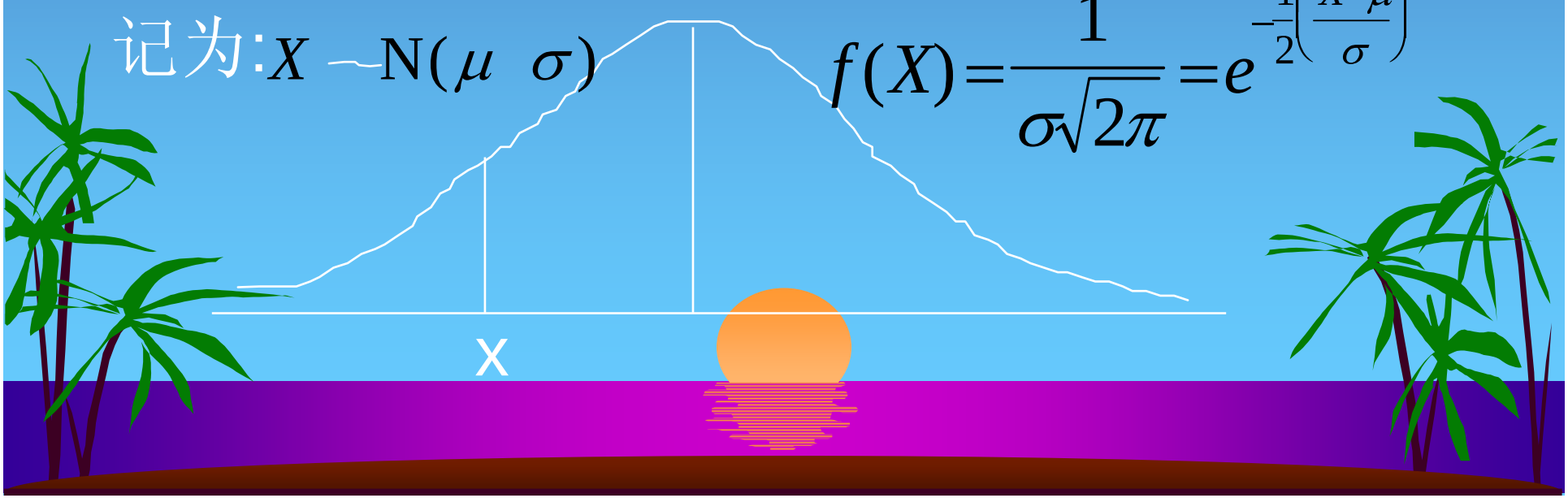
正态分布的概率密度函数

$$P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} = e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} dx$$

X服从正态分布

记为：$X \sim N(\mu\ \sigma)$

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} = e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$



标准 正态分布

$$P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

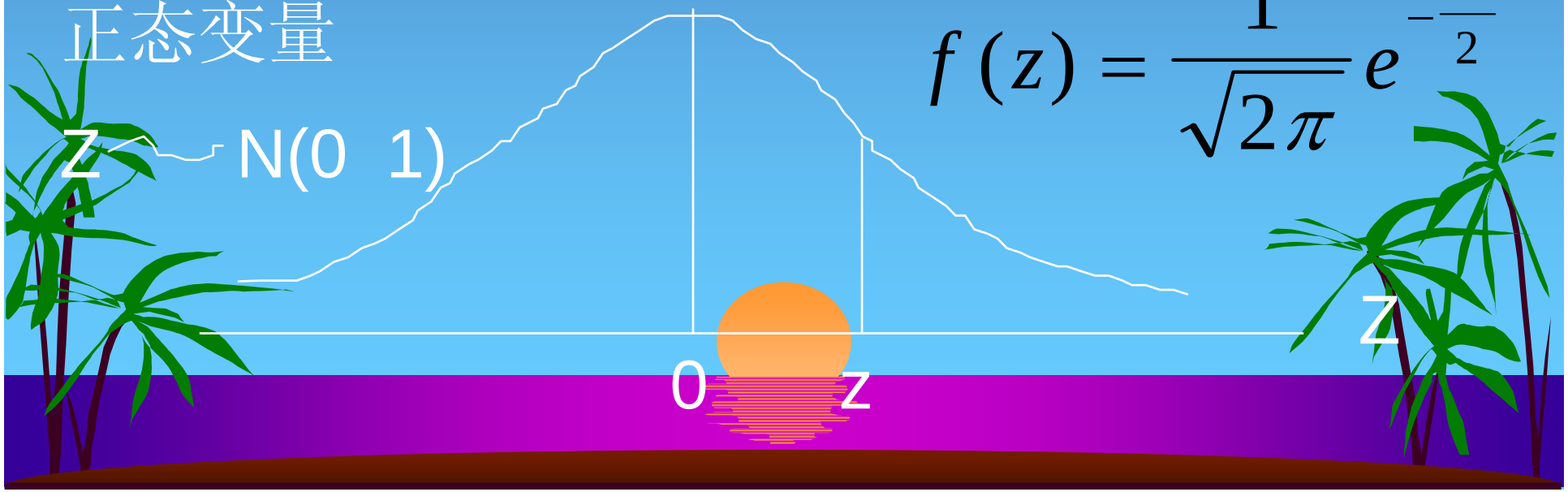
Z为标准

正态变量

$Z \sim N(0\ \ 1)$

$\sigma = 1$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$



正态分布$N(\mu\ \sigma)$和标准正态N(0 1)的关系

$X \sim N(\mu\ \sigma)$ 则 $\left(\frac{X-\mu}{\sigma}=z\right) \sim N(0\ 1)$

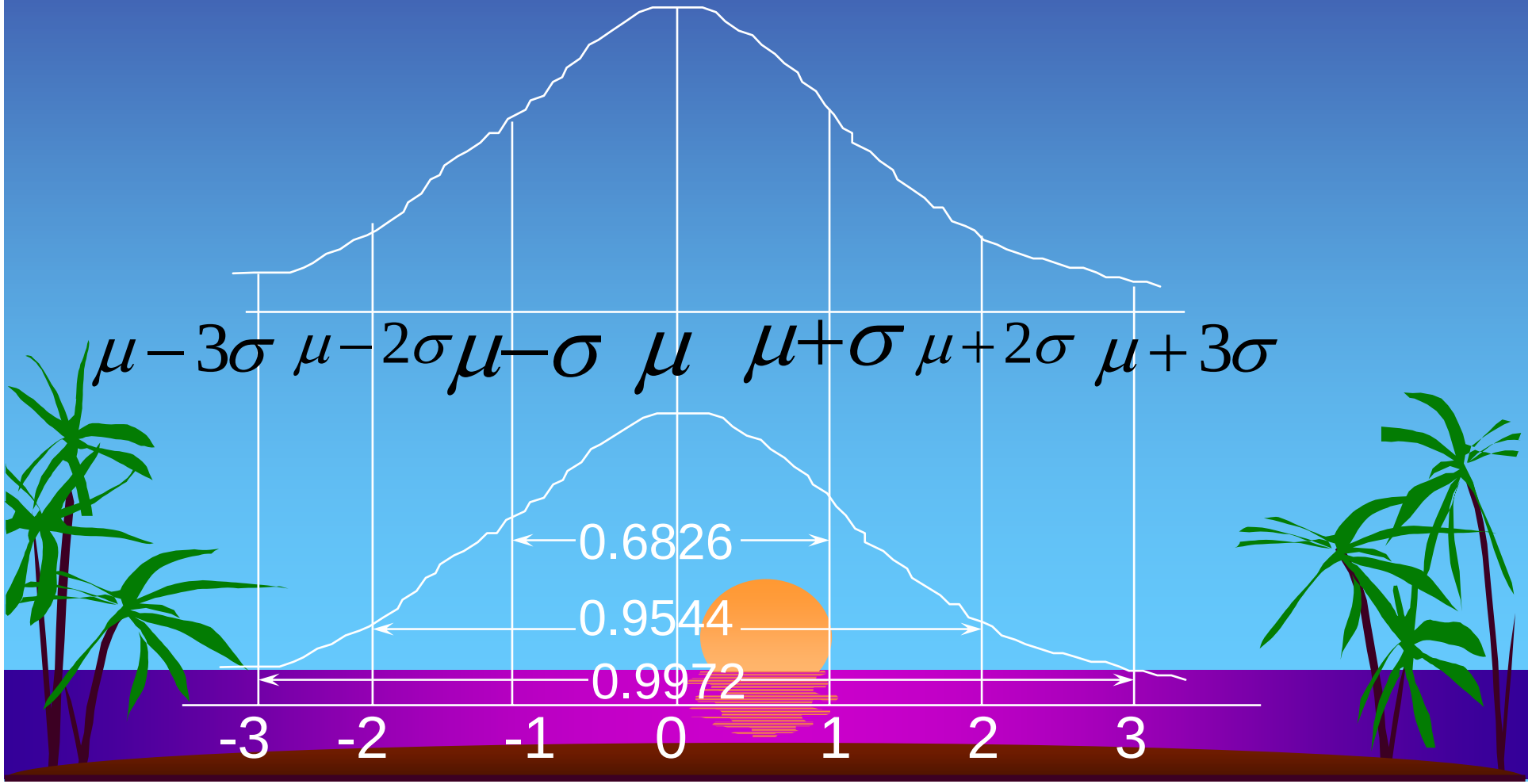
$x=\mu+\sigma$ 时 $Z=\frac{X-\mu}{\sigma}=\frac{\mu+\sigma-\mu}{\sigma}=1$

$x=\mu-\sigma$ 时 $Z=\frac{X-\mu}{\sigma}=\frac{\mu-\sigma-\mu}{\sigma}=-1$

$x=\mu+2\sigma$ 时 $Z=\frac{X-\mu}{\sigma}=\frac{\mu+2\sigma-\mu}{\sigma}=2$

$x=\mu-2\sigma$ 时 $Z=\frac{X-\mu}{\sigma}=\frac{\mu-2\sigma-\mu}{\sigma}=-2$

正态分布$N(\mu\ \ \sigma)$和标准正态N(0 1)的关系



正态分布的的应用

某企业生产目光灯,目光灯的使用寿命服从正态分布,其均值为1000小时,标准差为200小时,试求使用寿命在下列范围内的概率.

1.使用寿命在800--1200小时之间;

2. 使用寿命在600--1400小时之间;

3. 使用寿命在400--1600小时之间;

4. 使用寿命小于920小时;

上题计算结果

$$1.\ p(800 \le X \le 1200) = p\left(\frac{800-1000}{200} \le \frac{x-\mu}{\sigma} \le \frac{1200-1000}{200}\right)$$

$$= p(-1 \le z \le 1) = 0.6826$$

$$2.\ p(600 \le X \le 1400) = p\left(\frac{600-1000}{200} \le z \le \frac{1400-1000}{200}\right)$$

$$= p(-2 \le z \le 2) = 0.9544$$

$$3.\ p(400 \le X \le 1600) = p(-3 \le z \le 3) = 0.9972$$

$$4.\ p(X \le 920) = p\left(Z \le \frac{920-1000}{200}\right) = p(Z \le -0.4)$$

$$= 0.3446$$

第十章 抽样和抽样分布

本章重点：

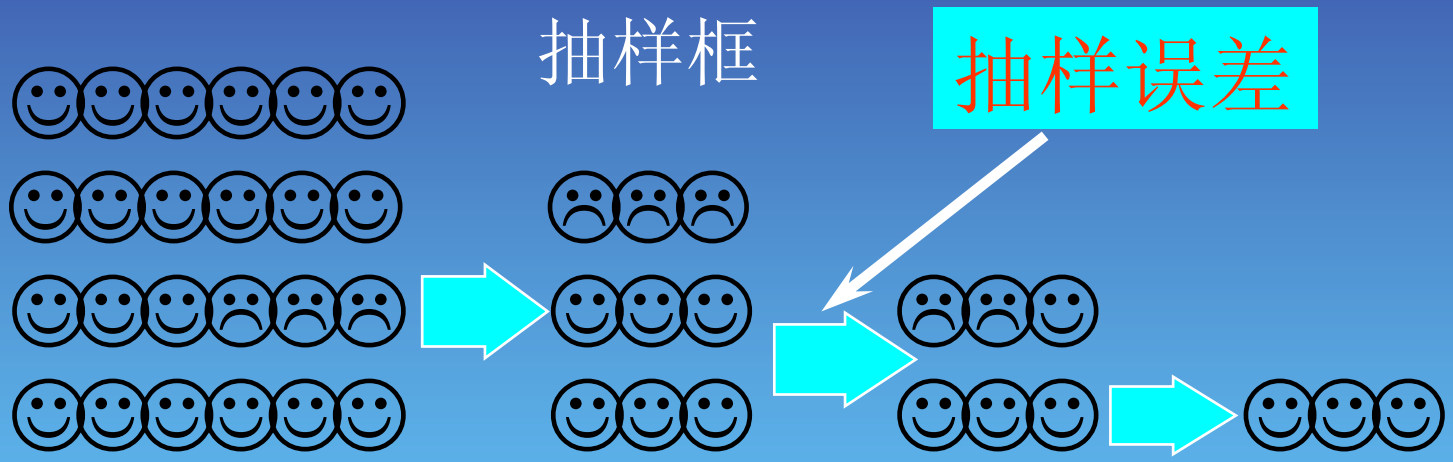
- 第一节：抽样的基本概念
  - 抽样调查的优越性和必要性
  - 抽样基本设计
- 第二节：抽样分布的基本原理
  - 总体参数和样本统计量
  - 抽样分布定理
    - 正态分布再生定理
    - 中心极限定理
  - 总体标准差不明确时样本平均数的抽样分布
  - 样本比率的抽样分布

第一节：抽样的基本概念

- 抽样调查的必要性和优越性
  - 经济性，时效性，必需性
- 统计误差

抽样调查的误差来源

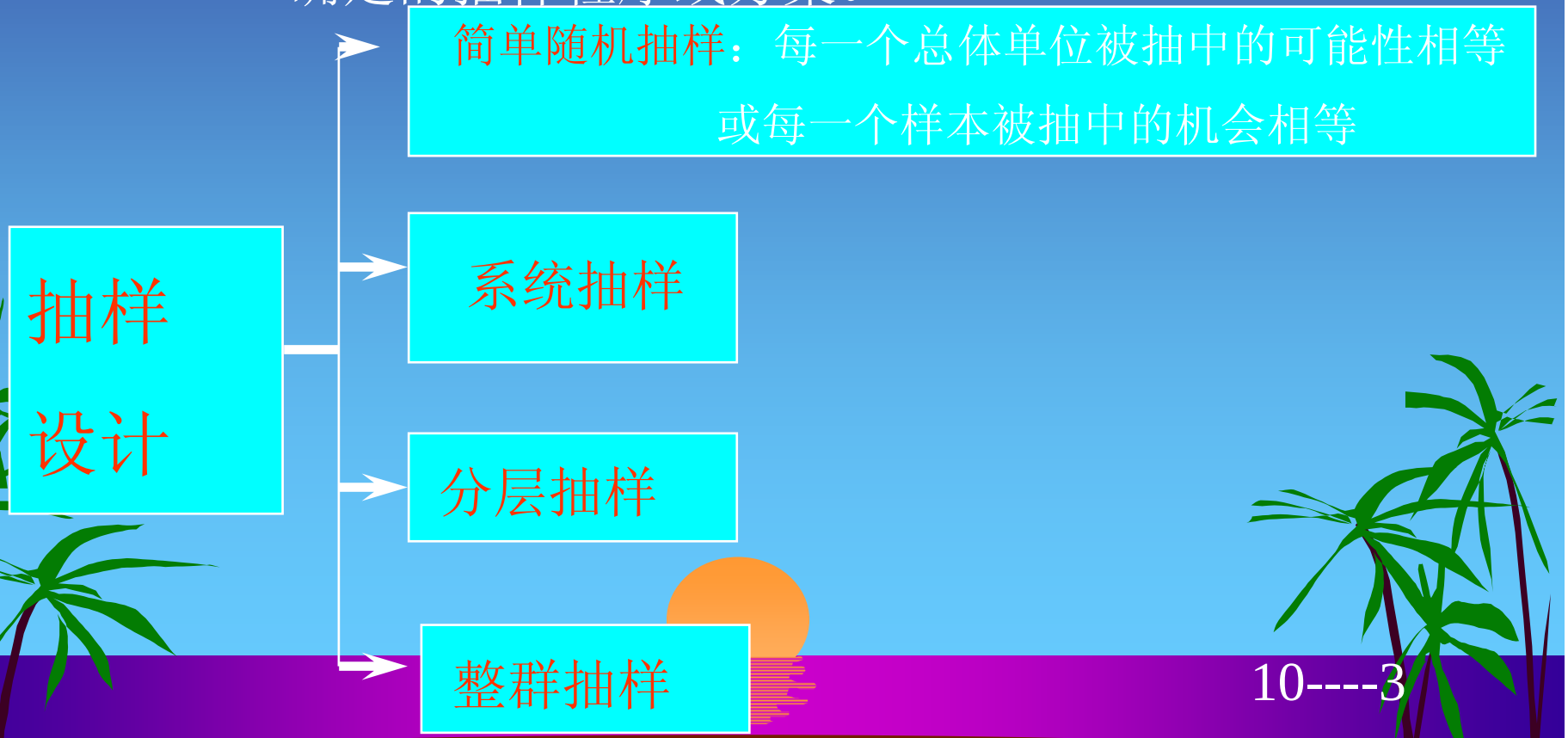
抽样调查的目的：据实际样本的信息推断目标总体



目标总体：	作业总体：	计划	实际样本
即我们所要研究的总体	抽取样本的总体	样本	实际被调查的样本

第一节：基本抽样设计

抽样设计：指在从所确定的总体中搜集样本数据之前，事先确定的抽样程序或方案。



系统抽样

- 把标志值根据标志排队
- 把所有的标志值分成K份
  - K=N/n
- 从1----K中随机抽取
  - 一个样本单位
- 以后每隔K距离抽取一个
  - 样本单位，直到抽到n个单位

分层抽样

- 把总体分成若干层
  - 使每一层层内差距缩小
  - 层间差距加大
- 从每一层中抽取相应的样本量
  - 比例分配法
  - 最优分配法

$$n_i = \frac{N_i}{N} n$$

$$n_i = \frac{N_i \partial_i}{\sum N_i \partial_i}$$

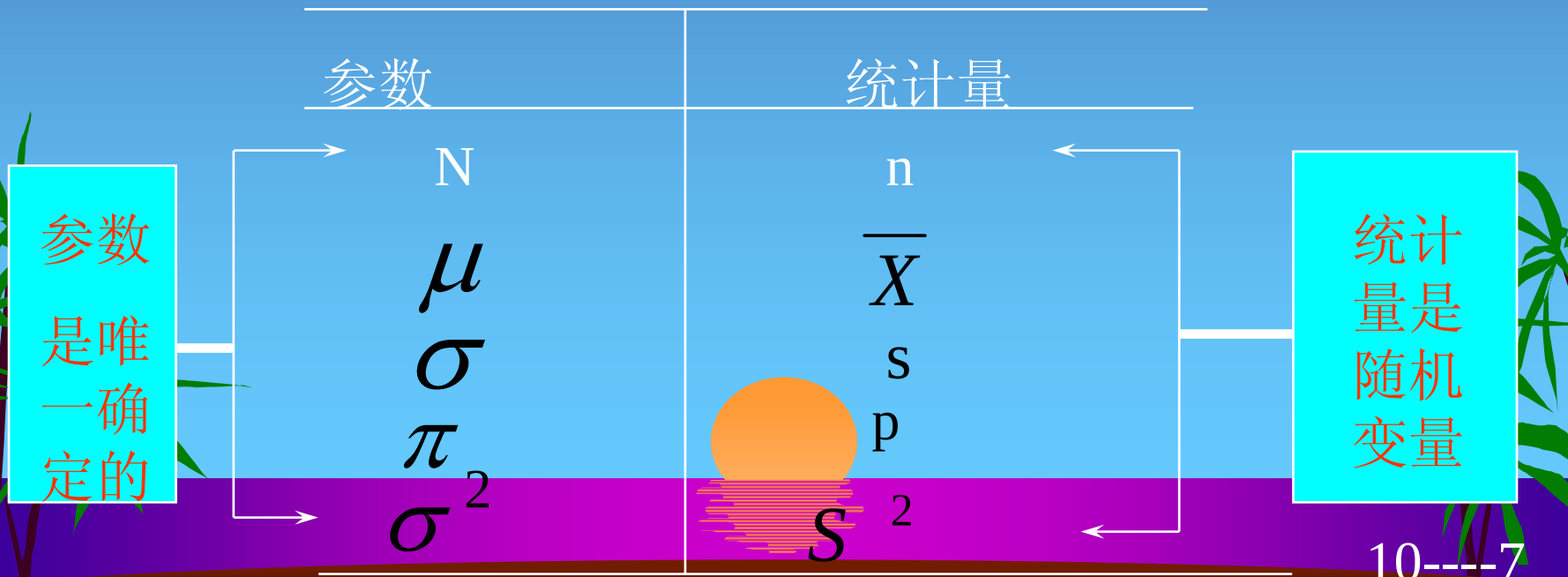
整群抽样

- 把总体分成若干群
  - 使群内各单位的差异与总体内各单位的差异相似
  - 群间差异小
- 根据简单随机抽样从中抽取一群或几群
- 对被抽中的群内的所有的单位进行全面观察

第二节 抽样分布的基本原理

1. 参数和统计量

- 参数：反映总体分布特征的指标
- 统计量：反映样本分布特征的指标



第二节 抽样分布的基本原理

2.样本平均数的概率分布

$\overline{X}$ 的概率 分布

当:X服从正态 分布和标准差已知时

$\overline{X}$ 也服从正态分布

正态分布再生定理

当X服从任意分布时,只要 n>30

$\overline{X}$ 逼近正态分布

中心极限定理

当X服从正态分布,标准差未知并且n<=30

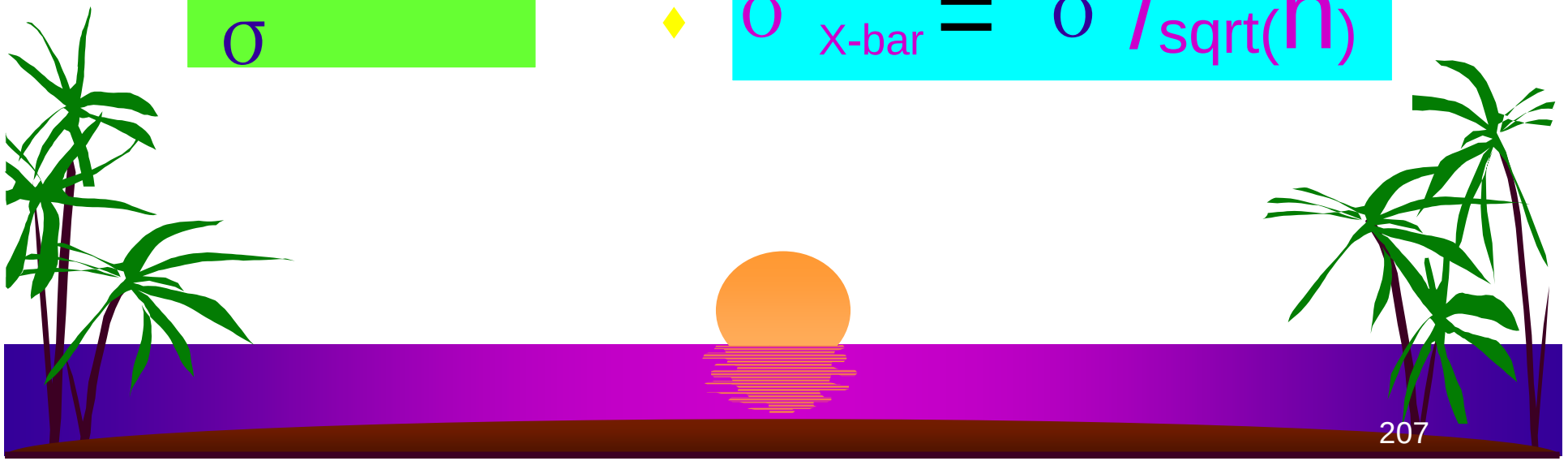
$\overline{X}$ 服从t 分布

小样本定理

样本平均数抽样分布的量数与总体参数的关系

- 总体平均数 μ
- 总体标准差 σ

- $\mu_{\text{x-bar}} = \mu$
- $\sigma_{\text{x-bar}} = \sigma / \text{sqrt}(n)$



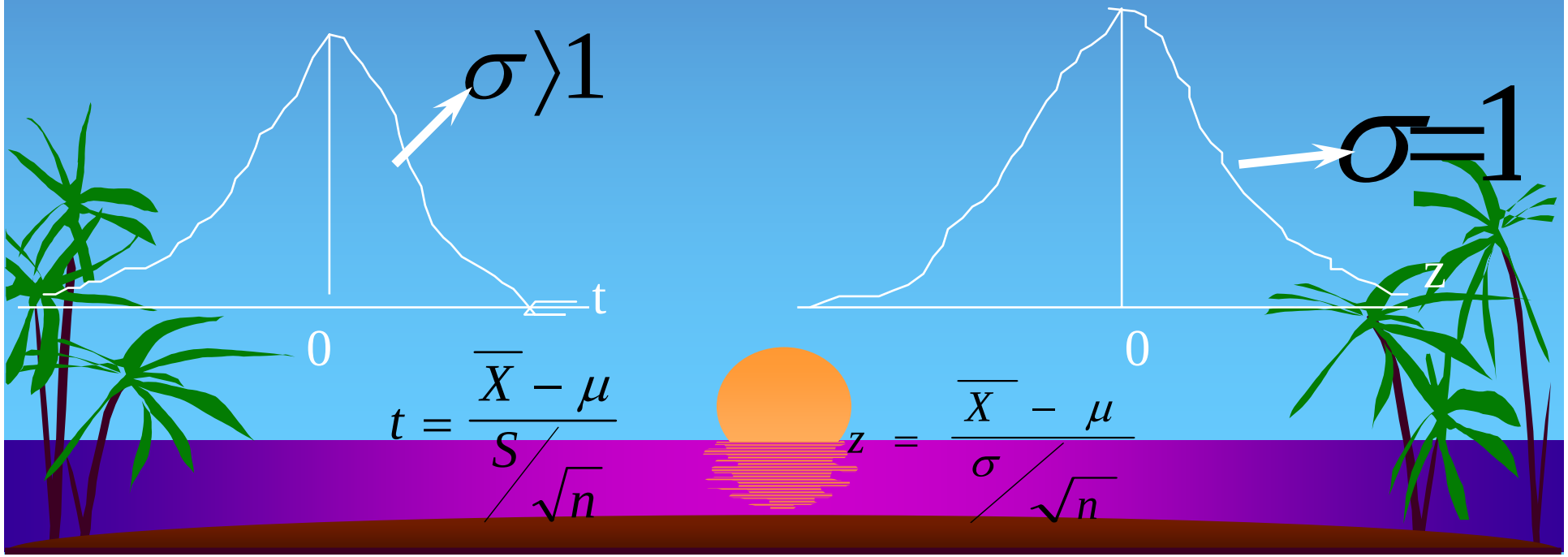
样本平均数服从t分布时的期望值和标准差

$$E\left(\overline{X}\right) = E\left(X\right) = \mu$$

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

其中

$$s = \sqrt{\frac{\sum\left(x-\overline{X}\right)^2}{n-1}}$$



$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

t分布(1)

t分布的性质：

（1）t分布是对称的钟形曲线，　　以E（t）= 0 为对称轴。

（2） t曲线的离散程度强于正态曲线，标准差大于1。当 $t = \dfrac{(\bar{X} - \mu)}{S/\sqrt{n}}$

n增大时，其标准差就趋于1，t分布逼近于正态分布。

$n \geq 30$

σ

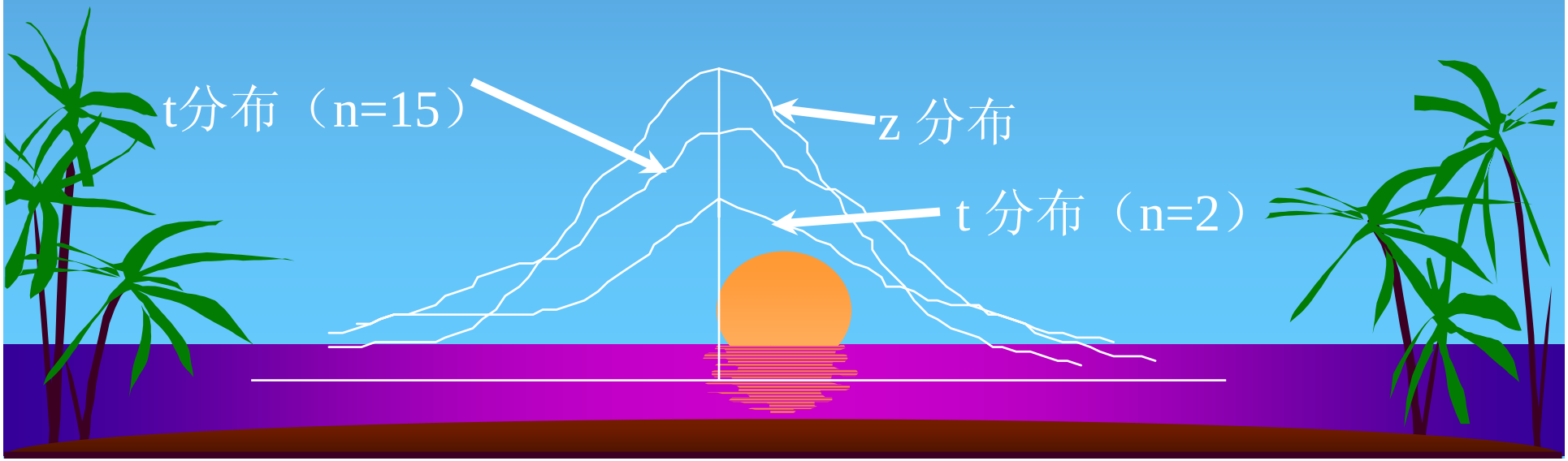
t分布(2)

（3）t分布是一个分布族，不同的自由度(n-1)

对应于不同的分布。但它们的均值都等于零。

（4）与标准正态分布相比。t分布的中心部位较低，两个尾部较高

t分布与z分布的比较图



自由度

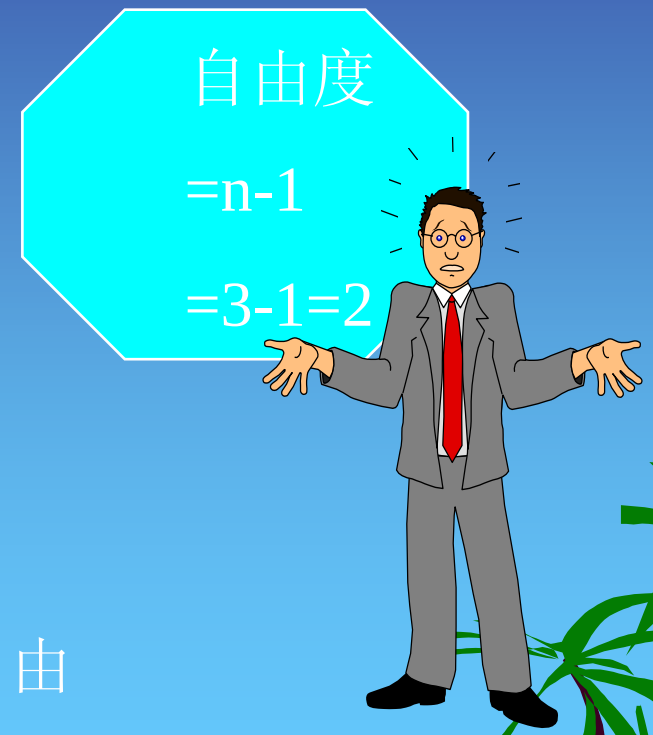
- 自由度是指可以自由选择的数值的个数
- 例如
  - 3个数的和等于6

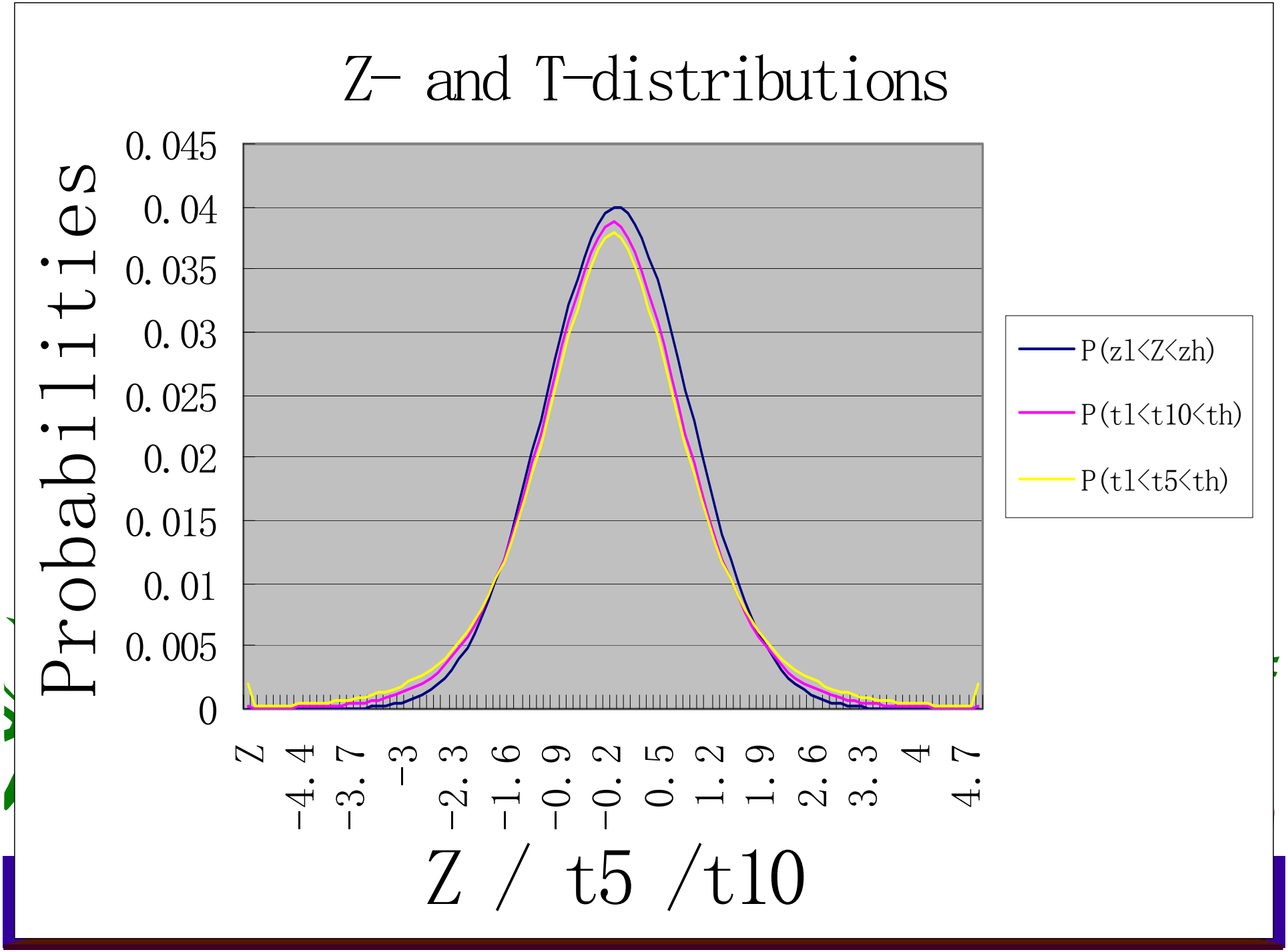
$X_1 = 1$ （可取任意数）

$X_2 = 2$ （可取任意数）

$X_3 = 3$ 只能取3，没有自由

6




Z- and T-distributions
Probabilities
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Z
-4.4
-3.7
-3
-2.3
-1.6
-0.9
-0.2
0.5
1.2
1.9
2.6
3.3
4
4.7
Z / t5 /t10
P(z1<Z<zh)
P(t1<t10<th)
P(t1<t5<th)

第二节 抽样分布的基本原理

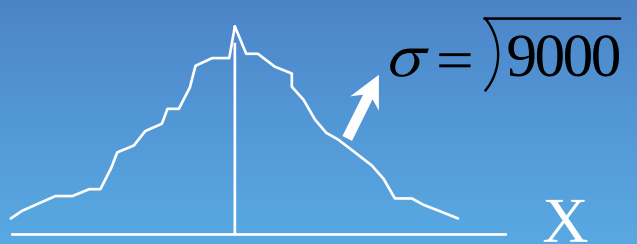
3. $\overline{X}$ 的期望值和标准差

$\overline{X}$ 服从正态分布时

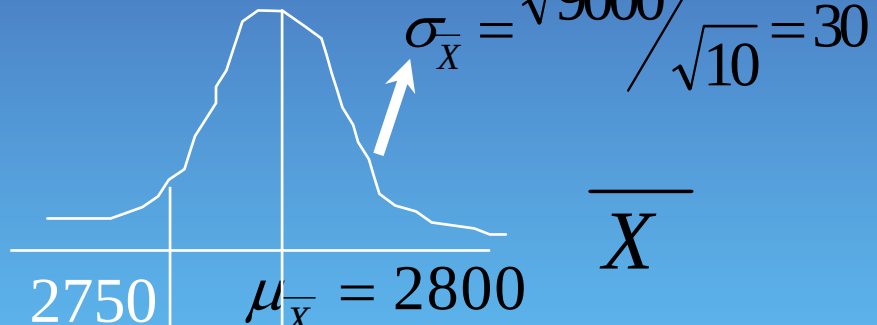
- 有放回抽样或 无放回抽样,但 N很大($N \geq 20n$)
  - $E(\overline{X}) = E(X) = \mu$
  - $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$
- 无放回抽样且 $N < 20n$
  - $E(\overline{X}) = E(X) = \mu$
  - $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} \sqrt{\dfrac{N-n}{N-1}}$

应用案例 1

X表示某类钢制产品的重量，它服从正态分布，并且知道其平均数为2800公斤，方差为9000公斤。现假设从该总体中抽出样本容量为10的随机样本，问这个样本的平均重量小于或等于2750公斤的概率为多少



总体分布

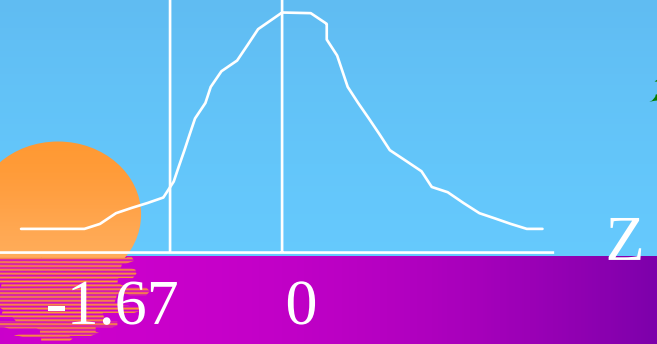


$\bar{X}$ 的抽样分布

$$p(\bar{X} \le 2750) = p\left(z \le \frac{2750-2800}{30}\right)$$

$$= p(z \le -1.67) = 0.5 - 0.4525$$

$$= 0.0475$$



应用案例2

从某地区统计中得知,该地区郊区平均每一家庭年收入为3160元,标准差为800元.从此郊区抽取50个家庭为一个随机样本

问这个样本的平均每个家庭年收入为以下数值的概率是多少:

(1) 多于3000元

(2) 少于3000元

(3) 在3200元到3300元之间

解:X为每一家庭的年收入 $\mu = 3160, \sigma = 800$

由于n=50大于30,所以 $\overline{X}$ ——〉 N(3160,800/$\sqrt{50}$)的正态分布

应用案例2的计算结果

$$(1).\ p\left(\overline{X}\rangle 3000\right)=p\left(\overline{X}\geq 3000\right)=p\left(Z\geq\frac{3000-3160}{800/\sqrt{50}}\right)$$

$$=p(Z\geq -1.41)=0.5+0.4207=0.9207$$

$$(2).\ p\left(\overline{X}\langle 3000\right)=p\left(\overline{X}\leq 3000\right)=1-0.9207=0.0793$$

$$(3).\ p\left(3200\leq\overline{X}\leq 3300\right)=p\left(\frac{3200-3160}{800/\sqrt{50}}\leq Z\leq\frac{3300-3160}{800/\sqrt{50}}\right)$$

$$p(0.35\leq Z\leq 1.24)=0.2557$$

应用案例3

从海外某地区进口一批大豆,总共为1000包,已知大豆平均每包的重量为100公斤,标准差为4公斤.现从这批大豆中按无放回抽样抽取样本容量为500的样本,问样本平均数小于或等于99.5公斤的概率

$$\mu_{\overline{X}} = 100$$

$$\sigma_{\overline{X}} = \frac{4}{\sqrt{500}}\sqrt{\frac{1000-500}{1000-1}} = 0.1266$$

$$p\left(\overline{X} \le 99.5\right) = p\left(Z \le \frac{99.5-100}{0.1266}\right) = p(Z \le -2.86)$$

$$= 0.00005$$

第二节 抽样分布的基本原理

3.比率的抽样分布(比率的概率分布)

p的概率分布 当 $n\pi \rangle 5$ $n(1-\pi) \rangle 5$ 时 p 服从正态分布

$$E(p) = \pi$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

应用案例4

假定我们已知办公室人员所填写的表格中有5%至少包括一处笔误.如果我们检查一个由475份表格组成的简单随机样本,其中至少含一处笔误的表格所占的比例在3%和7.5%之间的概率有多大.

$\pi = 0.05$

$n = 475$

$n\pi = 23.75$

$n(1-\pi) = 451.25$

大于5

p 服从正态分布

$E(p) = \mu_p = 0.05$

$\sigma_p = \sqrt{\frac{0.05 \times 0.95}{475}} = 0.01$

$$p(0.03 \le p \le 0.075) =$$

$$p\left(\frac{0.03 - 0.05}{0.01} \le p \le \frac{0.075 - 0.05}{0.01}\right)$$

$$= p(-2 \le z \le 2.5) = 0.971$$

第十一章 参数估计

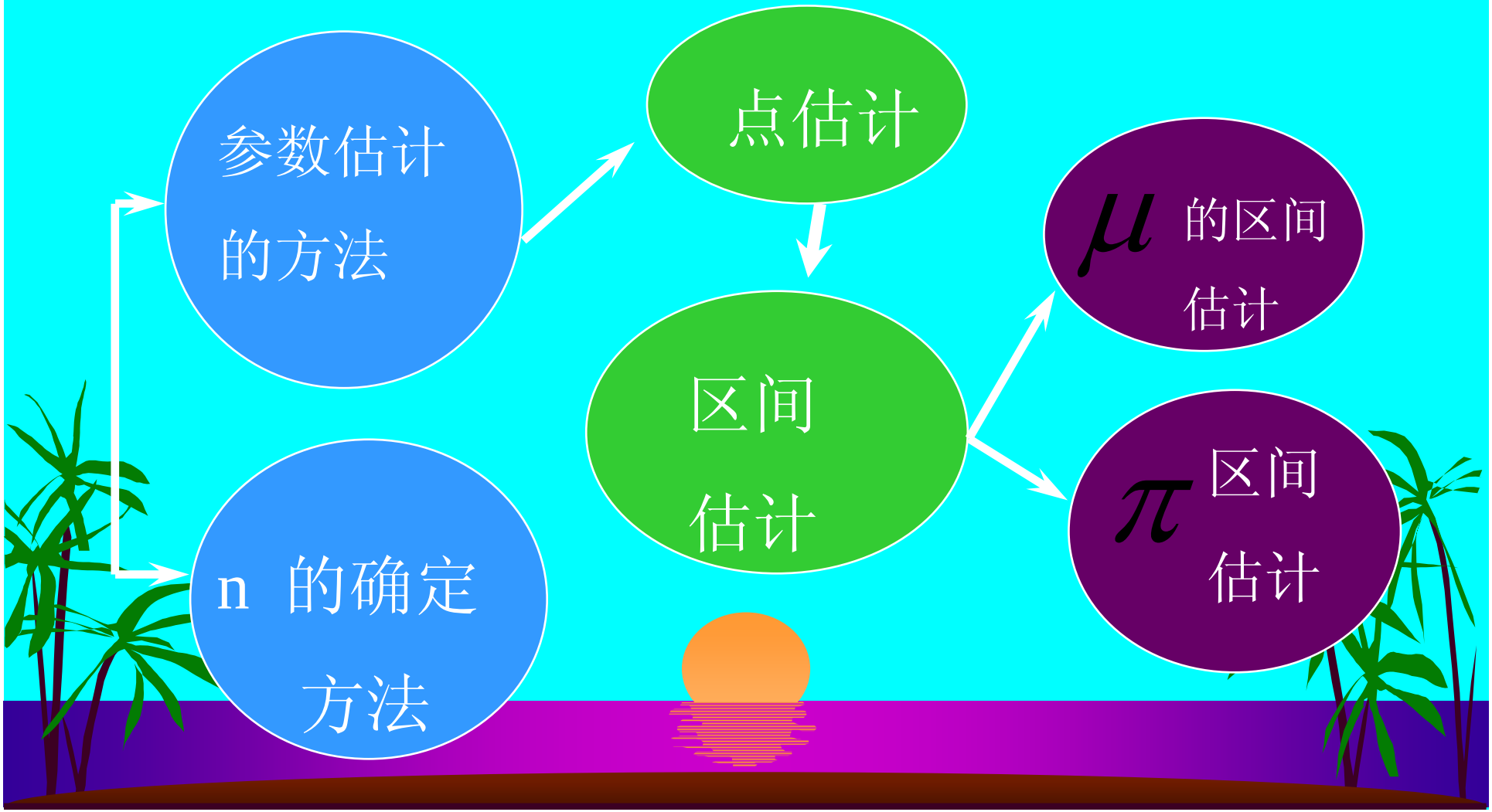
第十章的内容

样本 $\overline{X}$ 怎样 ?

本章内容

样本 $\overline{X}$ 已知

第十一章　参数估计----本章重点



优良估计量的选择标准

设θ为待估计参数，$\hat{\theta}$ 为估计量

1. 无偏性 $E(\hat{\theta}) = \theta$

2. 有效性 $Var(\hat{\theta}_1) > Var(\hat{\theta}_2)$, $\hat{\theta}_2$较为有效

3. 一致性 随着n增大， $Lim\ \hat{\theta}$ 趋近于θ

4. 充分性：充分利用了样本信息

区间估计的基本原理

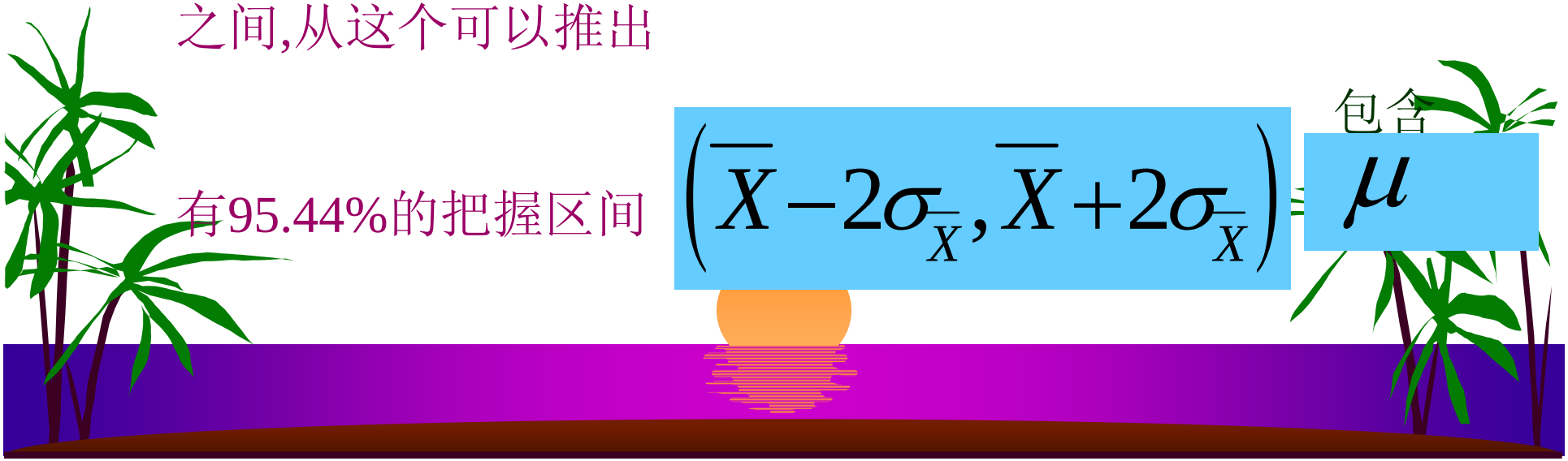
据上一章我们知道，如样本平均数服从正态分布

那么有 $p\left(\mu-2\sigma_{\overline{X}} \le \overline{X} \le \mu+2\sigma_{\overline{X}}\right)=0.9544$

也就是说有95.44%的样本平均数在 $\left(\mu-2\sigma_{\overline{X}},\mu+2\sigma_{\overline{X}}\right)$

之间,从这个可以推出

有95.44%的把握区间 $\left(\overline{X}-2\sigma_{\overline{X}},\overline{X}+2\sigma_{\overline{X}}\right)$ 包含 μ

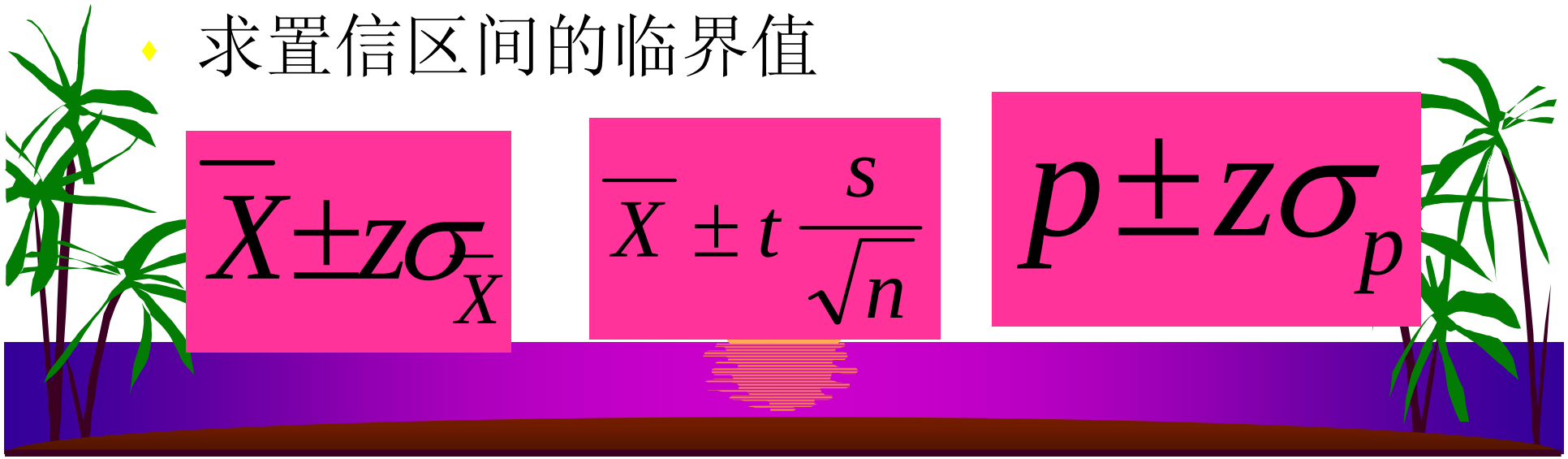


区间估计的几个关键概念

- 置信系数 $(1-\alpha)$使人相信区间包含总体均值的概率,一般取 0.95,0.90,0.99.它的大小说明估计的把握性
  的大小.
- 置信区间:在一定概率的保证下,包含总体均值的区间的宽窄，说明估计的可靠程度。区间越宽,估计的可靠程度越大; 反之就越小。
- 置信限: 确定置信区间起始的数值上限和下限
- 注意置信系数和 区间宽窄的关系

区间估计的步骤

- 选定置信系数
- 确定统计量的概率分布
- 抽取一个样本容量为n的样本
- 计算 $\overline{X}$ 当 σ 未知时还要计算s
- 求置信区间的临界值



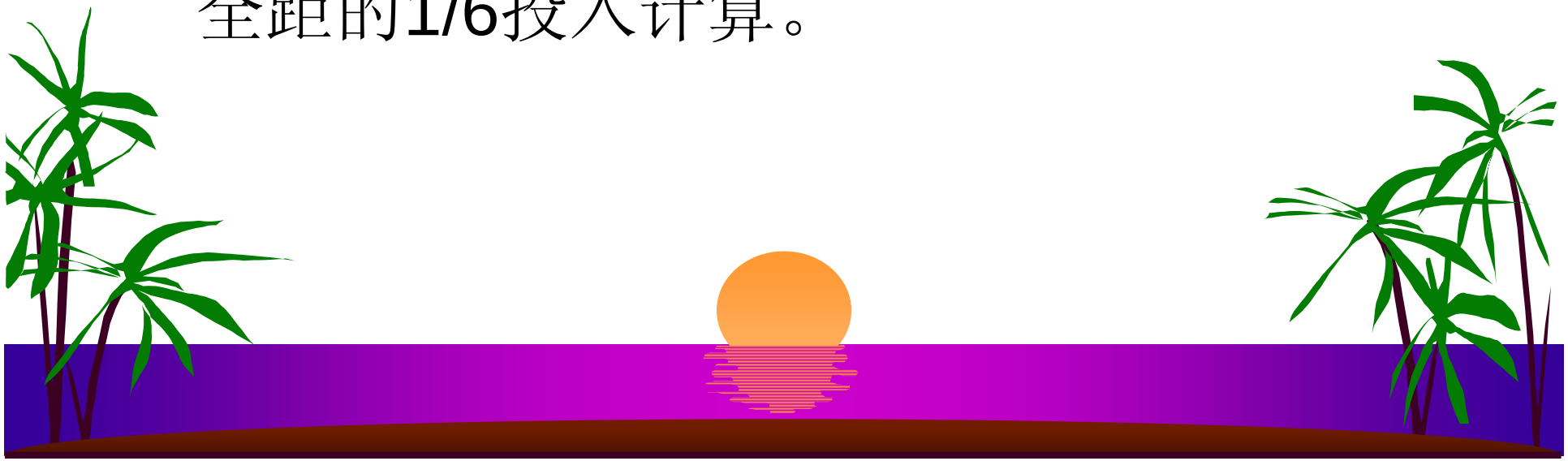
$$\overline{X} \pm z\sigma_{\overline{X}}$$

$$\overline{X} \pm t\frac{s}{\sqrt{n}}$$

$$p \pm z\sigma_p$$

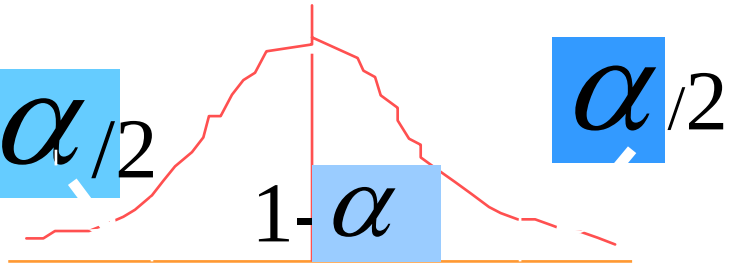
σ的获知源

- 从同一总体历史数据获知，条件是离散状况无大变化；
- 从类似的其它总体数据获知；
- 当本总体的最大值和最小值可知时，以全距的1/6投入计算。



置信系数和置信区间

$\overline{X}$ 的抽样分布



区间的临界值为

$\overline{X} \pm z\sigma_{\overline{X}}$

$(1-\alpha)\%$ 的区间包含 μ

$\alpha\%$ 不包括

区间估计的应用案例1

样本取自于总体标准差已知的正态分布

某质量管理部门的负责人估计一批原材料的平均重量。抽取样本容量为250的一个随机样本，测得　　样本平均数为为65千克。已知总体标准差为15千克，假设原材料每包的重量服从正态分布。求置信系数为95%的这批原材料平均重量的置信区间.

解：根据已知条件可知样本平均数服从正态分布

μ 的置信区间的临界值为 $65 \pm 1.96 \frac{15}{\sqrt{250}} = 65 \pm 1.86$

区间估计的应用案例2

样本取自总体标准差已知的非正态总体

某职业介绍所的职员从申请某一职业的1000名申请者中

采用不重服抽样方式随机抽取了200名申请人，借此来估计1000名申请者考试的平均成绩。样本平均数为78分，由已

往经验得知总体的方差为90分。求总体平均数的95%的置信区间。

解：由于n大于30，根据中心极限定理$\overline{X}$ 服从正态分布

$$\sigma_{\overline{X}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{(N-n)}{N-1}}=\frac{\sqrt{90}}{\sqrt{200}}\sqrt{\frac{1000-200}{1000-1}}=0.6$$

μ 的置信区间为 $(78-1.645*0.6,78+1.645*0.6)$

得出：有90%的把握总体均值在（77　79）区间之内。

区间估计的应用案例3

σ 未知，大样本时总体平均数的区间估计

某百货商店通过100位顾客的随机样本研究购买额。样本均值为247。5元，样本标准差为55元。求逛此商店的所有顾客平均购买额的99%的置信区间。

解：总体标准差未知时，严格来说，样本平均数应服从t分布，但由于n=100大于30。所以可以用正态分布逼近它。

μ 的置信区间为

$$\bar{X} \pm z\frac{S}{\sqrt{n}} = 247.5 \pm 2.58 * \frac{55}{10} = 247.5 \pm 14.19$$

$$= (233.31, 261.69)$$

间估计的应区用案例4

总体标准差未知，小样本的正态

总体平均数的区间估计

为了估计一分钟一次广告的平均费用，抽出了15个电视台的随机样本。样本的平均值为2000元，样本的标准差为1000元假定一分钟一次广告的费用服从正态分布，求总体平均数的95%的置信区间。

解：已知：X=2000元，S=1000元，n=15，自由度（n-1）

μ 间为

$$\overline{X} \pm t\frac{s}{\sqrt{n}} = 2000 \pm 2.14\frac{1000}{\sqrt{15}} = (1447.5, 2552.5)$$

区间估计的应用案例5
总体比率的区间估计

某企业在一项关于寻找职工流动原因的研究中，研究人员从该企业前职工的总体中随机抽取了200人的一个样本。在对他们进行访问时，有140人说他们离开该企业的原因是因为收入太低。求由于这种原因而离开该企业的人员的真正比率的95%的置信区间。

解：因np=200*0.7=140,n(1-p)=200*0.3=60.它们都大于5.所以p 可以用正态分布逼近它.

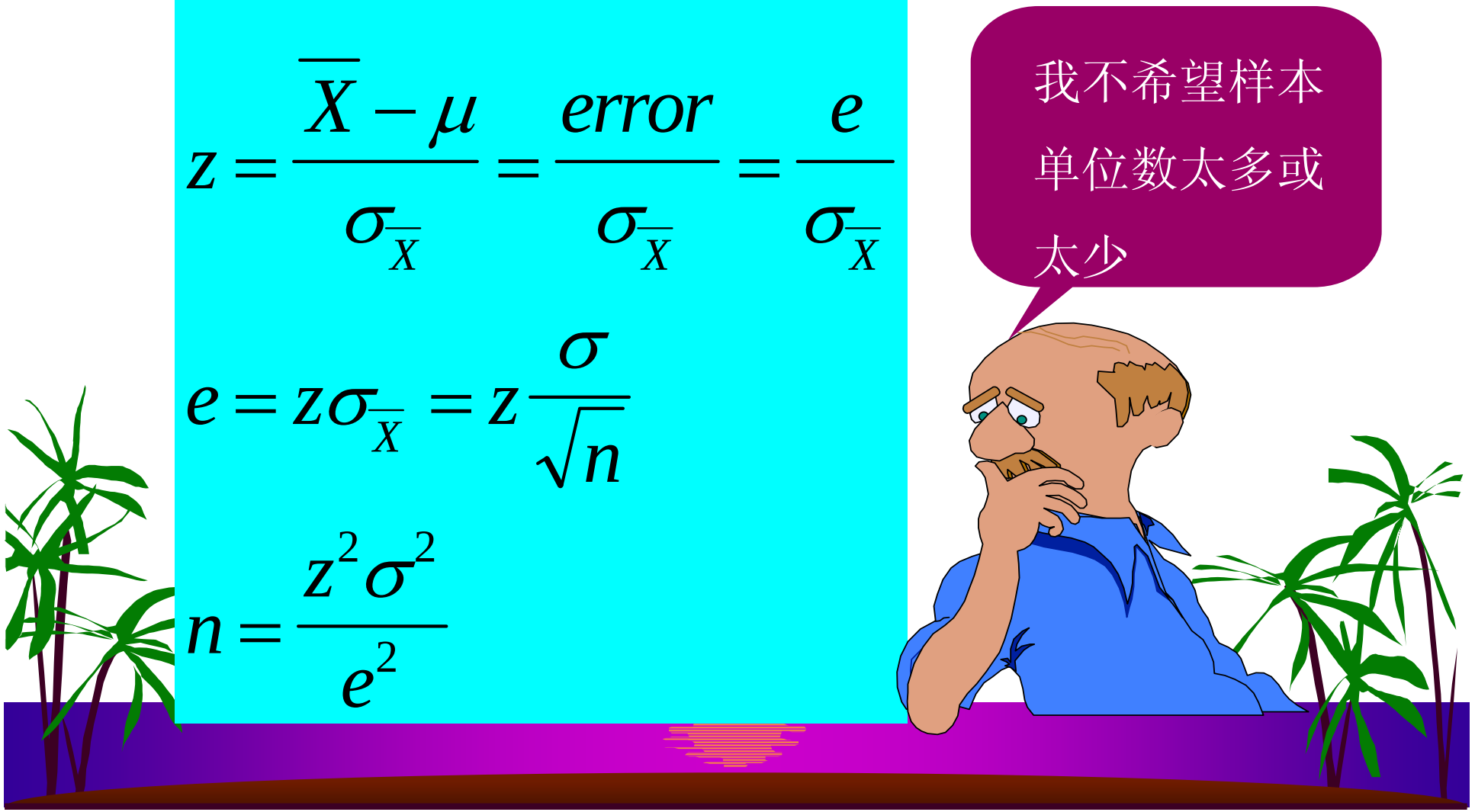
$$p \pm z\sigma_p = 0.7 \pm 1.96\sqrt{\frac{p(1-p)}{n}} = 0.7 \pm 1.96\sqrt{\frac{0.7*0.3}{200}}$$
$$= 0.7 \pm 0.064 = (63.6\%, 76.4\%)$$

估计μ 时样本容量的确定

$$z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{error}{\sigma_{\overline{X}}} = \frac{e}{\sigma_{\overline{X}}}$$

$$e = z\sigma_{\overline{X}} = z\frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z^2\sigma^2}{e^2}$$



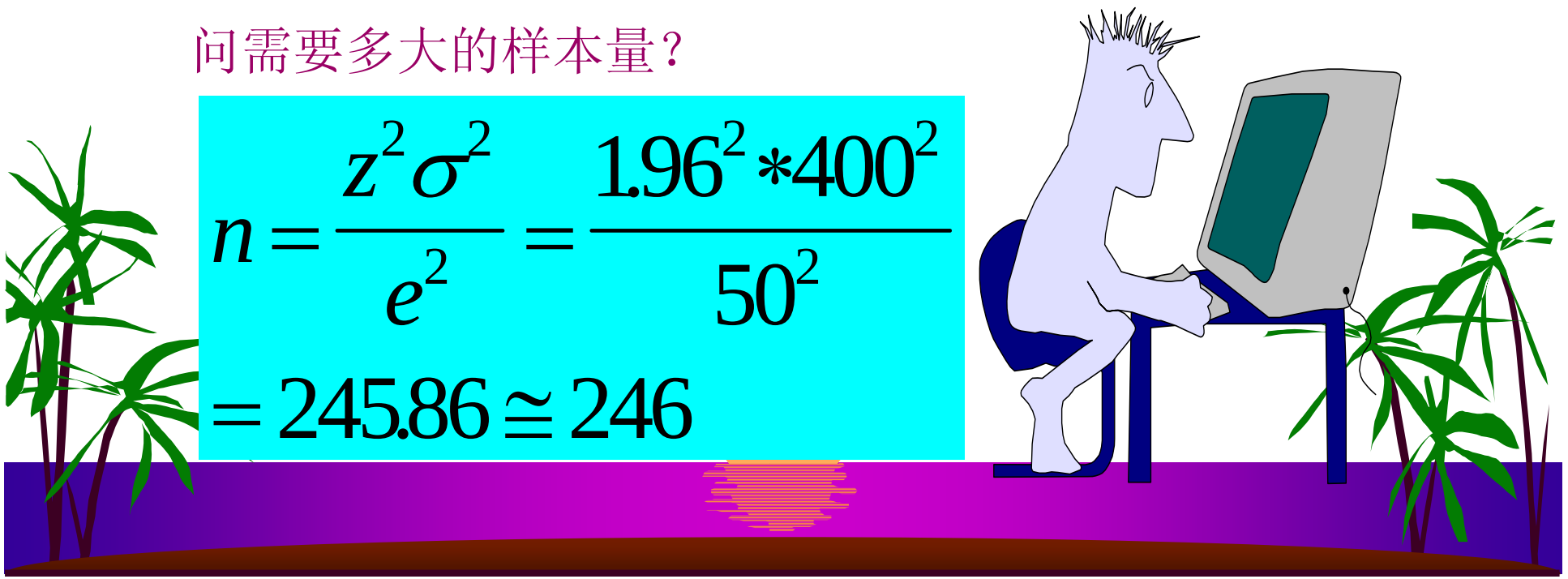
估计 μ 时样本容量确定的应用案例

你在某公司人事部门工作，计划调查雇员每年人均医疗费用，需要有95%的把握使估计误差控制在 $\pm$ 50 元之间，根据以往经验得知总体标准差为400元。

问需要多大的样本量？

$$n=\frac{z^2\sigma^2}{e^2}=\frac{1.96^2*400^2}{50^2}$$

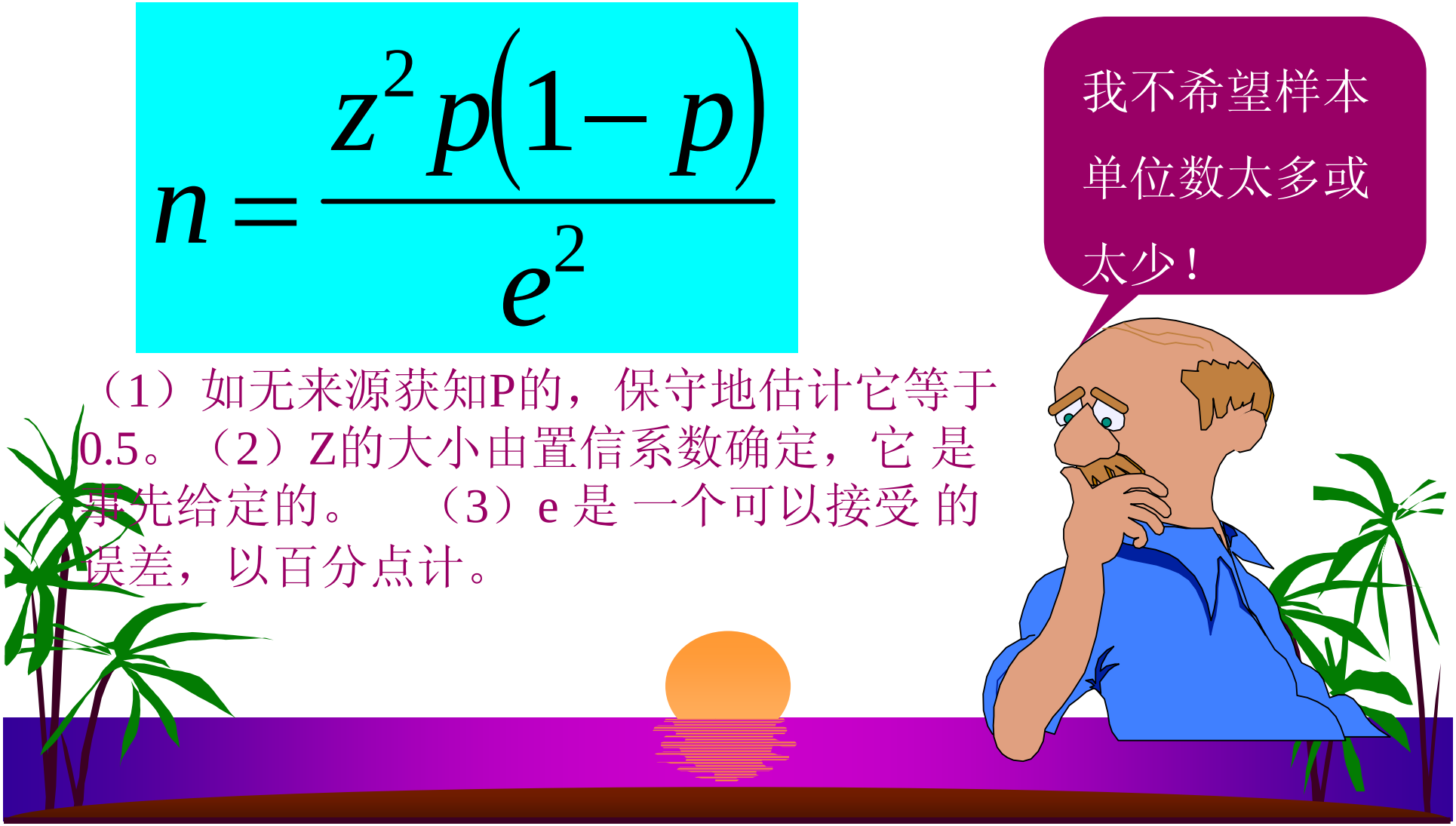
$$=245.86\cong 246$$



估计π时样本容量的确定

$$n = \frac{z^2 p(1-p)}{e^2}$$

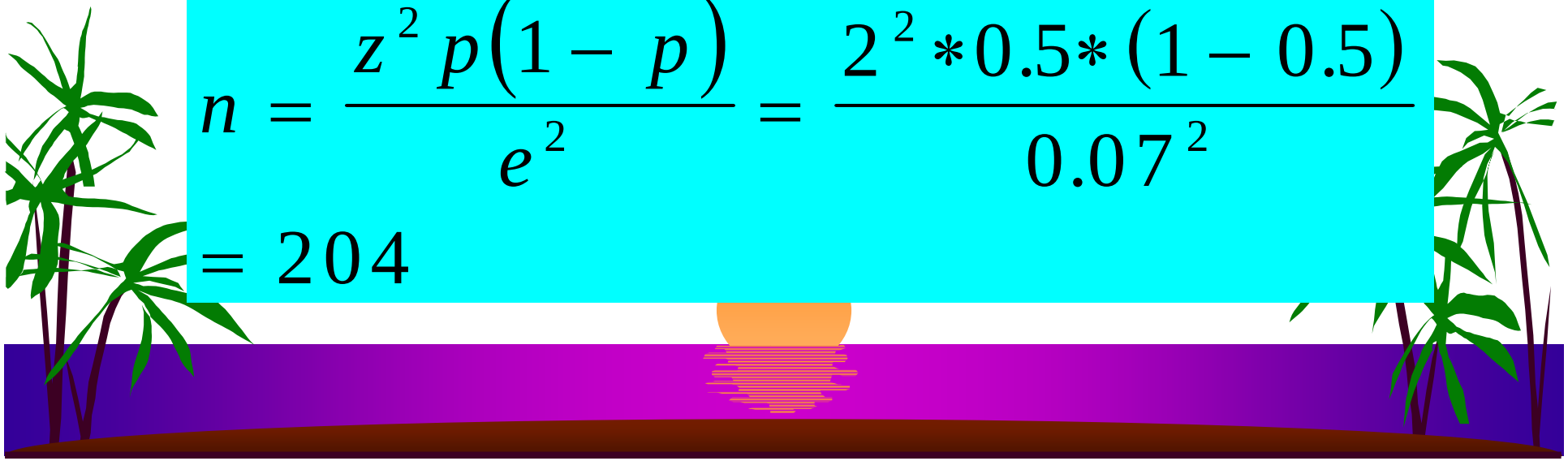
（1）如无来源获知P的，保守地估计它等于0.5。（2）Z的大小由置信系数确定，它 是事先给定的。 （3）e 是 一个可以接受 的误差，以百分点计。



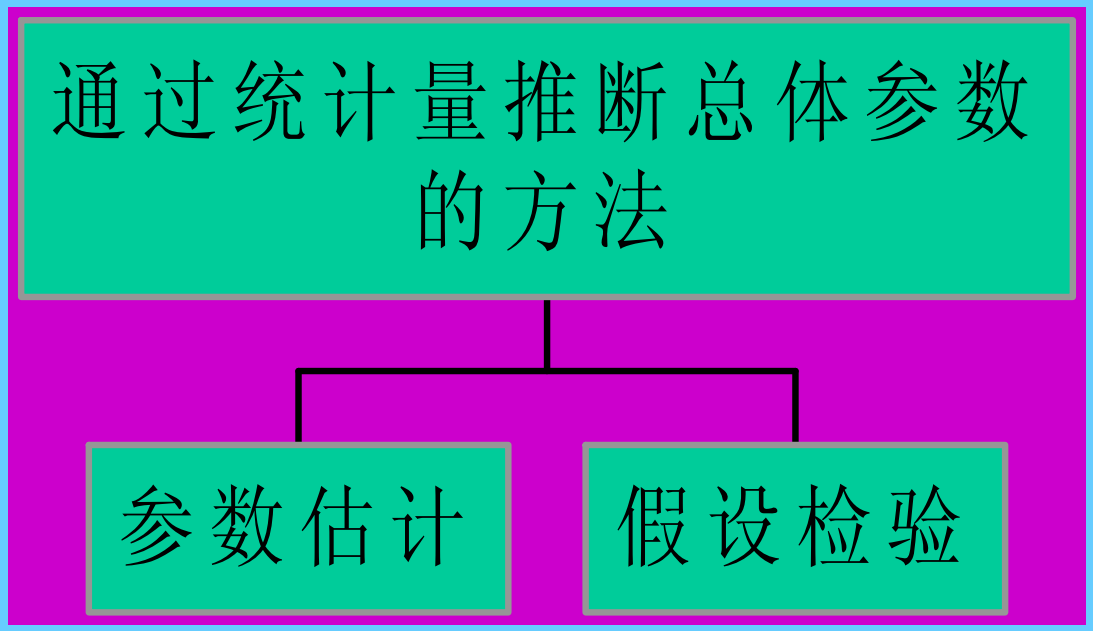
估计 π 时样本容量确定 的应用案例

一家市场调查公司希望估计某地区有29英寸彩色电视机的家庭所占的比率。该公司希望对 π 的估计误差不超过0.07，置信系数为95.44%,但没有可利用的比率的估计值.问应抽取多大容量的样本。

$$n = \frac{z^2 p(1-p)}{e^2} = \frac{2^2 * 0.5 * (1-0.5)}{0.07^2} = 204$$



第十二章 假设检验

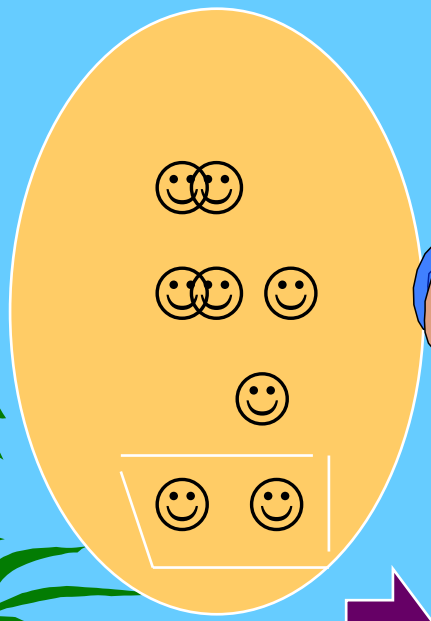


通过样本直接推断参数的单值，进而以一定把握程度确定其存在区间

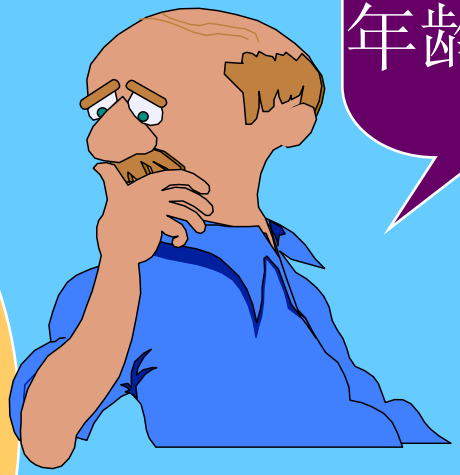
先假设总体参数具有某特征值，然后看样本提供的信息是否支持假设

假设检验

总体



我认为总体平均年龄是50（假设）



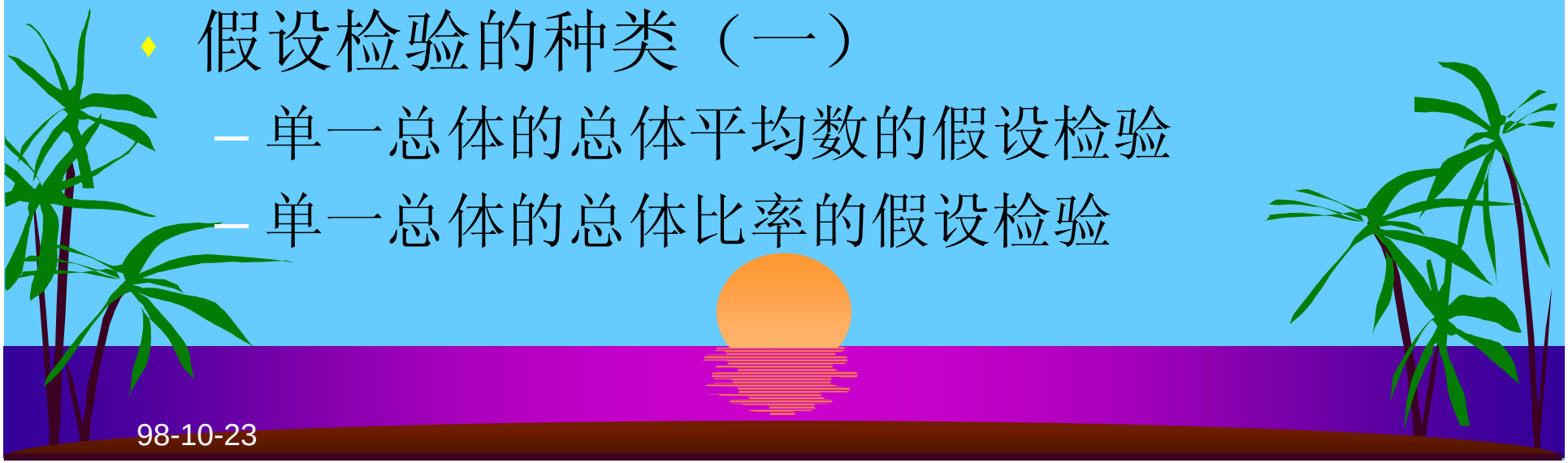
随机抽样

样本平均数为20

拒绝假设？

本章重点

- 基本概念
  - 假设
  - 显著性水平
  - 假设检验的两类误差
- 假设检验的基本程序
- 假设检验的种类（一）
  - 单一总体的总体平均数的假设检验
  - 单一总体的总体比率的假设检验



假设检验的基本原理：小概率事件在一次实验中几乎不可能发生。

若总体平均净重确是250克，样本平均数落到249.4克以下或250.6克之上的概率只有4.56%——这被认为是一个小概率事件，几乎不可能发生。

现在居然发生了样本平均数落到249.4克以下或250.6克之上的事，我们便有足够的理由怀疑总体平均净重并非是250克。

假设检验的概念

$$\leq , \geq$$

$$H_0 : \mu = 250$$

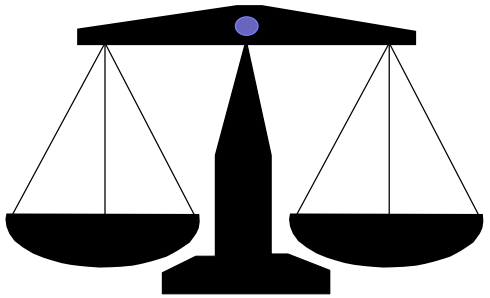
$$H_1 : \mu \neq 250$$

假设：是关于总体参数（数值）的一种陈述

- 零假设（Null Hypothesis）：
- 是受检验的基本假设，它总是带等号，可以有 =，记为：H_0
  - 如：
- 备择假设（Alternative hypothesis）：零假设的对立假设，符号不带等号，记为：H_1
  - 如

无罪推定

法官判案的例子



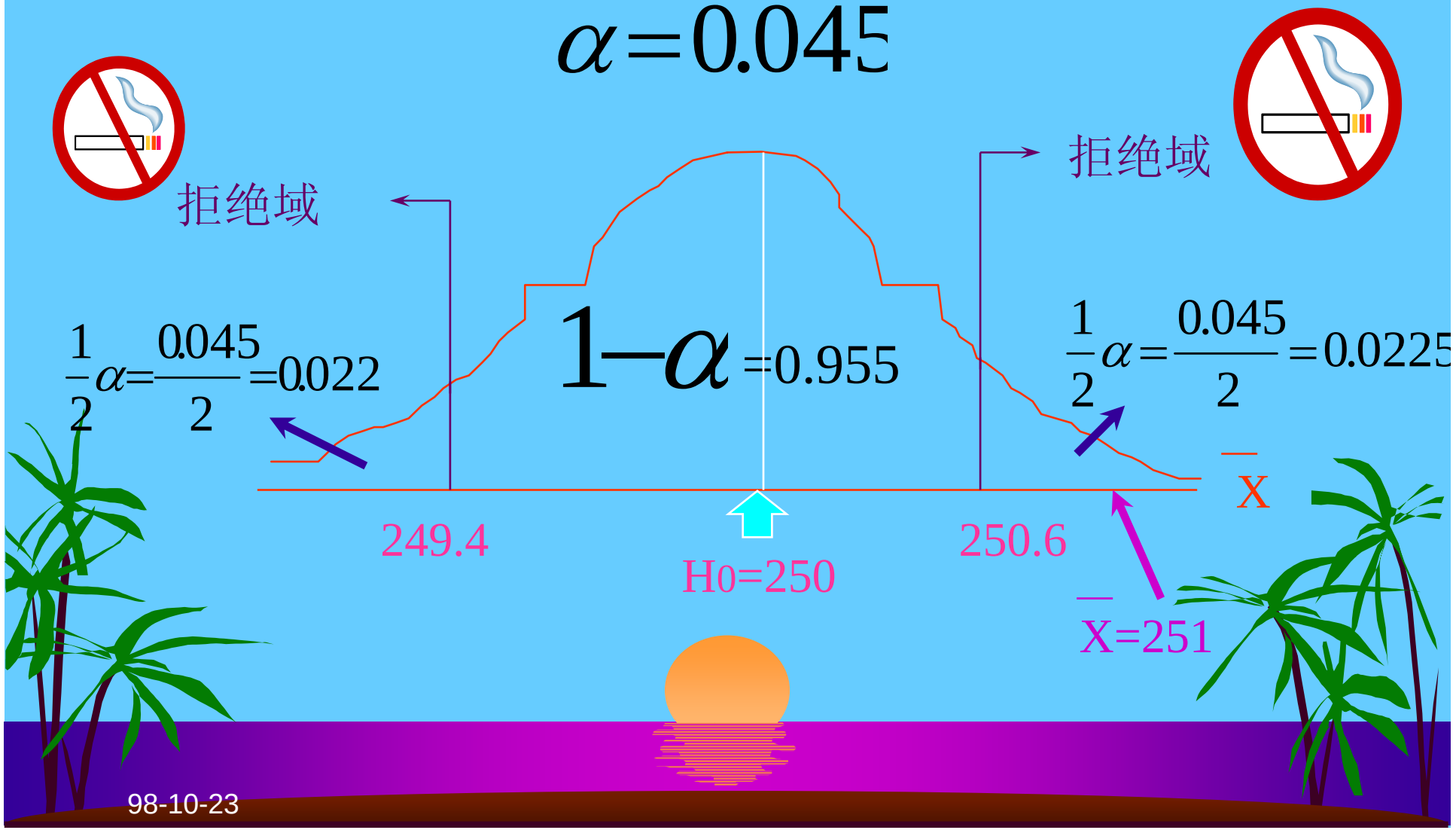
审判前的假设是 H0： 被指控人无罪

法官的可能判决	被指控人的实际情况	
	无罪?（H0是真的）	有罪（H0是假的
1 无罪	正确抉择 $(1-\alpha)$	错误抉择？犯了II型误差（ β ）
2 有罪	错误抉择？犯了I 型误差（ α ）	正确抉择（1 - β ）

假设检验的两类误差

- I 型误差
  - 拒真错误:拒绝一个真的零假设.
  - 犯这个错误的后果是很严重的.
  - 犯 这个错误的概率为 α ,又称显著性水平.
- II型误差
  - 取伪错误,即:接受了一个伪的零假设.
  - 犯 这个错误的概率为 β

拒绝域的图示


$\alpha=0.045$
拒绝域
拒绝域
$\frac{1}{2}\alpha=\frac{0.045}{2}=0.022$
$1-\alpha=0.955$
$\frac{1}{2}\alpha=\frac{0.045}{2}=0.0225$
$\bar{X}$
249.4
H0=250
250.6
$\bar{X}=251$

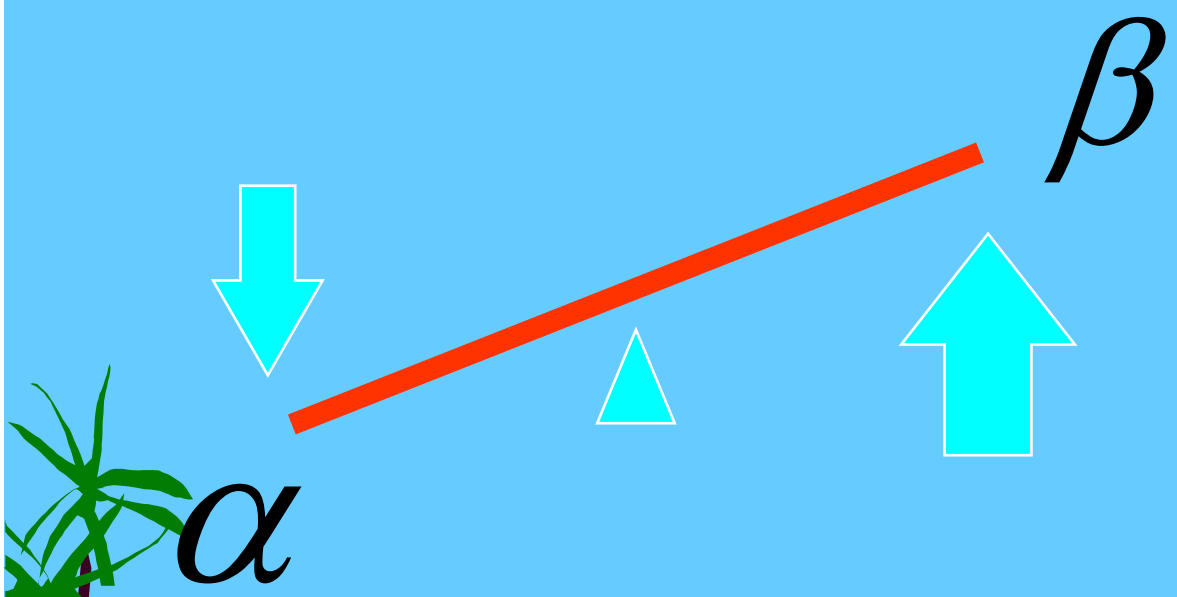
显著性水平

- 如果零假设是正确的，而根据样本的信
  - 息却拒绝了零假设的概率，叫 I 型误差，又称拒真错误。
- 用 α 表示
  - 一般取 0.01，0.05，0.10
- 显著性水平是事先确定的

α 和 β的 关 系

你能同时减少犯

两种错误的概率吗



假设检验的基本程序

- 确定零假设和备择假设
- 选定 显著性水平
- 抽取样本容量为n的样本
- 确定统计量的抽样分布
- 确定决策规则
- 计算相应的统计量
- 计算临界值
- 根据决策规则判断是否接受零假设
- 结合具体情况作出结论

假设检验的种类

- 双尾检验——拒绝域在两个尾部

  H0 $\mu = \mu_0$

  H1 $\mu \neq \mu_0$

- 单尾检验——拒绝域在一个尾部

  左 单尾检验　拒绝域在左边的尾部

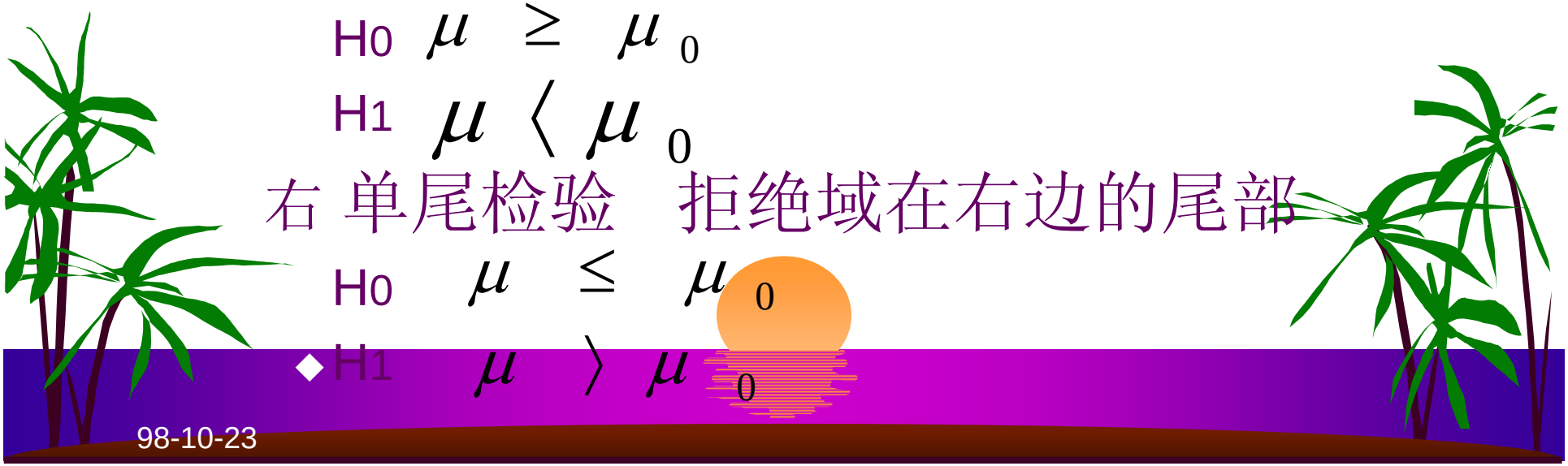
  H0 $\mu \geq \mu_0$

  H1 $\mu < \mu_0$

  右 单尾检验　拒绝域在右边的尾部

  H0 $\mu \leq \mu_0$

  H1 $\mu > \mu_0$



单一总体的总体平均数的假设检验的应用案例1----双尾检验（第404页）

Z 检验法

$H_0: \mu = 250$

$H_1: \mu \neq 250$

$\alpha = 0.05$

$n = 100$

$$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{251 - 250}{3 / \sqrt{100}}$$

$= 3.33$

$$\alpha = 0.05 \Rightarrow z = \pm 1.96$$

抉择规则：

如 $z > 1.96$ 或 $z < -1.96$

则拒绝H0

如 $-1.96 \leq z \leq 1.96$

则接受H0

兹有Z=3.33大于1.96

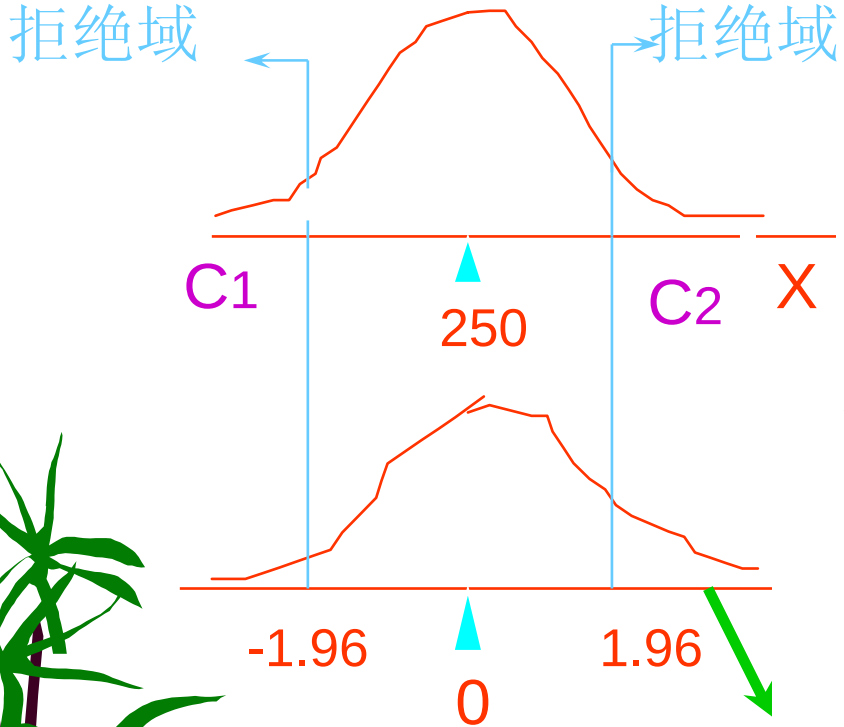
所以拒绝H0

结论:有95%的把握这批出口罐头的平均重量不等于250克

案例1续

$$C_1=\mu-1.96\frac{\sigma}{\sqrt{n}}=250-1.96*0.3=249.412$$

$$C_2=\mu+1.96\frac{\sigma}{\sqrt{n}}=250+1.96*0.3=250.558$$



抉择规则

如 $C_1 \le \overline{X} \le C_2$ 则接受H_0 否则拒绝H_0

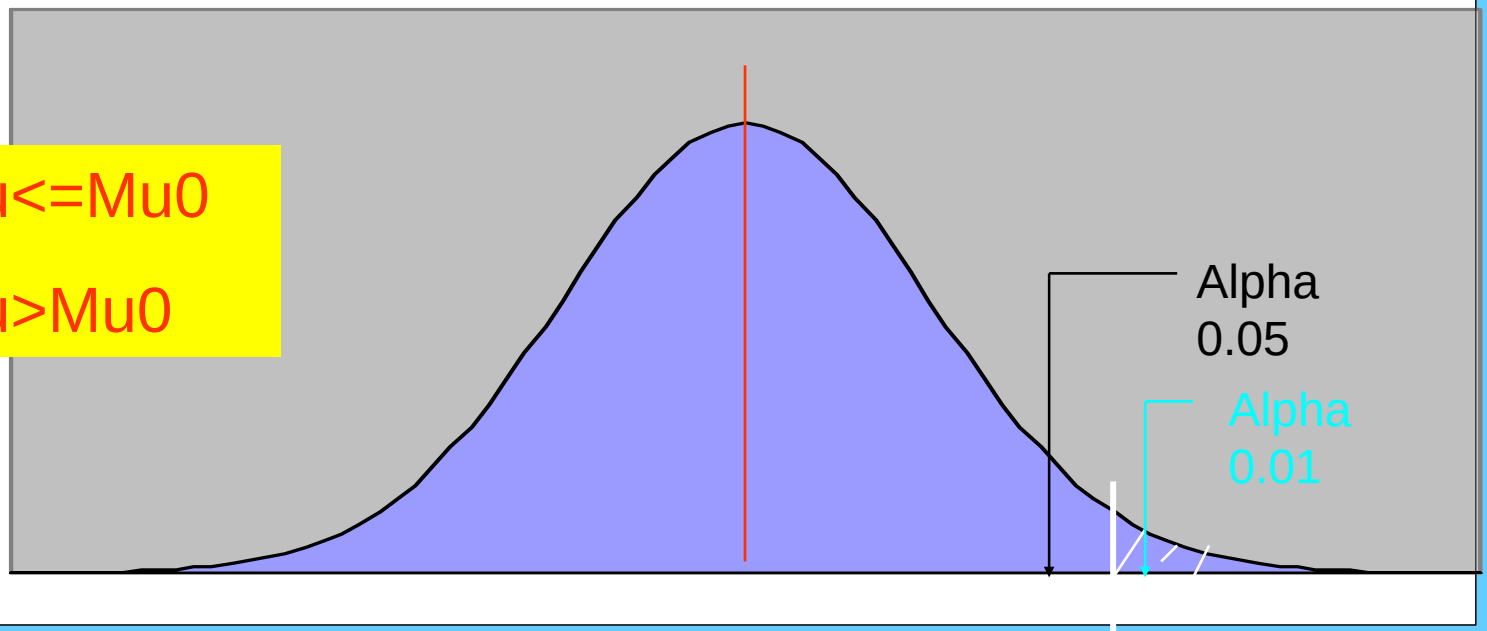
因 $\overline{X}=251\rangle C_2\left(250.558\right)$

所以拒绝H_0 接受H_1

P-值决策法

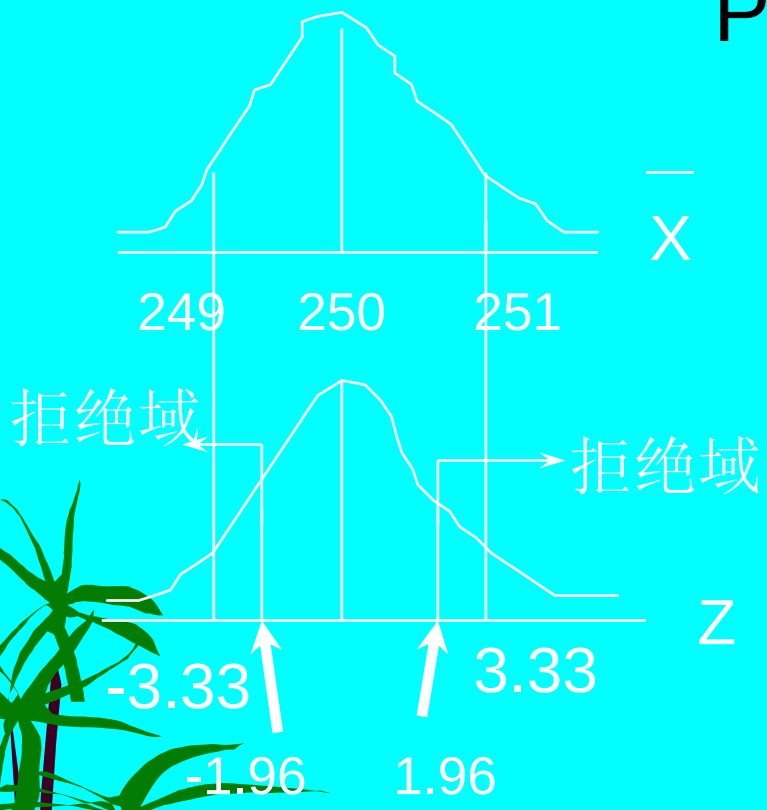
H0: Mu<=Mu0

H1: Mu>Mu0



一个较小的P-值使得决策者有较强的信心拒绝零假设；一个较大的P-值使得决策者没有较强的信心去拒绝零假设。

P值检验法应用(案例1续)



$$\text{P值} = p\left(\overline{X} \geq 251\right) + p\left(\overline{X} \leq -249\right)$$

$$= p(z \geq 3.33) + p(z \leq -3.33)$$

$$= 0.0004 + 0.0004 = 0.0008$$

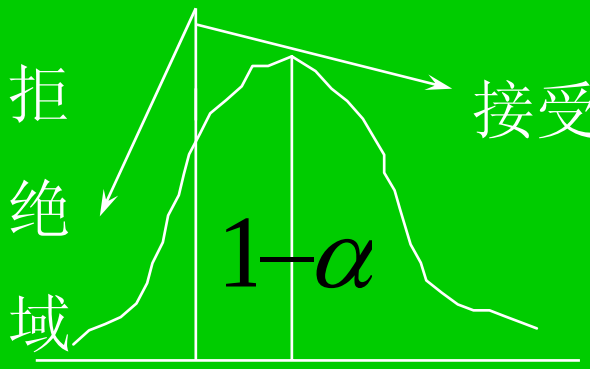
抉择规则:

P值< α 则统计量落在拒绝域内 导致拒绝H0

P值> α 则统计量落在接受域里 导致接受H0

单一总体的总体平均数的假设检验的应用案例2----左单尾检验（第410页)

Z 检验法



1. $H_0 \geq 0.2$

   $H_1 < 0.2$

2. $n = 100$

   $\bar{X} = 0.19$

   $\alpha = 0.05$

   $\sigma = 0.01$

3. $\because n > 30 \quad \therefore \bar{X}$服从正态分布

4. 计算统计量 $Z = \dfrac{0.19 - 0.20}{0.01/\sqrt{100}}$

   $= -10$

5. 如果z < -1.96,则拒绝H_0

   否则接受H_0

6. 兹有z = -10<-1.96, 所以拒绝H_0

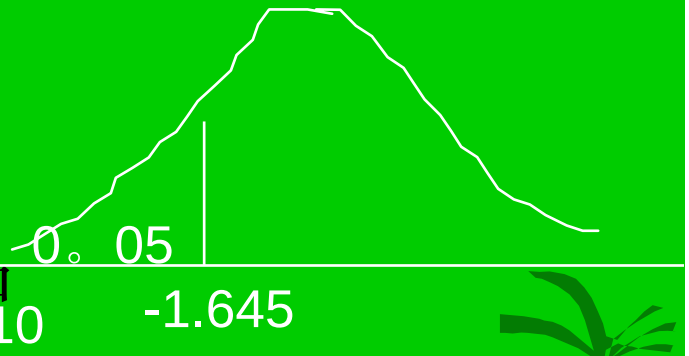
7. 结论：

P值检验法应用(案例2续)

$$P\text{值} = p(X \le 0.19) = p\left(z \le \frac{0.19-0.20}{0.01/\sqrt{100}}\right)$$

$$= p(z \le -10) \approx 0$$

$\because \alpha = 0.05 > 0$. 统计量落在拒绝域中



$\therefore$ 拒绝 H_0, 接受 H_1.

结论有近于100%的把握这种充气包的抗压力不超过0.2秒

单一总体的总体平均数的假设检验的应用案例3----右单尾检验（第413页)

Z 检验法

1. $H_0: \mu \leq 30$

$H_1: \mu > 30$

2. $n = 100, \sigma = 10$

$\alpha = 0.05$

3. $\bar{X} = 32, \sigma_{\bar{X}} = 1$

4. $z = \dfrac{32-30}{1} = 2$

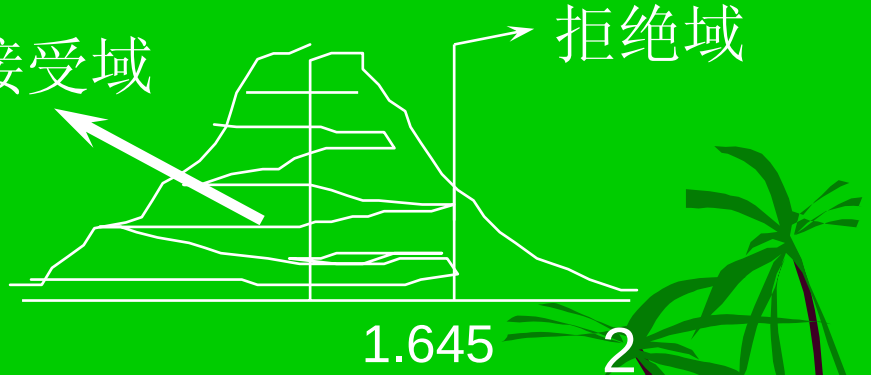
5. 如统计量z>1.645

则拒绝H_0否则接受H_0

6. $\Theta \quad z = 2 > 1.645$

$\therefore$ 拒绝H_0

7. 结论：



P----值检验法应用(案例3续)

$$P\text{值} = p\left(\bar{X} \geq 32\right) = p\left(z \geq \frac{32-30}{1}\right)$$

$$= p\left(z \geq 2\right) = 0.5 - 0.4772 = 0.0228$$

如$\alpha = 0.05 > 0.0228$ 则拒绝H_0

如$\alpha = 0.01 < 0.0228$ 则接受H_0

小样本时单一总体的总体平均数的假设检验的应用案例4

某公司人事部为一项工程上马在社会上招大批青年工人.在文化程度考核后,经理问人事部情况怎样,回答说:很好,估计平均成绩可达90分.经理随机从中抽出20份,发现平均成绩为83分,样本标准差为12分.如果经理想在0.01的显著性水平下检验人事部所做的推测的准确性,应该怎样处理?

案例4的计算结果

1. $H_0: \mu = 90 \quad H_1: \mu \neq 90$
2. $n = 20,\ s = 12 \overline{X} = 83, \alpha = 0.01$
3. 假设X服从正态分布, σ 未知, 所以 $\overline{X}$ 服从T分布
4. 统计量 $t = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} = \dfrac{83 - 90}{12/\sqrt{20}} = -2.609$
5. $\Theta\ \alpha = 0.01$ n = 20 自由度 = 19. 查表得临界值 $t^* = -2.86$
6. 如果 $-t^* \leq t \leq t^*$ 则接受 H_0, 否则拒 H_0
7. Θ t = -2.609〉- 2.86, 所以接受 H_0
8. 结论

总体比率假设检验的应用案例5

某企业的产品畅销于国内市场。据以往调查，购买该产品的顾客有50%是30岁以上的男子。该企业的负责人关心这个比例是否发生了变化，而无论是增加还是减少。于是，该企业委托了一家咨询机构进行调查，这家咨询机构从众多的购买者中随机抽选了400名进行调查，结果有210名为30岁以上的男子，该厂负责人希望在显著性水平为0.05下检验“50%的顾客是30岁以上的男子”这个假设。

案例5 的计算结果

1. $H_0:\pi=0.05,\ H_1:\pi\neq 0.05$

2. $n=400,\alpha=0.05,p=210/400=0.525$

3. $\Theta\ n\pi=400*0.5=200\rangle 5$且$n(1-\pi)=200\rangle 5$

$\therefore$ p服从正态分布

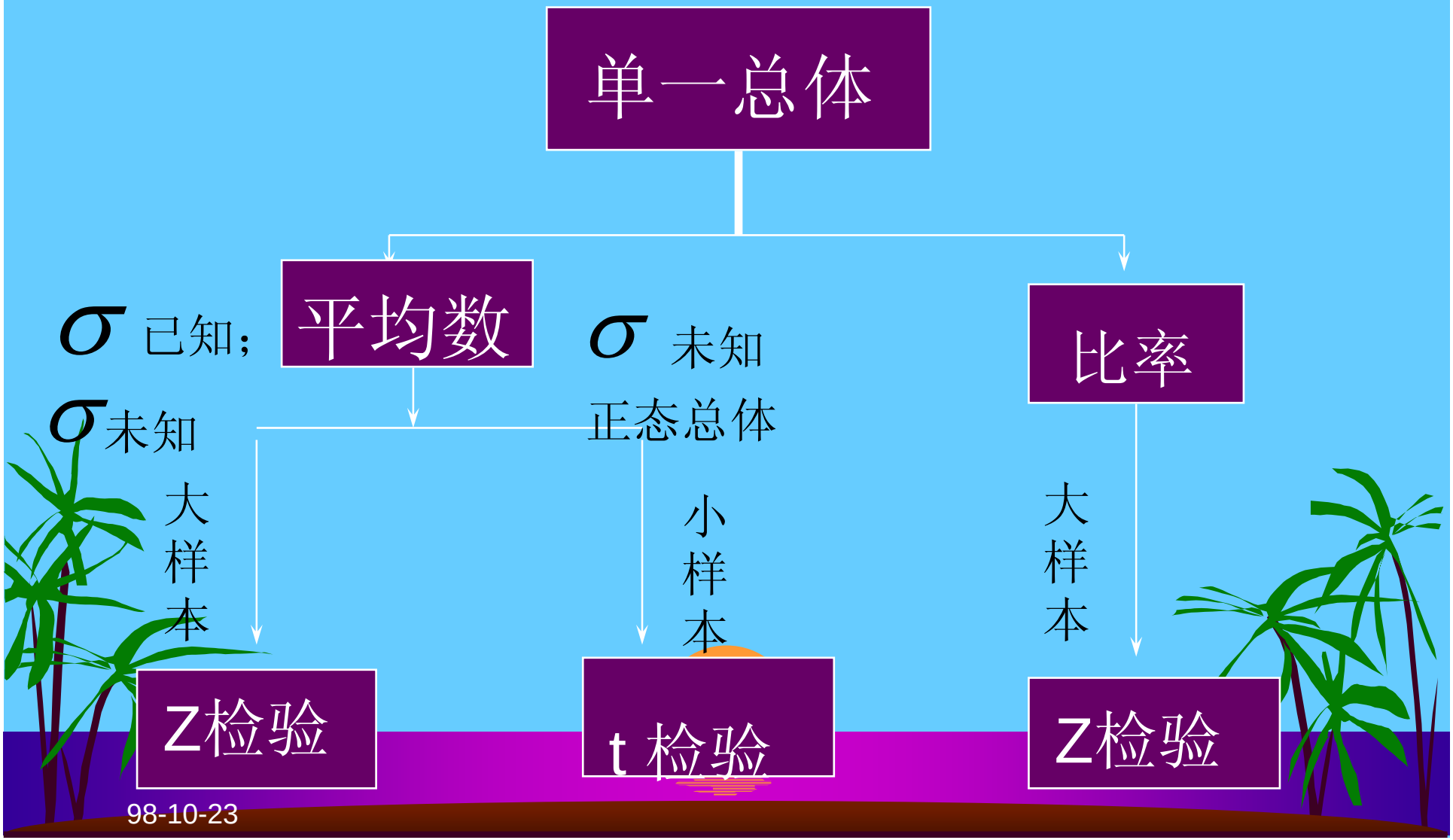
4. $z=\dfrac{p-\pi}{\sigma_p}=\dfrac{0.525-0.5}{\sqrt{\dfrac{0.5(1-0.5)}{400}}}=1$

5. 如$z<-1.96$或$z>1.96$则拒绝H_0；否则接受H_0

6. $\Theta\ z=1\leq 1.96,\therefore$接受$H_0$

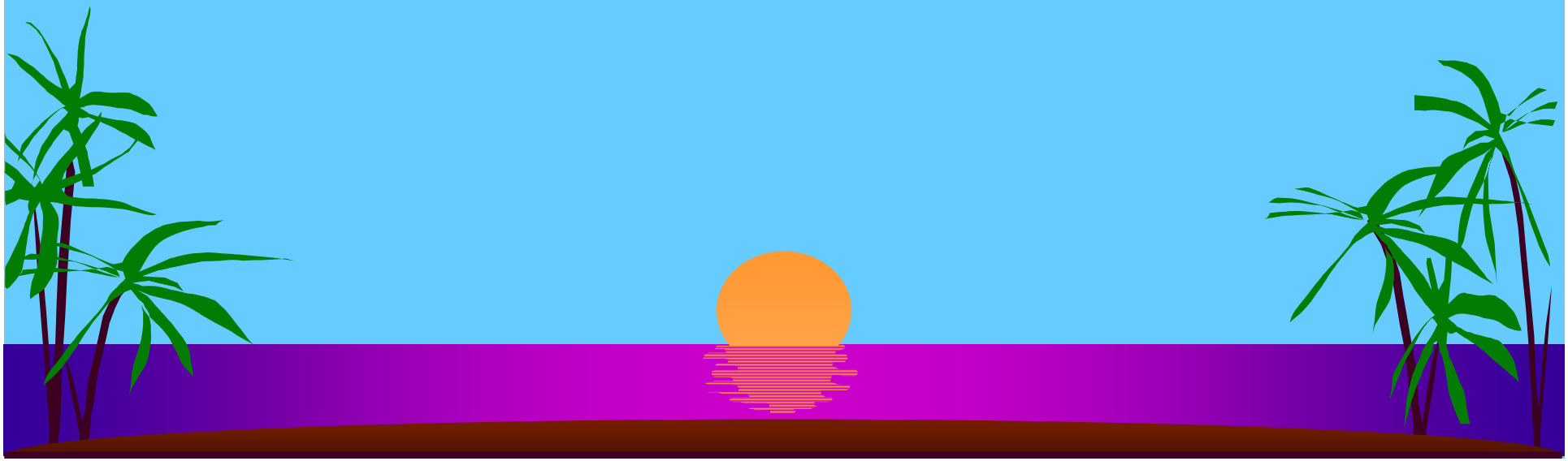
7. 结论：

单一总体假设检验总结



第十三章　简单回归和相关分析

研究两个变量之间的关系



本章重点

- 什么是线性回归模型
- 建立线性回归模型的步骤
- 解释最小平方法
- 计算回归系数
- 样本回归方程在统计推断中的作用
- 如何衡量变量之间关系的密切程度

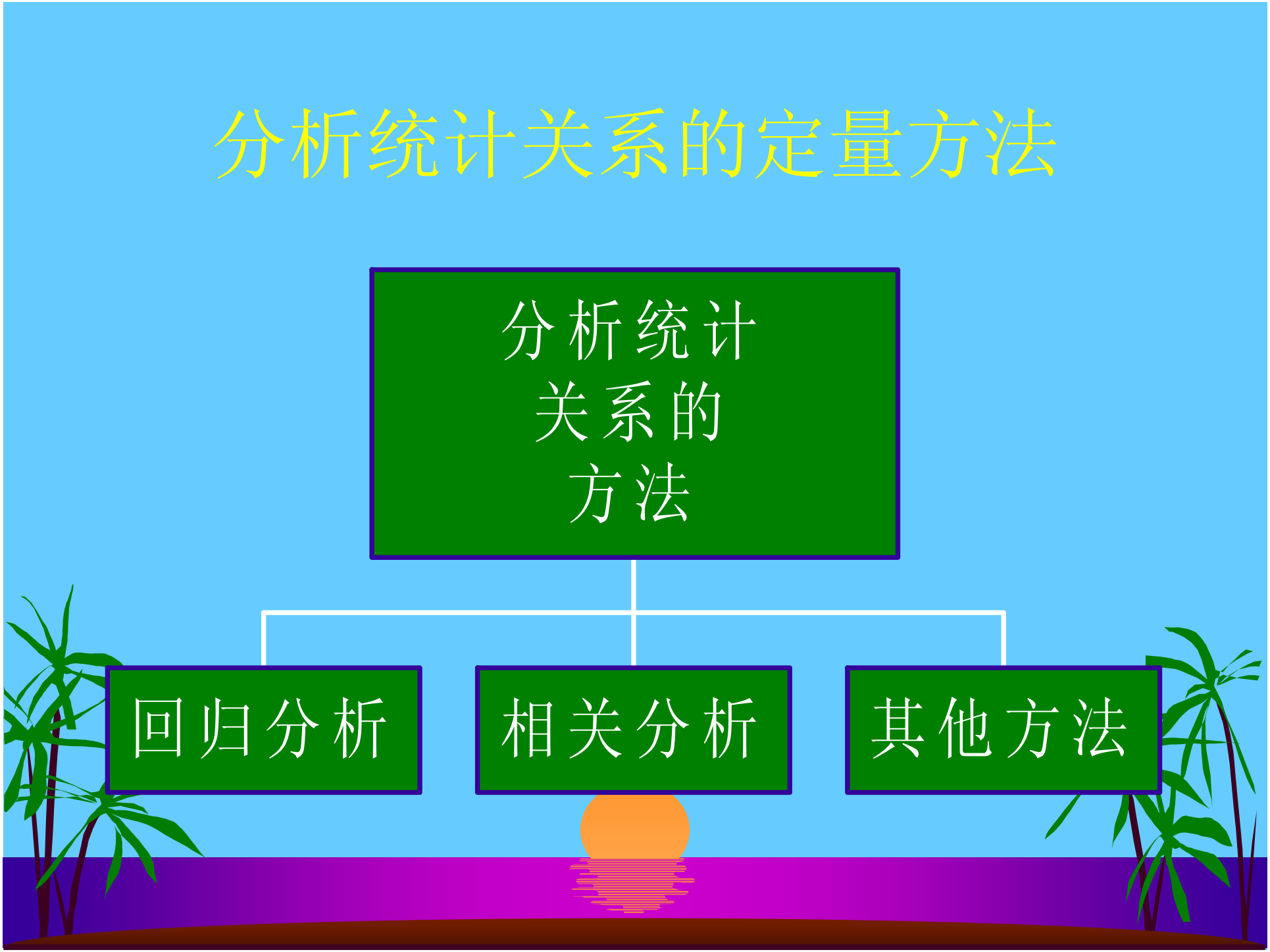
函数关系和统计关系

- 函数关系：
  两变量的数量表现在一定条件下是完全确定的。
  如： 圆的面积和半径的关系 $S = \pi * r^2$
- 统计关系（相关关系）： 两变量的数量表现尽管存在着密切关系，但却不是完全确定的。
  如：成本和利润的关系

统计关系的种类

- 按涉及变量的多少可分
  - 简单相关回归关系（一个自变量和一个因变量）
  - 复相关复回归关系（一个因变量和多个自变量）
- 按变量关系在图形上的形态可分
  - 线性相关回归
  - 非 线性相关回归
- 按 两变量变动的方向可分
  - 正相关回归
  - 负相关回归

分析统计关系的定量方法



1996年12个沿海省、直辖市、自治区
大型零售、批发贸易业企业利润额与销售额

单位：亿元

省、市、区	销售总额	利润总额
北　京	147	71
天　津	64	20
河　北	87	40
辽　宁	108	59
上　海	206	120
江　苏	277	122
浙　江	209	88
福　建	64	29
山　东	173	91
广　东	214	105
广　西	44	19
海　南	53	24

贸易业企业利润额与销售额
相关与回归分析数据计算

单位：亿元

省、市、区	销售额 （X）	利润额（Y）	X^2	Y^2	XY
北　京	147	71	21609	5041	10437
天　津	64	20	4096	400	1280
河　北	87	40	7569	1600	3480
辽　宁	108	59	11664	3481	6372
上　海	206	120	42436	14400	24720
江　苏	277	122	76729	14884	33794
浙　江	209	88	43681	7744	18392
福　建	64	29	4096	841	1856
山　东	173	91	29929	8281	15743
广　东	214	105	45796	11025	22470
广　西	44	19	1936	361	836
海　南	53	24	2809	576	1272
合　计	1646	788	292350	68634	140652

12省市自治区销售额与利润额的相关

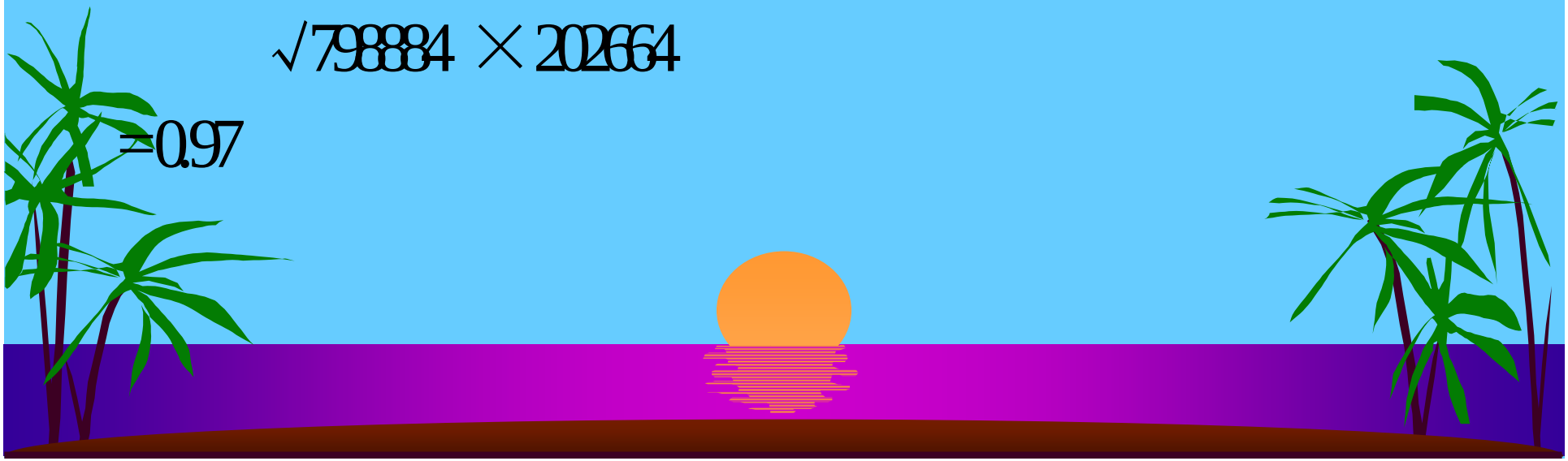
$$n\Sigma XY-(\Sigma X)(\Sigma Y)=390776$$

$$n\Sigma X^2-(\Sigma X)^2=798884$$

$$n\Sigma Y^2-(\Sigma Y)^2=202664$$

$$r=\frac{390776}{\sqrt{798884\times 202664}}$$

$$=0.97$$

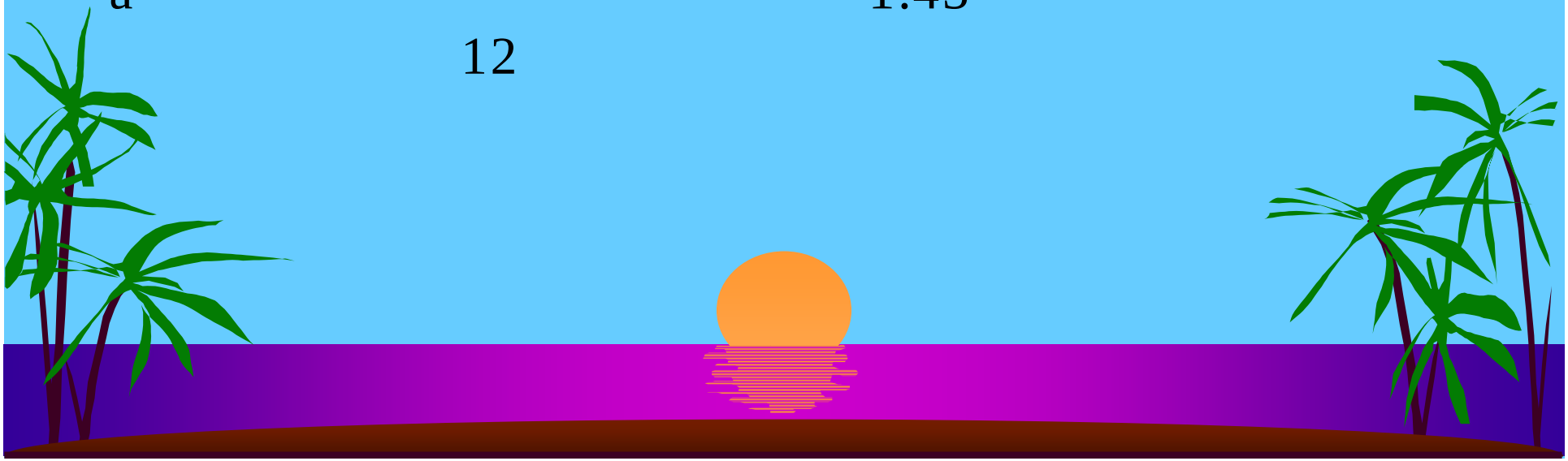


12省市自治区销售额与利润额的回归

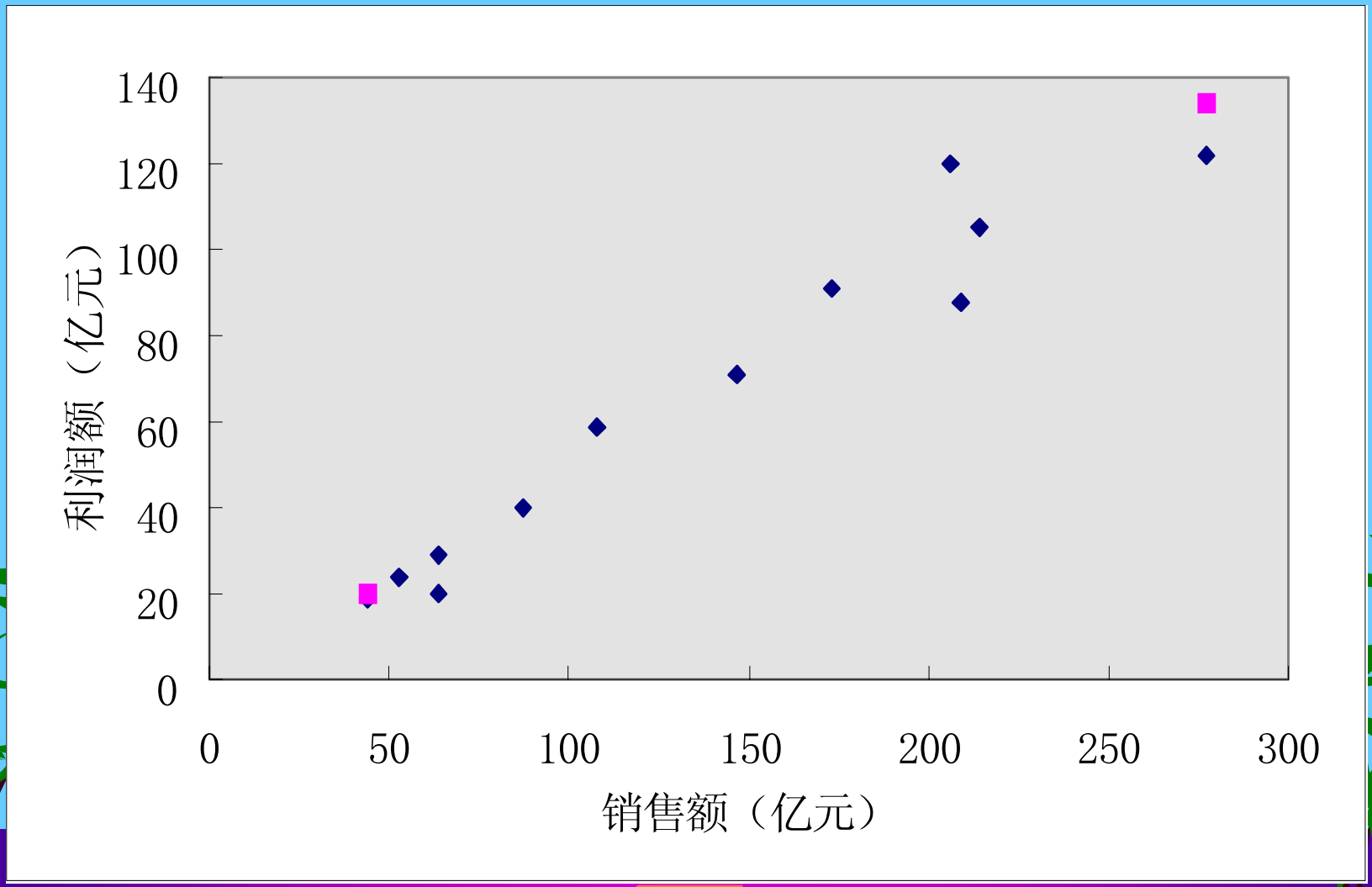
回归系数的计算：

$$b = \frac{390776}{798884} = 0.49$$

$$a = \frac{788 - 0.49 \times 1646}{12} = -1.43$$



12省利润额对销售额的散点图及回归



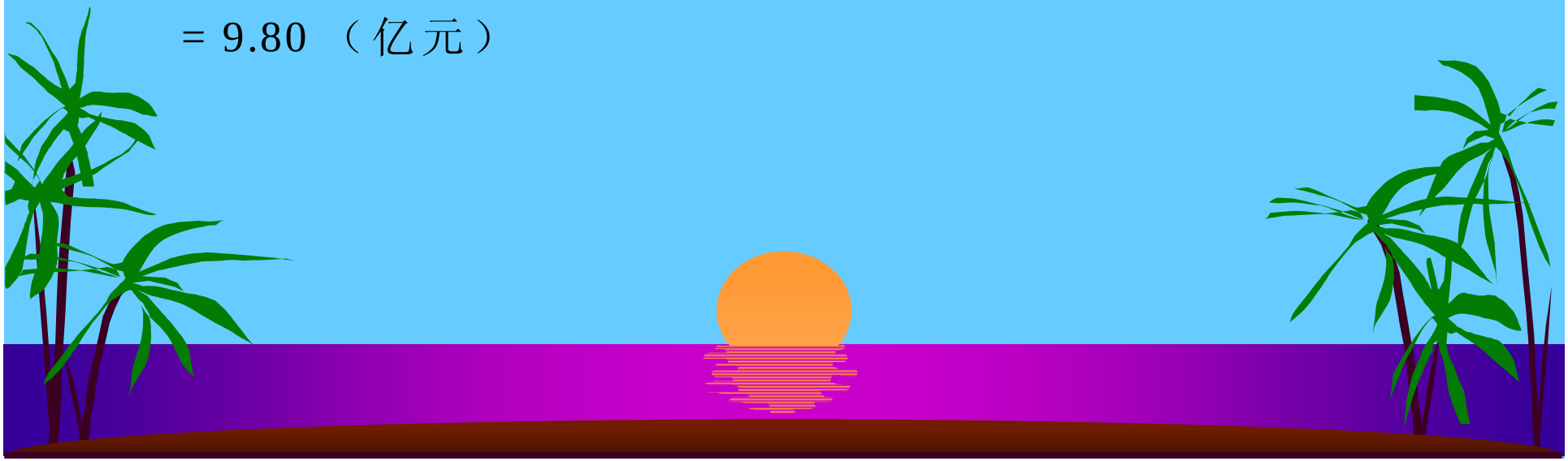
销售额为 200 亿元时利润额的平均值

$$Y_c = -1.43 + 0.49 \times 200 = 96.57 \text{（亿元）}$$

估计标准误

$$S_{Y.X} = \frac{68634 - (-1.43) \times 788 - 0.49 \times 140653}{12 - 2}$$

$$= 9.80 \text{（亿元）}$$

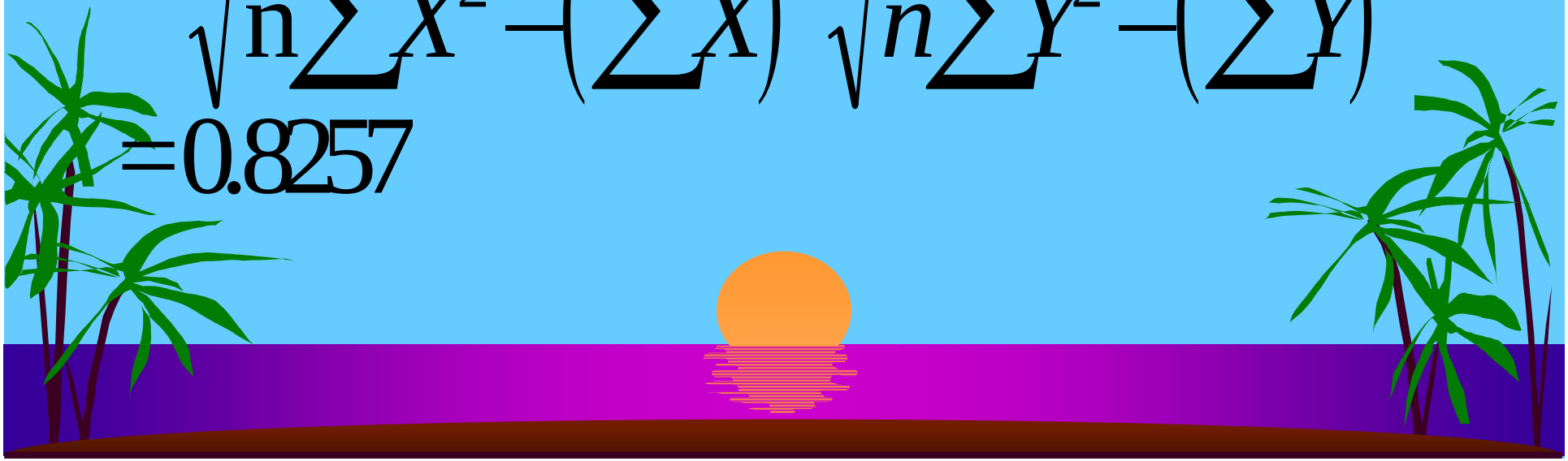


相关系数的计算

$$r=\sqrt{r^2}$$

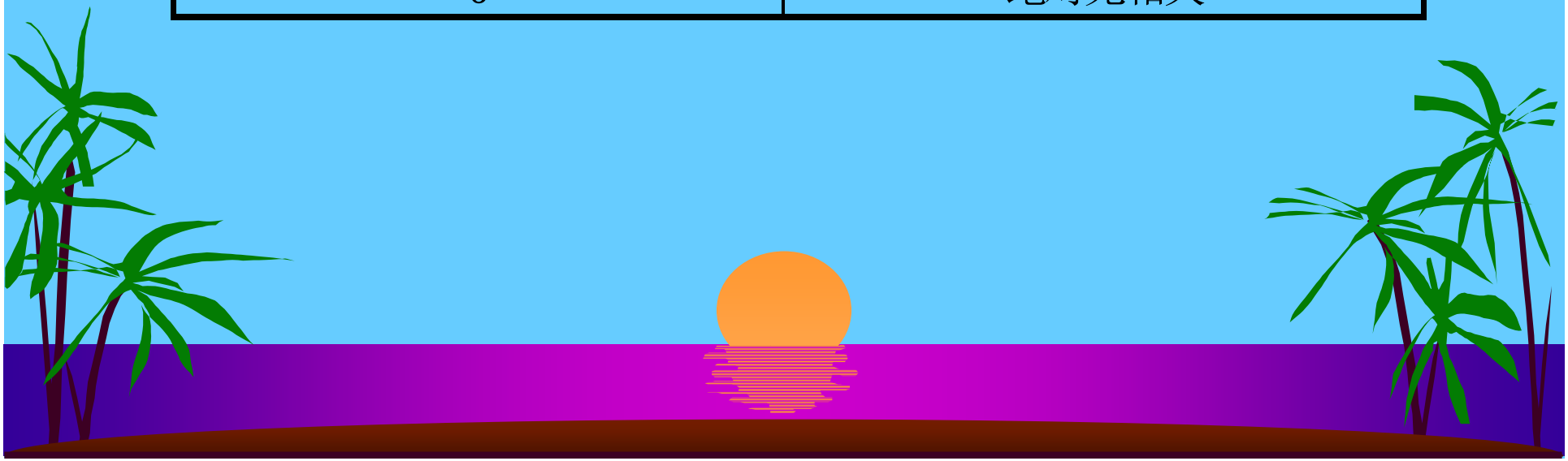
$$=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\sum X^2-\left(\sum X\right)^2}\sqrt{n\sum Y^2-\left(\sum Y\right)^2}}$$

$$=0.8257$$



相关系数对样本相关关系的计量

｜r｜值	相关程度
1	绝对相关
0.8～1	高度相关
0.5～0.8	中度相关
0.3～0.5	低度相关
0～0.3	无相关
0	绝对无相关

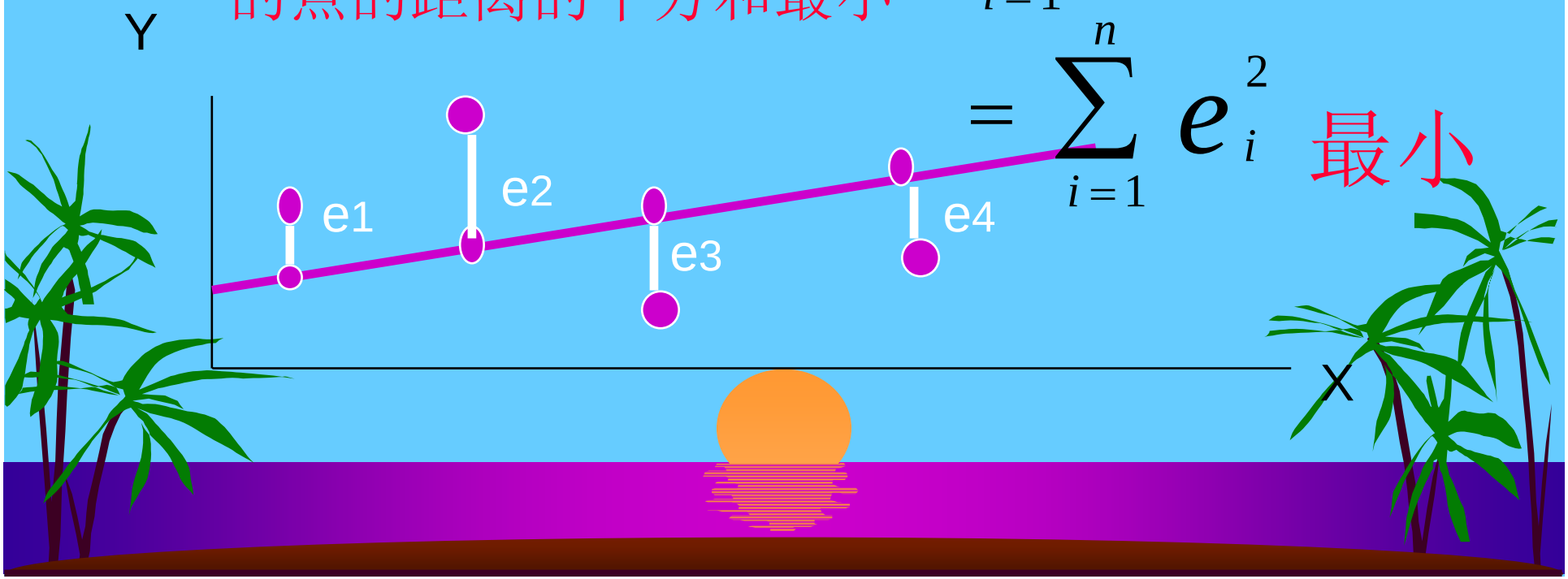


建立 样本线性回归模型的方法----最小平方法

实际观察值与样本回归线上的点的距离的平方和最小

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

$$=\sum_{i=1}^{n}e_i^2$$ 最小



样本回归系数的计算公式

$$\hat{y} = b_0 + b_1 X$$

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$$

线性回归分析

目的；

在因变量和自变量之间建立一个数学模型，根据这个模型可以根据自变量的变动预测因变量的变动。

应注意的问题：

1. 建立模型的目的
2. 谁将用这个模型
3. 建立 模型用的资料是否合适
4. 如何利用模型

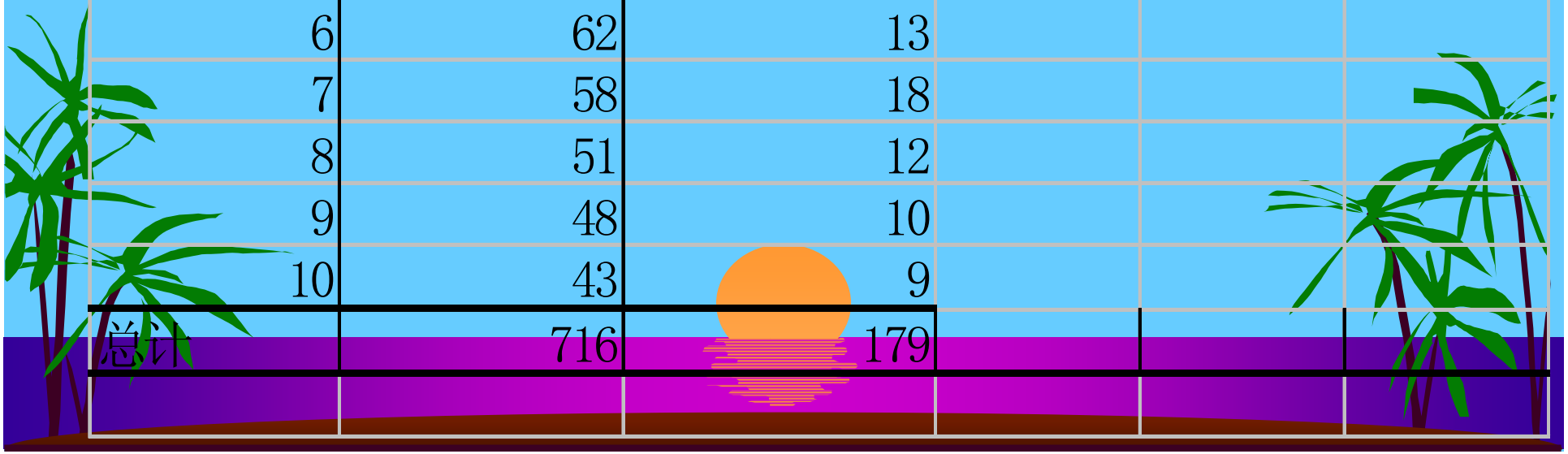
建立 样本线性回归模型的实际例子1

现有10个企业的销售额和利润的资料

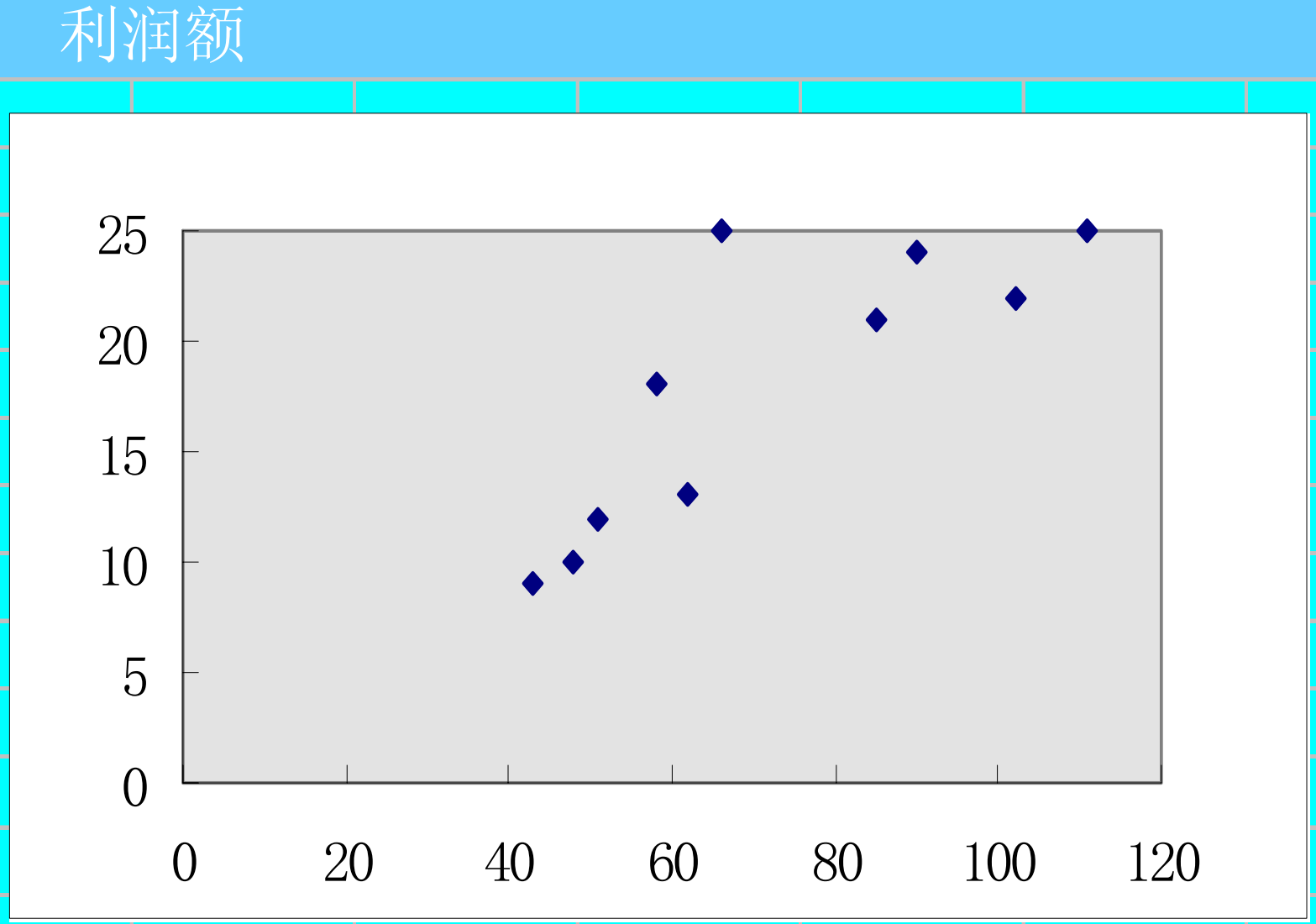
序号	销售额	利润额
1	111	25
2	102	22
3	90	24
4	85	21
5	66	25
6	62	13
7	58	18
8	51	12
9	48	10
10	43	9
总计	716	179

问：

利润额和销售额之间存在什么样的关系



销售额和利润额的散点图



实际例子的计算1

序号	销售额	利润额	x^2	xy	y^2
1	111	25	12321	2775	625
2	102	22	10404	2244	484
3	90	24	8100	2160	576
4	85	21	7225	1785	441
5	66	25	4356	1650	625
6	62	13	3844	806	169
7	58	18	3364	1044	324
8	51	12	2601	612	144
9	48	10	2304	480	100
10	43	9	1849	387	81
总计	716	179	56368	13943	3569

实际例子的计算2

$$\hat{y} = b_0 + b_1 X$$

$$b_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{10*13943 - 716*179}{10*56368 - 716^2} = 0.22$$

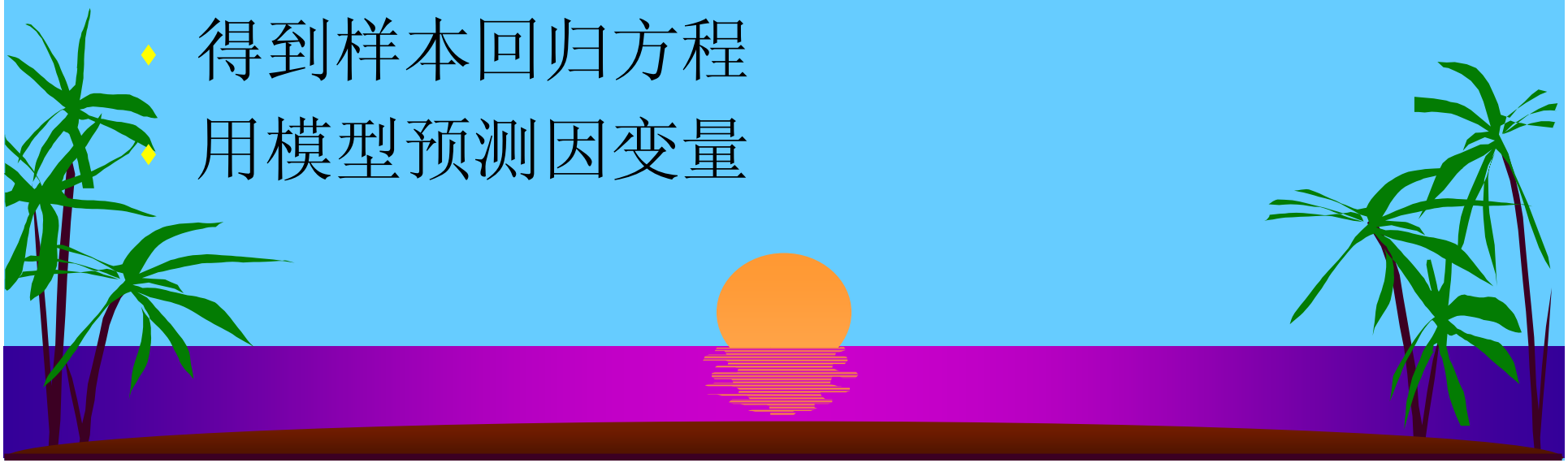
$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \frac{179}{10} - 0.22*\frac{716}{10} = 2.15$$

$$\hat{Y} = 2.15 + 0.22X$$

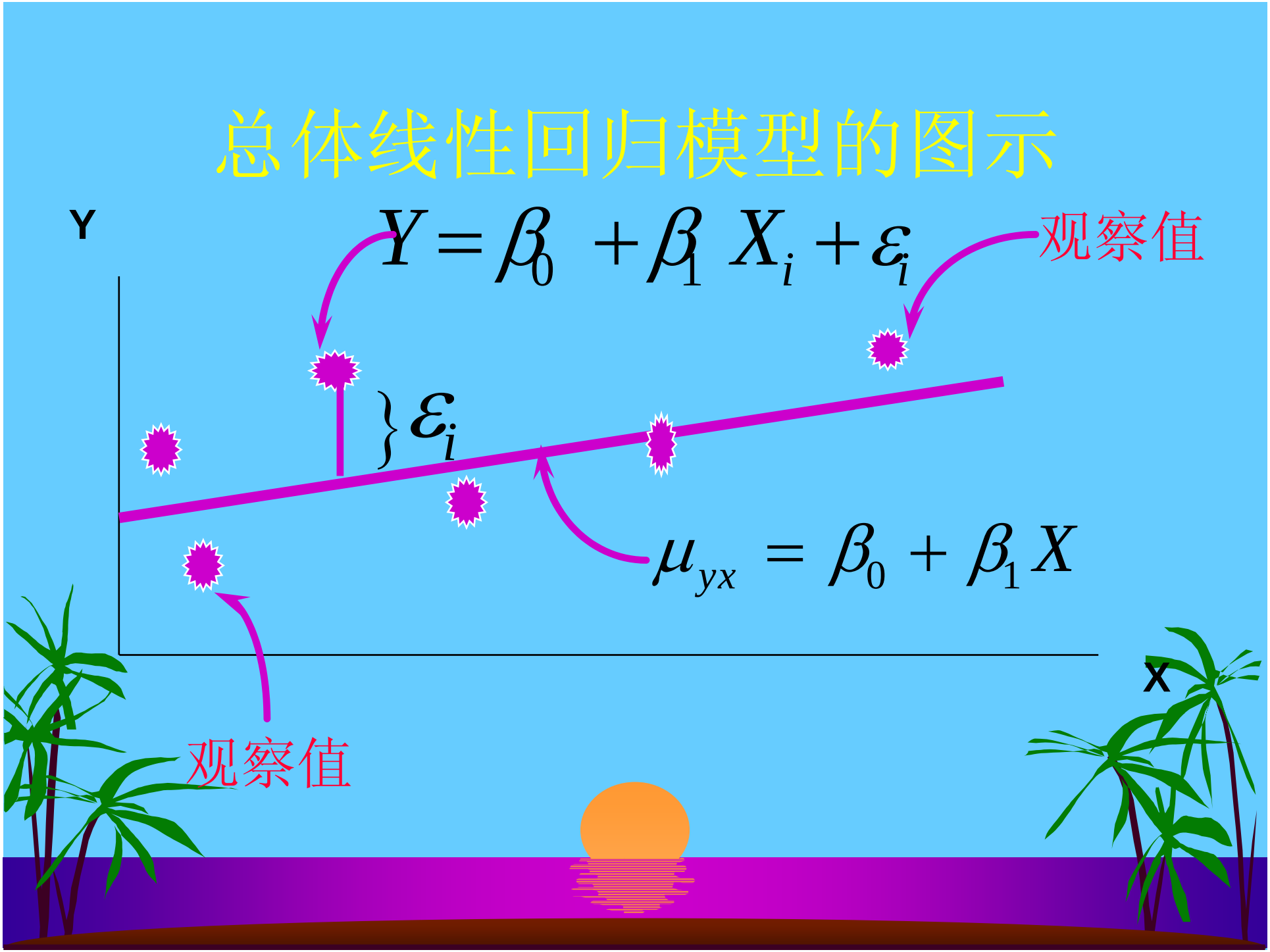
表示当销售额增加或减少1亿元时，利润额平均增加或减少0.22亿元

建立 线性回归模型的步骤

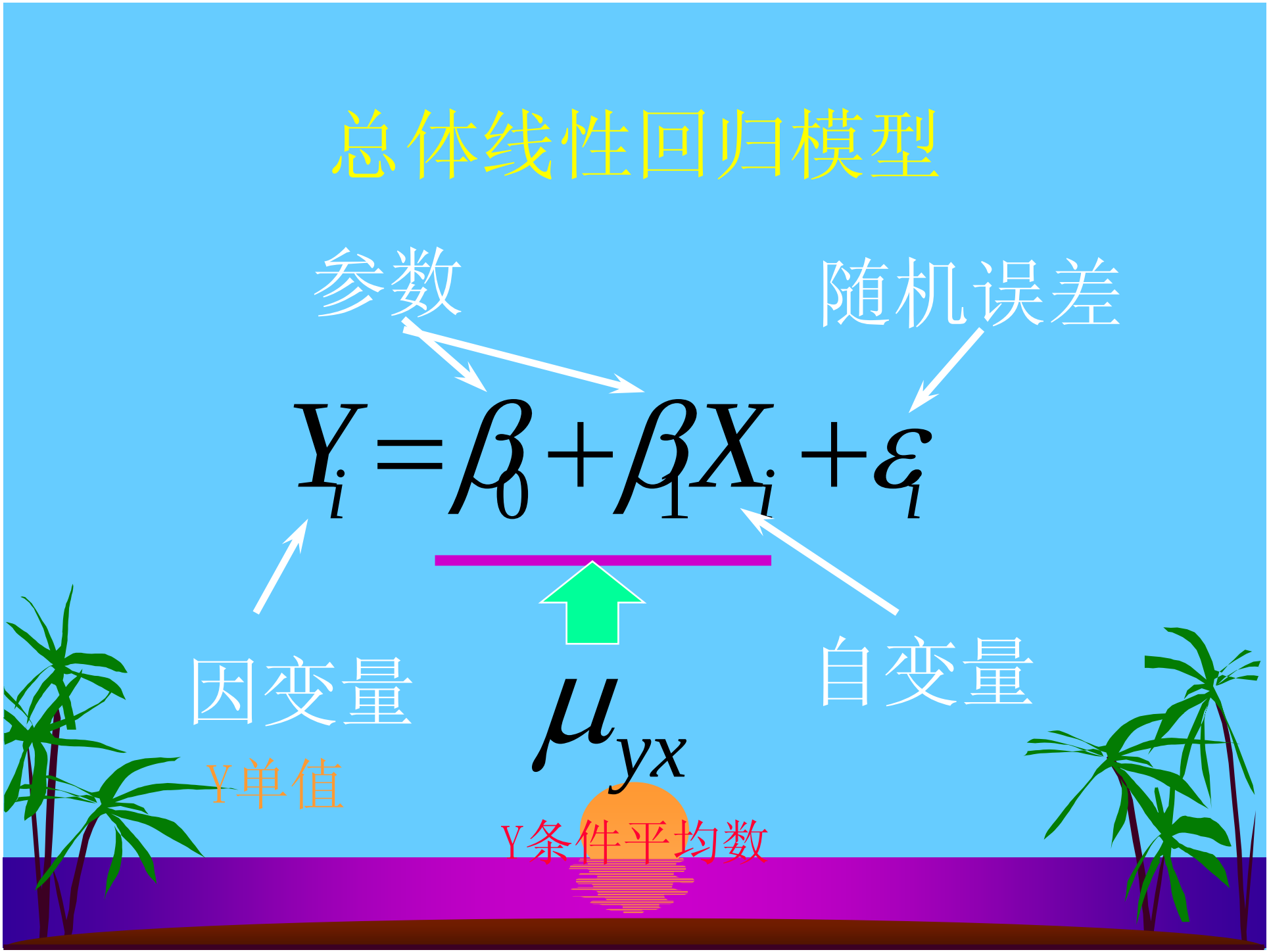
- 确定研究的问题
- 设样本回归模型(如: $\hat{Y} = a + bx$)
- 搜集样本资料(数据资料)
- 估计未知参数(计算统计量)
- 得到样本回归方程
- 用模型预测因变量



总体线性回归模型的图示

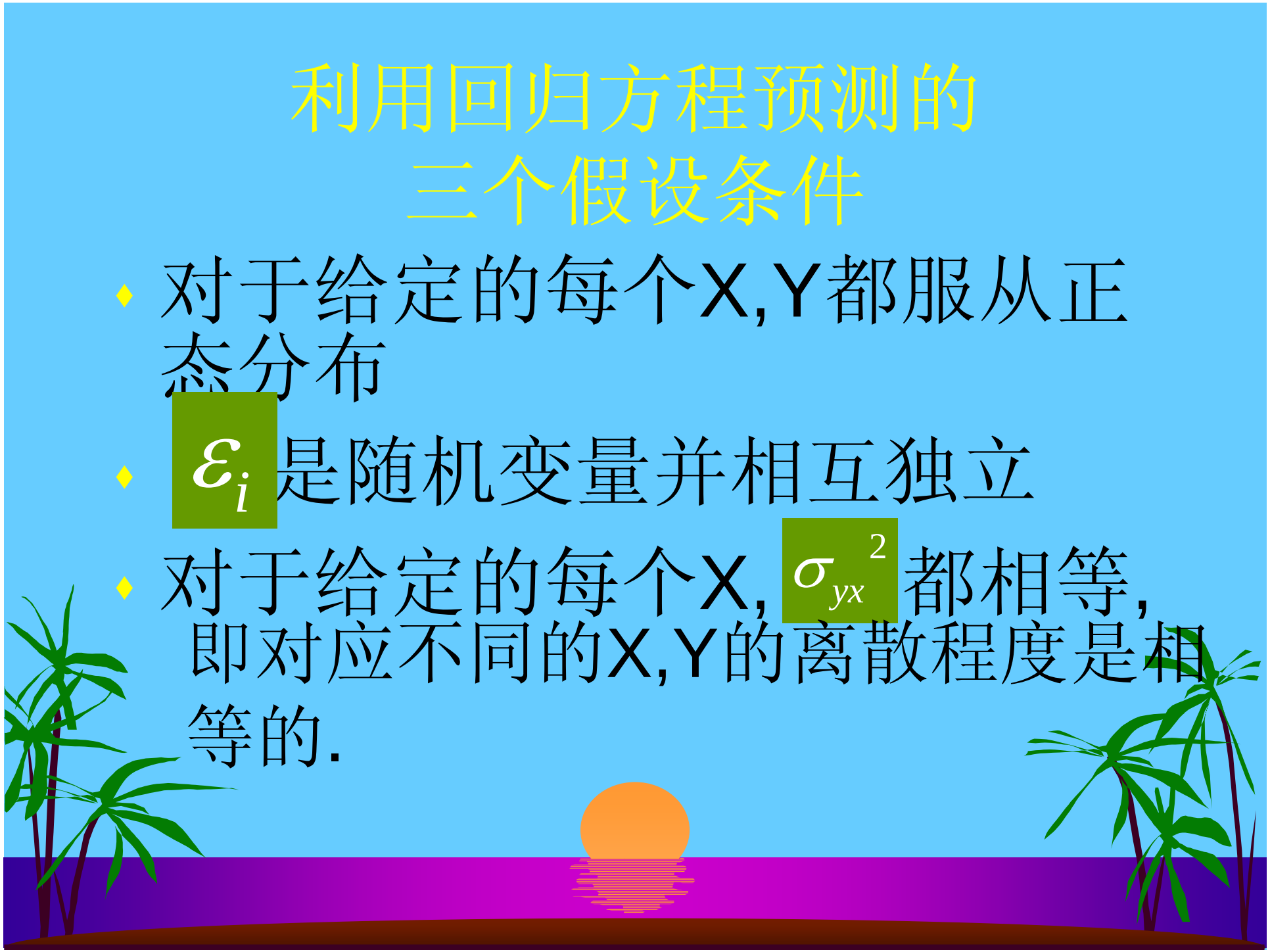


总体线性回归模型


参数
随机误差
$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$
因变量
自变量
μ_{yx}
Y单值
Y条件平均数

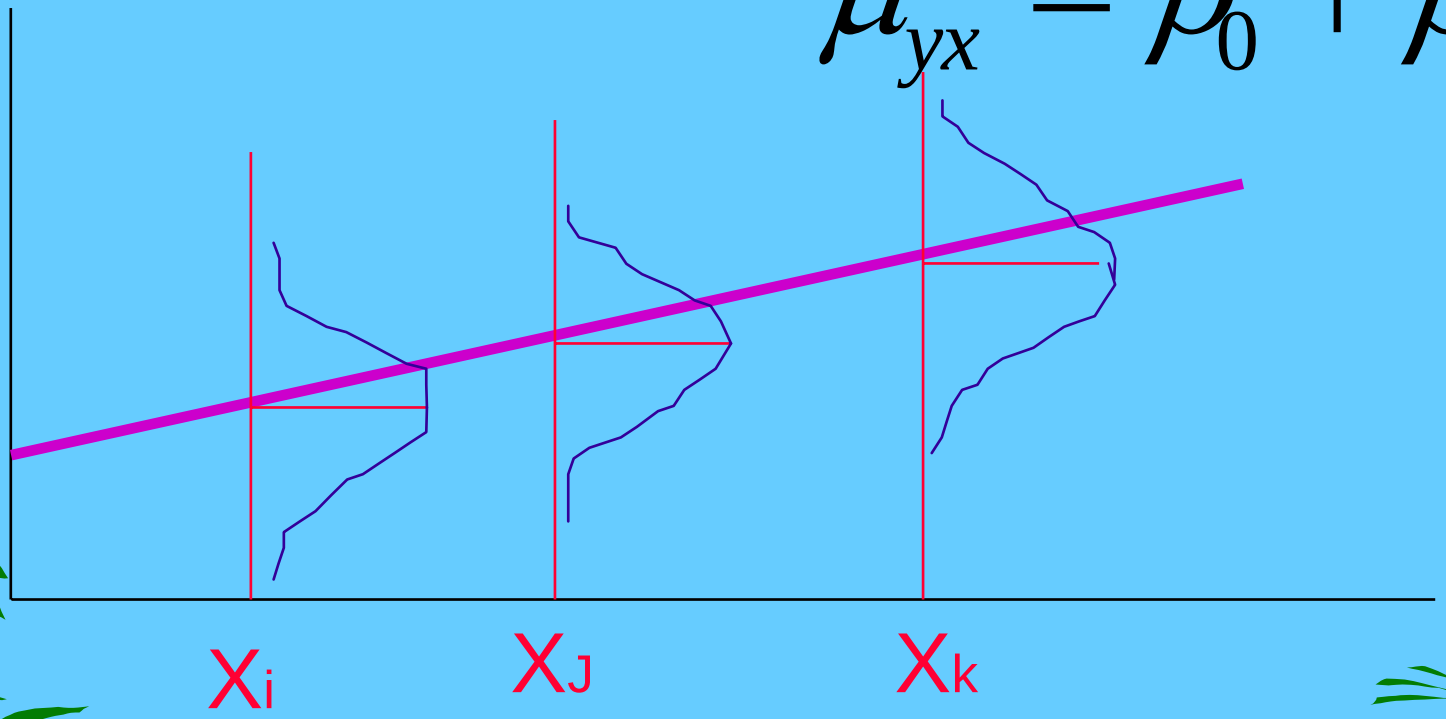
利用回归方程预测的三个假设条件

- 对于给定的每个X,Y都服从正态分布
- ε_i 是随机变量并相互独立
- 对于给定的每个X, σ_{yx}^2 都相等，即对应不同的X,Y的离散程度是相等的.



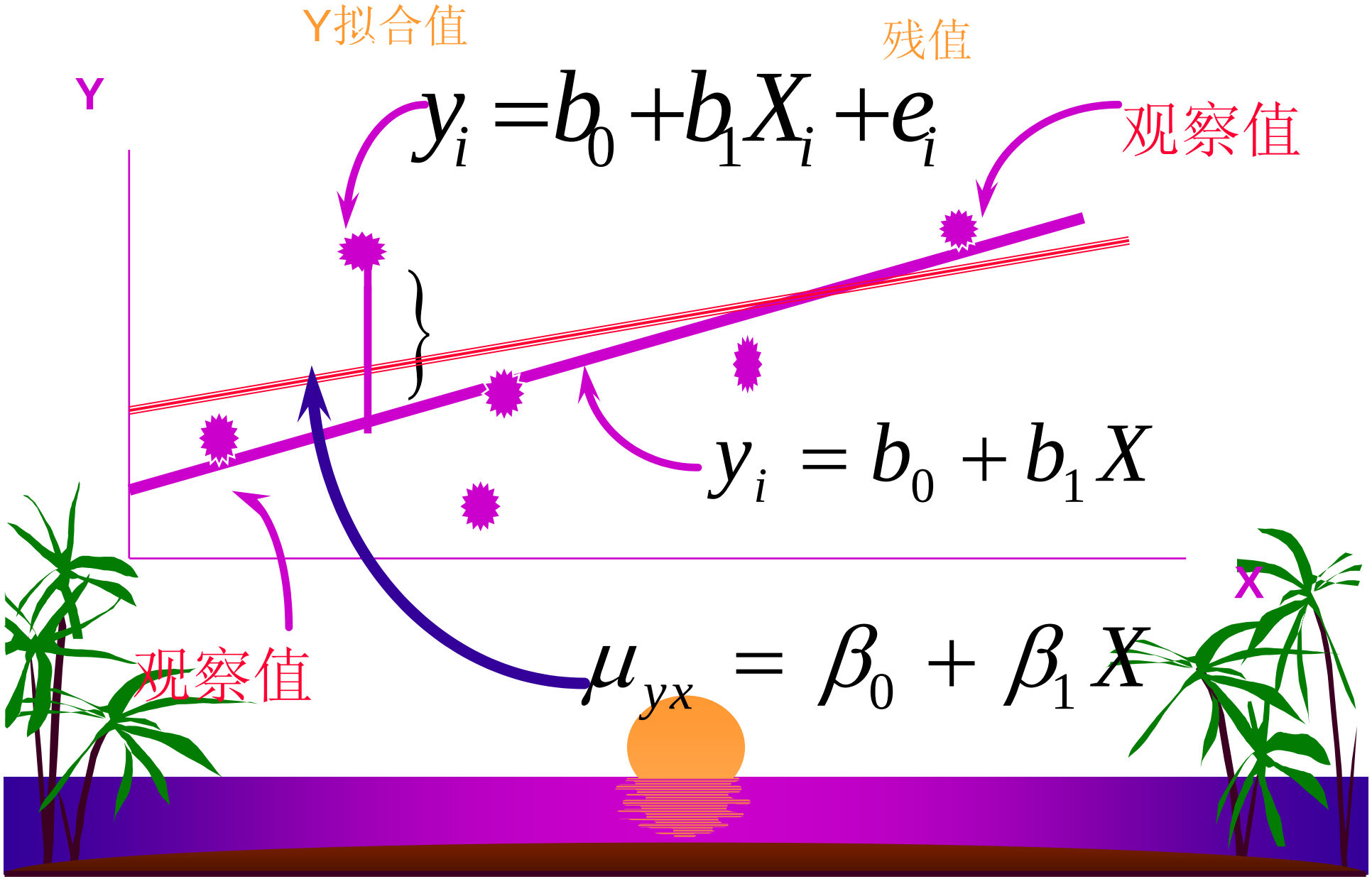
三个假设条件的图示

$$\mu_{yx} = \beta_0 + \beta_1 X$$



$$\sigma_{yx_i} = \sigma_{yx_j} = \sigma_{yx_k}$$

总体回归模型与样本回归方程



估计标准误差

- 估计标准误差: 实际观察值Y与 Y的平均离差
- 它可用来估计Y值围绕总体回归线的离散 程度

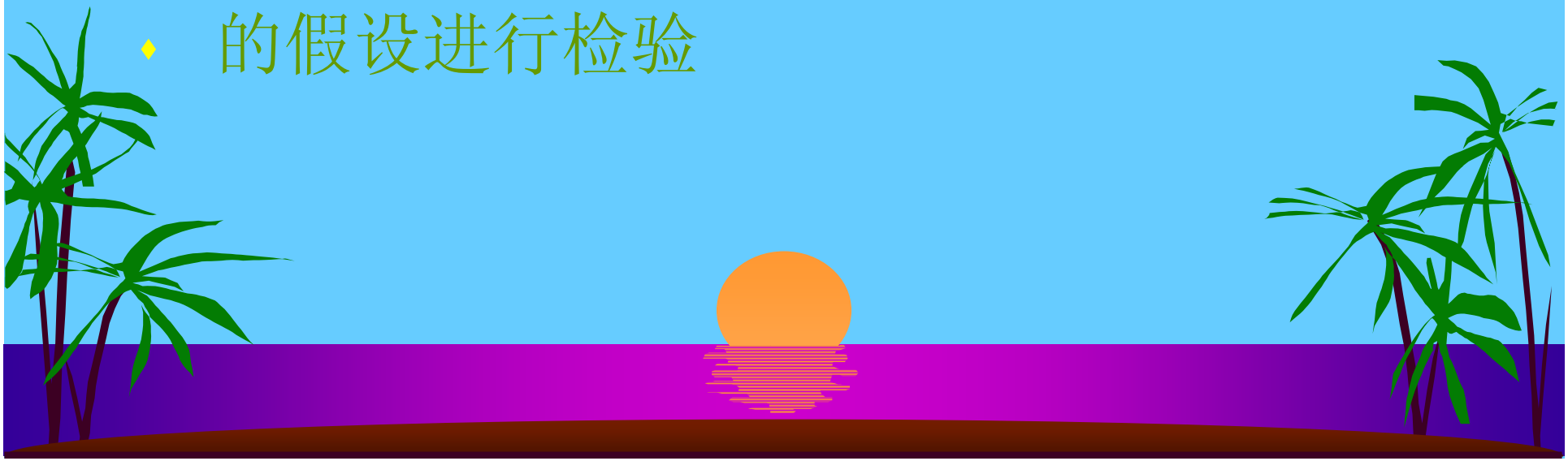
$$\sigma_{yx} = \sqrt{\frac{\sum_{1}^{N} \varepsilon_i^2}{N}} = \sqrt{\frac{\sum \left(Y_i - \mu_{yx}\right)^2}{N}}$$

$$S_{yx} = \sqrt{\frac{\sum \left(y_i - \hat{y}\right)^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

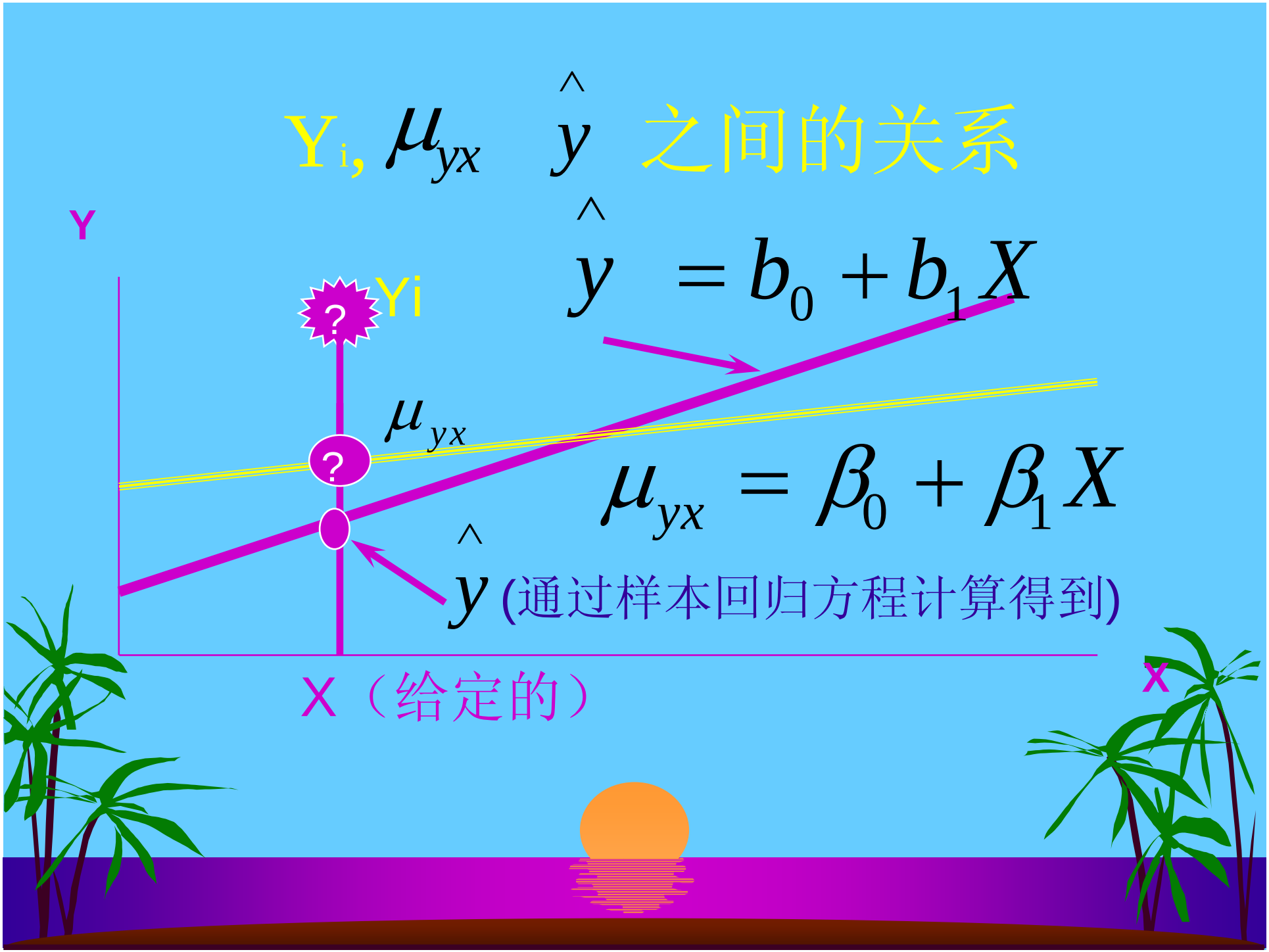
$$= \sqrt{\frac{\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i}{n-2}}$$

利用回归方程对总体进行推断

- 对给定的X，求 μ_{yx} 的置信区间
- 对给定的X，求单个 Y_i 的置信区间
- 求 β_1 的置信区间
- 根据样本回归方程对 $\beta_1 = 0$
- 的假设进行检验



Y_i, μ_{yx} $\hat{y}$ 之间的关系



对给定的X，求 μ_{yx} 的置信区间

$$\hat{y} - t_{n-2}\sigma_{\hat{y}} \le \mu_{yx} \le \hat{y} + t_{n-2}\sigma_{\hat{y}}$$

$$\sigma_{\hat{y}} = \sigma_{yx}\sqrt{\frac{1}{n} + \frac{\sum\left(X_i - \overline{X}\right)^2}{\sum X^2 - n\overline{X}^2}}$$

$$S_{\hat{y}} = S_{yx}\sqrt{\frac{1}{n} + \frac{\sum\left(X_i - \overline{X}\right)^2}{\sum X^2 - n\overline{X}^2}}$$

$$\hat{y} - t_{n-2}S_{\hat{y}} \le \mu_{yx} \le \hat{y} + t_{n-2}S_{\hat{y}}$$

μ_{yx}的置信区间

$$X_0 = 66, \alpha = 0.05 \quad S_{yx} = 3.81,$$

$$\hat{y}_{(66)} = 2.15 + 0.22 \times 66 = 15.35$$

$$S_{\hat{y}} = 3.18\sqrt{\frac{1}{10} + \frac{(66 - 71.6)^2}{56368 - 10 \times 71.6^2}} = 1.2413$$

μ_{yx}的置信区间为

$15.35 \pm t_8 1.2413$　　$15.35 \pm 2.306 * 1.2431$

(12.49　18.21)

请解释结果：

Y_i 的推算区间

$$\hat{Y} - t_{n-2}\sigma_{y_i} \le Y_i \le \hat{Y} + t_{n-2}\sigma_{y_i}$$

$$\sigma_{y_i} = \sigma_{yx}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum X^2 - n\overline{X}^2}}$$

$$S_{y_i} = S_{yx}\sqrt{1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum X^2 - n\overline{X}^2}}$$

Y_i的区间为 $\hat{Y} - t_{n-2}S_{y_i} \le Y_i \le \hat{Y} + t_{n-2}S_{y_i}$

Y_i 的推算区间

$$X_0 = 66, \alpha = 0.05 \quad S_{yx} = 3.81,$$

$$\hat{y}_{(66)} = 2.15 + 0.22 \times 66 = 15.35$$

$$S_{y_i} = 3.18\sqrt{\frac{1}{10} + 1 + \frac{(66 - 71.6)^2}{56368 - 10 \times 71.6^2}} = 4.0071$$

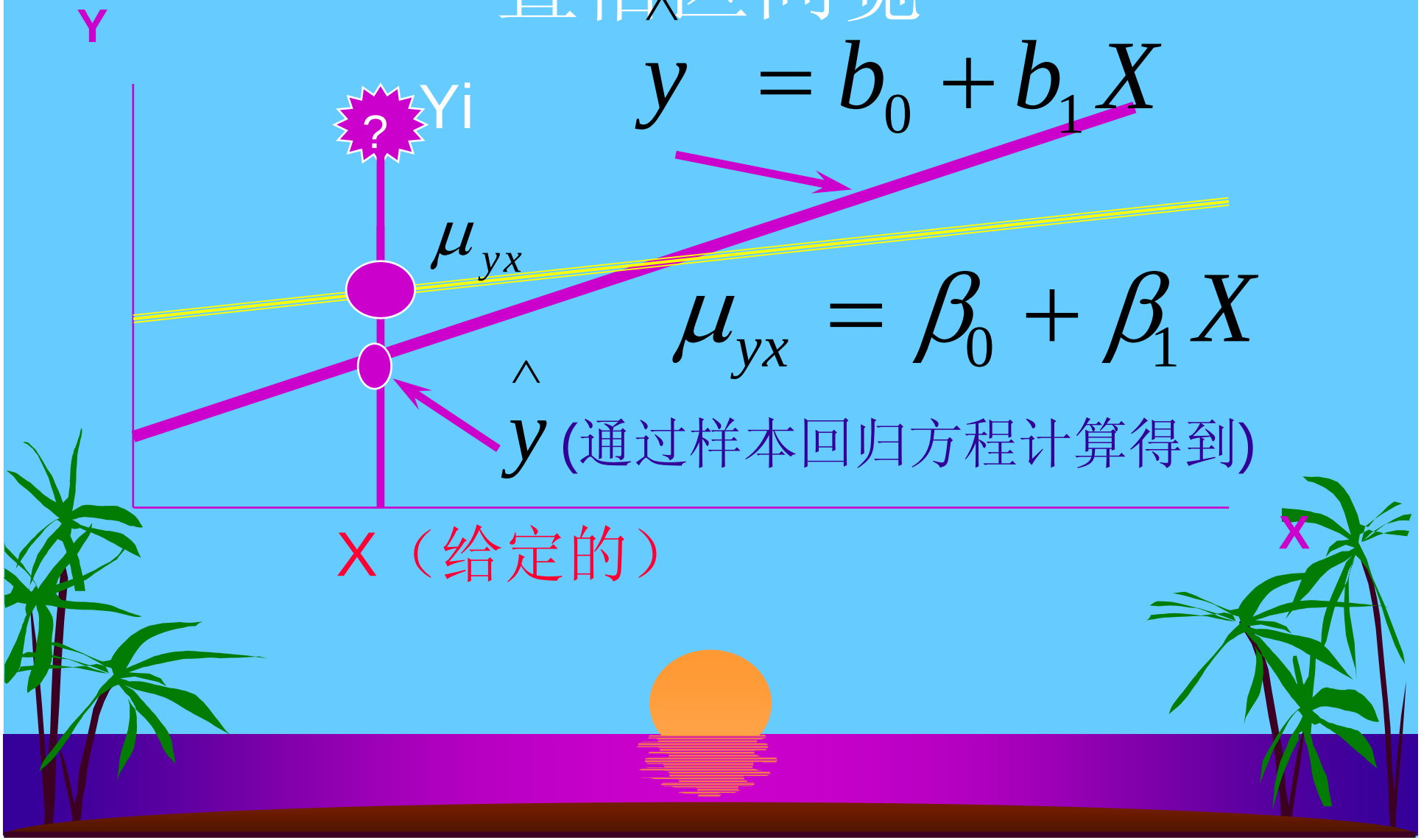
Y_i的置信区间为

$15.35 \pm t_8 4.0071 \qquad 15.35 \pm 2.306 * 4.0071$

(6.11　　24.54)

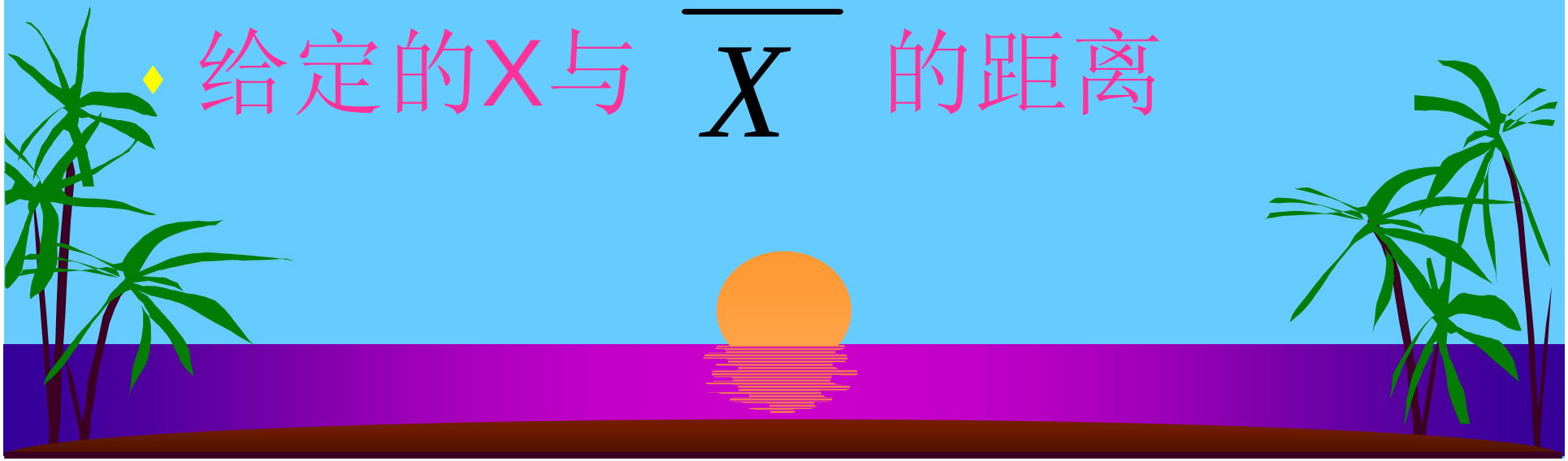
请解释结果：

为什么 Y_i 的置信区间比 μ_{yx} 的置信区间宽



影响区间宽度的因素

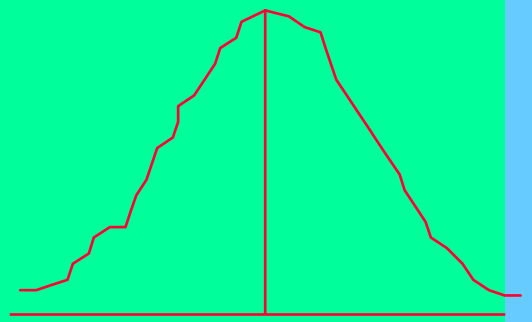
- 置信系数
- Y的变异程度
- 样本容量的大小
- 给定的X与 $\overline{X}$ 的距离



对总体回归系数的假设检验

1. $H_0: \beta_1 = 0$
   $H_1: \beta_1 \neq 0$
2. $\alpha \Rightarrow t_{n-2}$
3. 计算统计量 $t = \dfrac{b_1 - \beta_1}{s_{b_1}} = \dfrac{b_1}{s_{b_1}}$

$$s_{b_1} = \frac{s_{yx}}{\sqrt{\sum X^2 - n\overline{X}^2}}$$



4. 如果 $-t_{n-2} \leq t \leq t_{n-2}$ 则接收 H_0，否则拒绝 H_0

对总体回归系数的假设检验的例子

$\hat{Y} = 2.15 + 0.22X$

1. $H_0: \beta_1 = 0$　$H_1: \beta_1 \neq 0$

2. $\alpha = 0.05 \Rightarrow t_8 = 2.306$

3. $s_{b_1} = \dfrac{3.81}{\sqrt{56368 - 10*71.6^2}} = 0.053$

$$t = \frac{b_1}{s_{b_1}} = \frac{0.22}{0.053} = 4.15$$

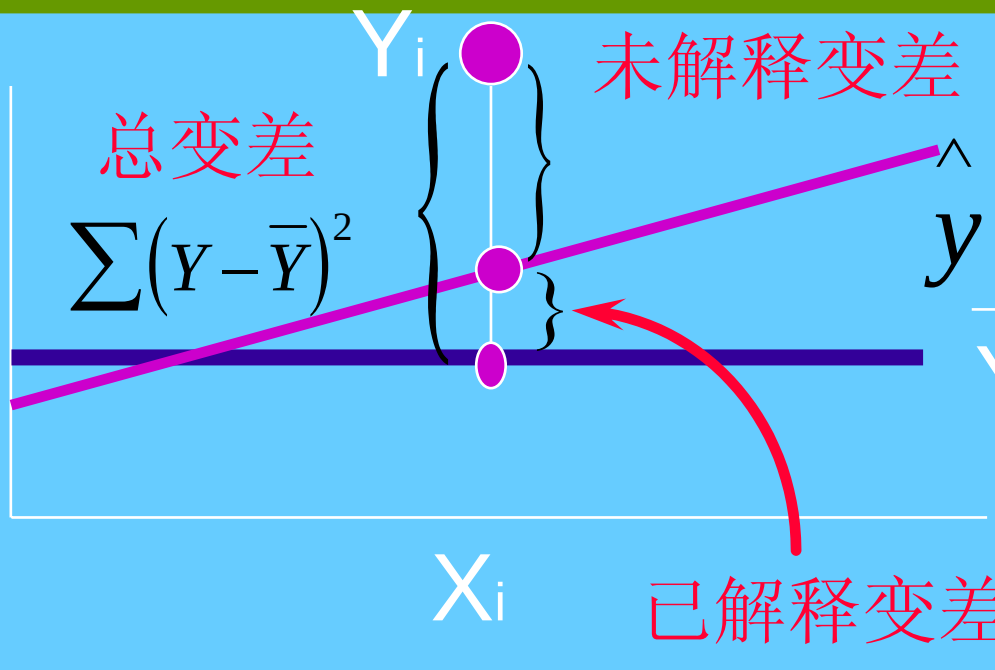
4. t = 4.15〉2.306　∴ 拒绝H_0

5. 结论：样本回归方程能代表总体回归方程.
也就是说从总体上X与Y之存在线性关系，
可以根据样本回归方程通过X预测Y.

可决系数

- 作用：衡量回归对Y变异的解释程度。
- 总变差=已解释变差+未解释变差。
- $$\text{可决系数} = \frac{\text{已解释变差}}{\text{总变差}}$$
- 经调整的可决系数

总变差，已解释变差，未解释变差的关系



$$\sum\left(Y-\hat{Y}\right)^2$$

$$\sum\left(\hat{Y}-\bar{Y}\right)^2$$

SST = SSE + SSR

$$\sum(Y-\bar{Y})^2=\sum\left(Y-\hat{Y}\right)^2+\sum\left(\hat{Y}-\bar{Y}\right)^2$$

可决系数

- 定义：已解释变差与总变差的比值，在估计Yi时，在总变差中可被X解释的比率，它越大，拟合回归方程的解释作用越强。
- 公式：样本可决系数

$$r^2 = \frac{\sum\left(\hat{Y}-\bar{Y}\right)^2}{\sum\left(Y-\bar{Y}\right)^2} = 1-\frac{\sum\left(Y-\hat{Y}\right)^2}{\sum\left(Y-\bar{Y}\right)^2} =$$

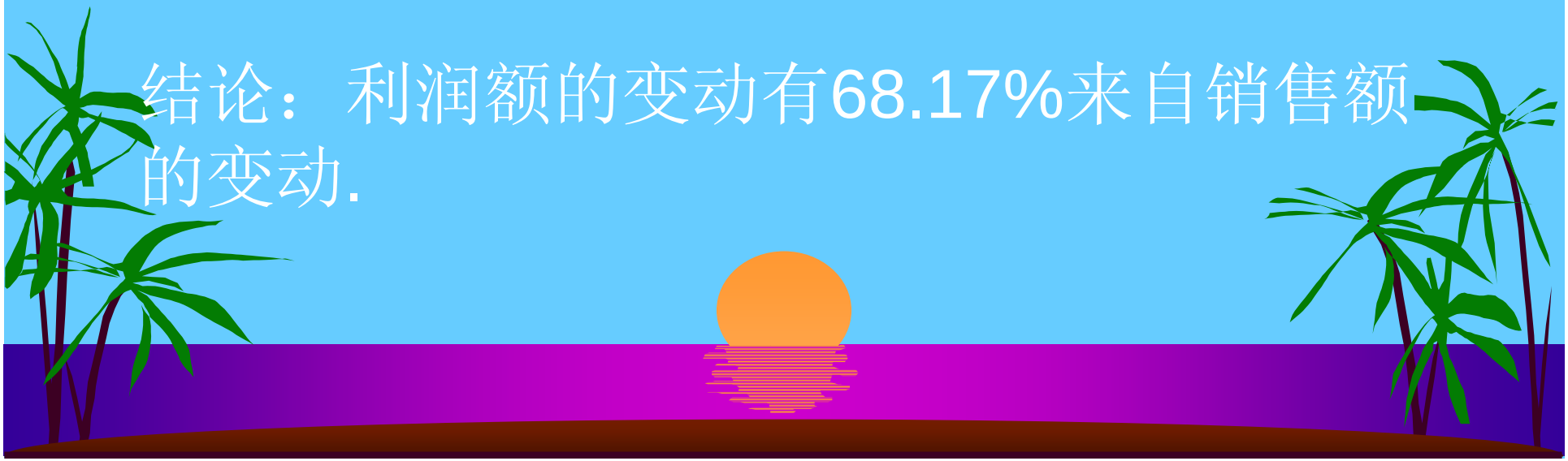
$$\frac{b_0\sum Y + b_1\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}$$

可决系数的例题

$$r^2 = \frac{2.15*179+0.22*13943-10*17.9^2}{3569-10*17.9^2}$$

$$=0.6817$$

结论：利润额的变动有68.17%来自销售额的变动.



相关系数---- 可决系数的平方根

$$r=\sqrt{r^2}$$

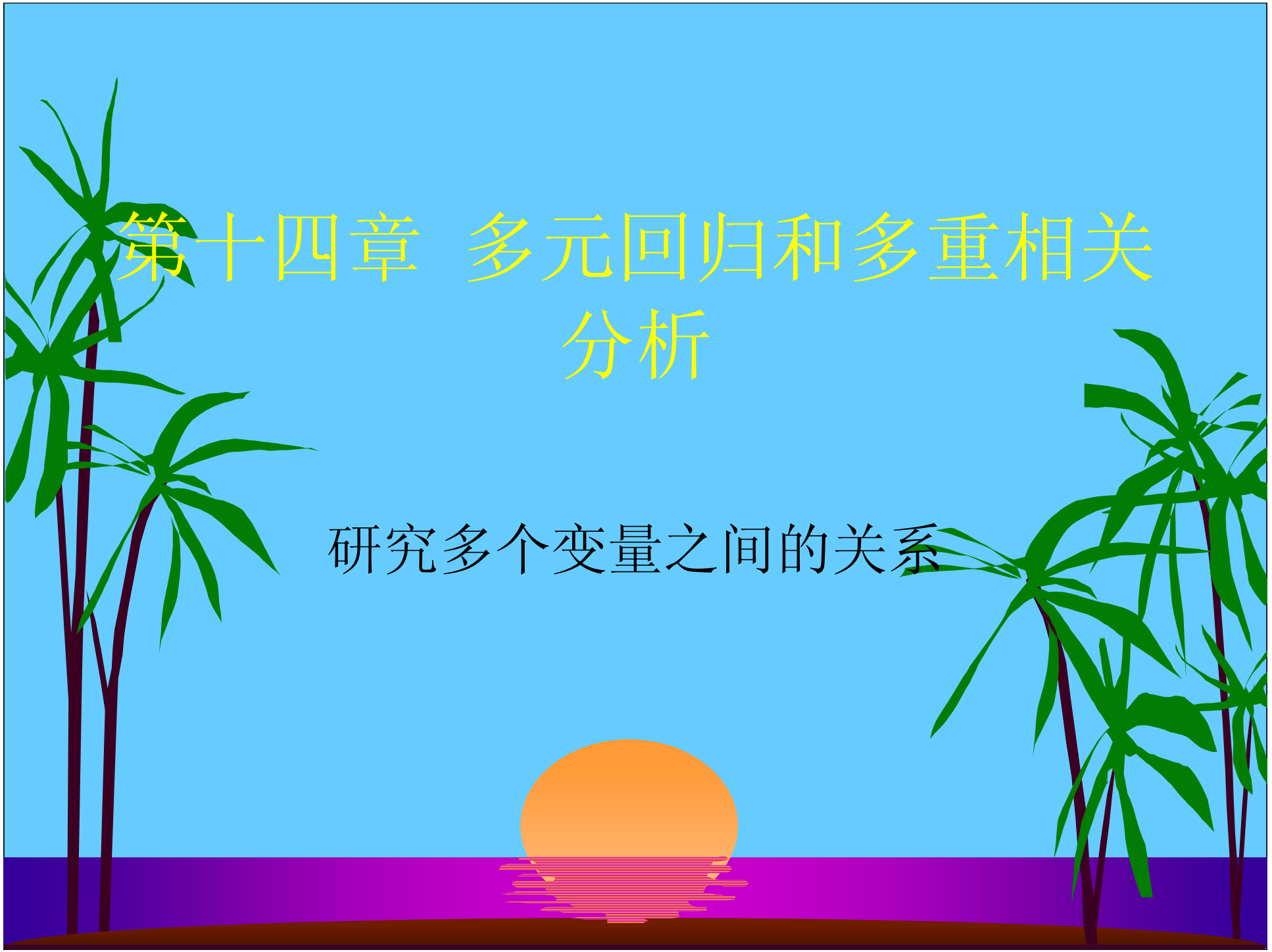
$$=\frac{n\sum XY-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\sum X^2-\left(\sum X\right)^2}\sqrt{n\sum Y^2-\left(\sum Y\right)^2}}$$

$$=0.8257$$

经调整 可决系数

$$r^2_{adj} = 1 - \frac{\sum\left(Y-\hat{Y}\right)^2 \Big/ n-2}{\sum\left(Y-\bar{Y}\right)^2 \Big/ n-1}$$

$$= 0.6419$$


第十四章 多元回归和多重相关分析
研究多个变量之间的关系

多元线性回归方程
一个因变量和多个自变量

总体回归方程

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \Lambda + \beta_k X_{ki} + \varepsilon_i$$

$$\mu_{y \cdot 123\Lambda k} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \Lambda + \beta_k X_{ki}$$

样本回归方程

$$y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \Lambda + b_k X_{ki} + e_i$$

$$\hat{y} = b_0 + b_1 X_{1i} + b_2 X_{2i} + \Lambda + b_k X_{ki}$$

多重可决系数和 多重相关系数

多重可决系数

$$r^2_{y\cdot 123\Lambda\, k} = \frac{\sum\left(\hat{y} - \bar{y}\right)^2}{\sum\left(y - \bar{y}\right)^2} = \frac{SSR}{SST}$$

多重相关系数

$$r_{y\cdot 123\Lambda\, k} = \sqrt{r^2_{y\cdot 123\Lambda\, k}}$$

多重回归分析中的F检验

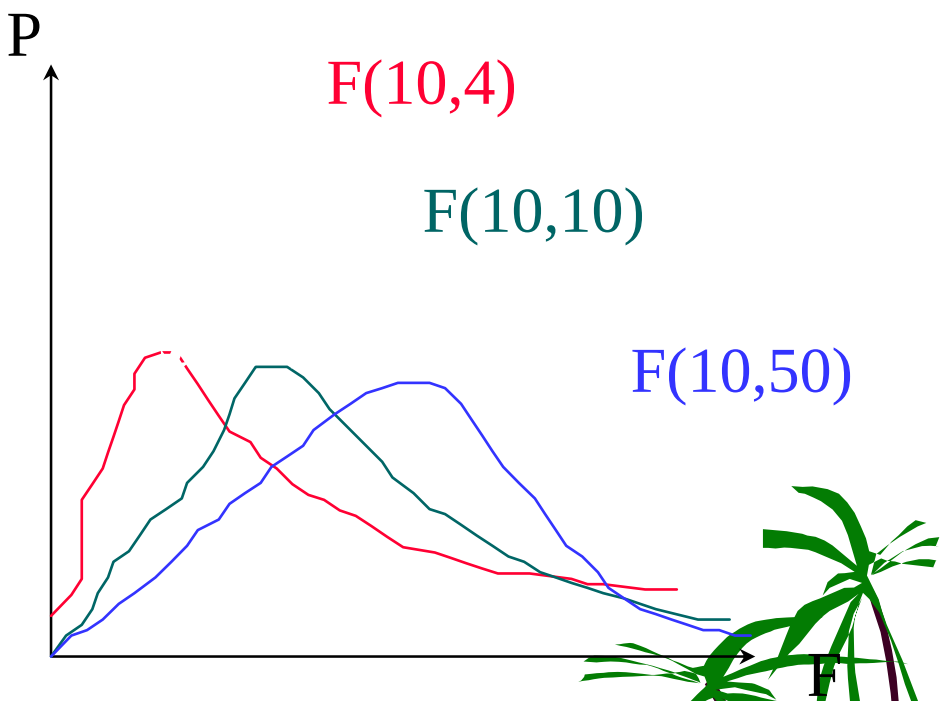
- 总检验：检验某一因变量和k个自变量在总体上是否有显著的线性关系

- 偏检验：检验因变量Y与新引入回归方程的自变量X_k是否有显著的偏相关关系

F分布

两个独立的 t分布被各自的自由度去除，所得之商的比率服从F分布.它是一种非对称分布,图形的形状取决于分子和分母的自由度.

F分布的图形



多元回归模型的总检验

1. $H_0: \beta_1 = \beta_2 = \Lambda = \beta_k = 0$

   H_1: 并非所有的β都为零

2. 据给定的$\alpha \Rightarrow F^*$

3. 根据样本资料计算统计量F

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{SSR/k}{SSE/n-1-k} = \frac{\Sigma\left(\hat{y}-\bar{y}\right)^2/k}{\Sigma\left(y-\bar{y}\right)^2/n-1-k}$$

4. 如果$F \rangle F^*$则拒绝H_0, 否则接收H_0

5. 得出结论:

多元回归模型的偏检验

1. $H_0: \beta_k = 0$

   $H_1: \beta_k \neq 0$

2. 据给定的$\alpha \Rightarrow F^*$

3. 计算统计量F

$$F = \frac{\left[SSR\left(X_1, X_2, \Lambda\ X_k\right) - SSR\left(X_1, X_2, \Lambda\ X_{k-1}\right)\right] / 1}{SSE\left(X_1, X_2, \Lambda\ X_k\right) / n-1-k}$$

4. 如果$F \rangle F^*$则拒绝H_0, 否则接收H_0

5. 结论:

分子为引入第K个变量后可解释变差的增加量，或者说为引入第K个变量后不可解释变差的减少量

分母为引入第K个变量后 不可解释变差

F检验的应用1----总检验

资料来自书中第486页

$$\hat{Y} = 7.7702 + 0.2022X_1 + 0.3209X_2 - 0.3842X_3$$

1. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

   H_0: 并非$\beta_1, \beta_2, \beta_3$都等于零.

2. $\alpha = 0.05 \Rightarrow F^*(3,12) = 3.49$

3. $$F = \frac{SSR/k}{SST/n-1-k} = \frac{3962.4/3}{284.5/12} = 55.7$$

4. $\Theta\ F > F^*$ 所以拒绝H_0

5. 结论: 总体回归方程通过总体检验, 说明Y和诸自变量之间存在显著回归, 即认为所拟合的样本回归方程在总体上有一定的意义.

F检验的应用2----偏检验

Y：利润额；　　X1：销售部　　X2 ：代销额

X3：合同批数

1. $H_0: \beta_3 = 0$

   $H_1: \beta_3 \neq 0$

2. $\alpha = 0.05 \Rightarrow F^*_{(1,12),0.05} = 4.75$

3. $$F = \frac{\left[SSR\left(X_1, X_2, X_3\right) - SSR\left(X_1, X_2\right)\right]/1}{SSE\left(X_1, X_2, X_3\right)/n-1-k}$$

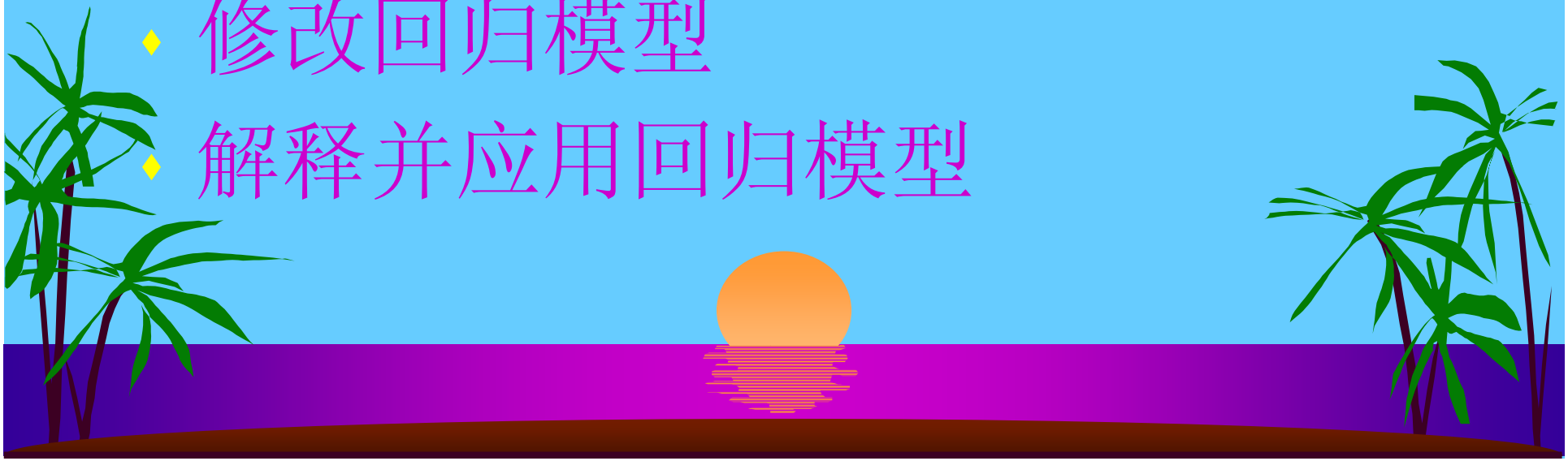
$$= \frac{3962.4 - 3624.2}{284.5/16-1-3} = 9.51$$

4. Θ F〉F^*，所以拒绝H_0

5. 结论：合同批数对利润额有显著的偏回归.

建立回归模型的步骤

- 找出被选变量
- 试建回归模型
- 评核回归模型
- 修改回归模型
- 解释并应用回归模型



spss的输出结果（资料来自第517页）

Equation Number 1　Dependent Variable.. 年外汇收入

---------- Variables in the Equation ----------

Variable ?	B	SE B	T	Sig T
X1（侨胞旅游人数）	4.917499	1.003854	4.899	.0006
X2（外国旅游人数）	-15.762767	16.185008	-.974	.3531
(Constant)	6.825275	6.953243	.982	.3495

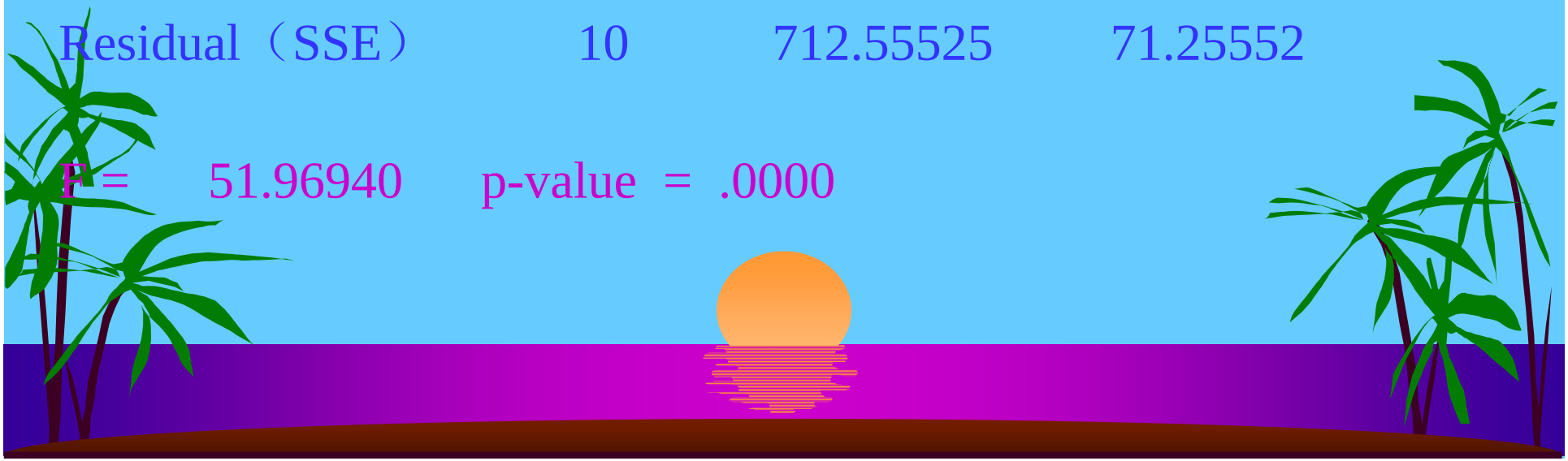
相关系数	Multiple R	.95511
可决系数	R Square	.91223
经调整的 可决系数	Adjusted R Square	.89468
估计标准误差	Standard Error	8.44130

spss的输出结果续

Analysis of Variance（方差分析）

	DF（自由度）	Sum of Squares	Mean Square
Regression（SSR）	2	7406.21398	? 3703.10699
Residual（SSE）	10	712.55525	71.25552

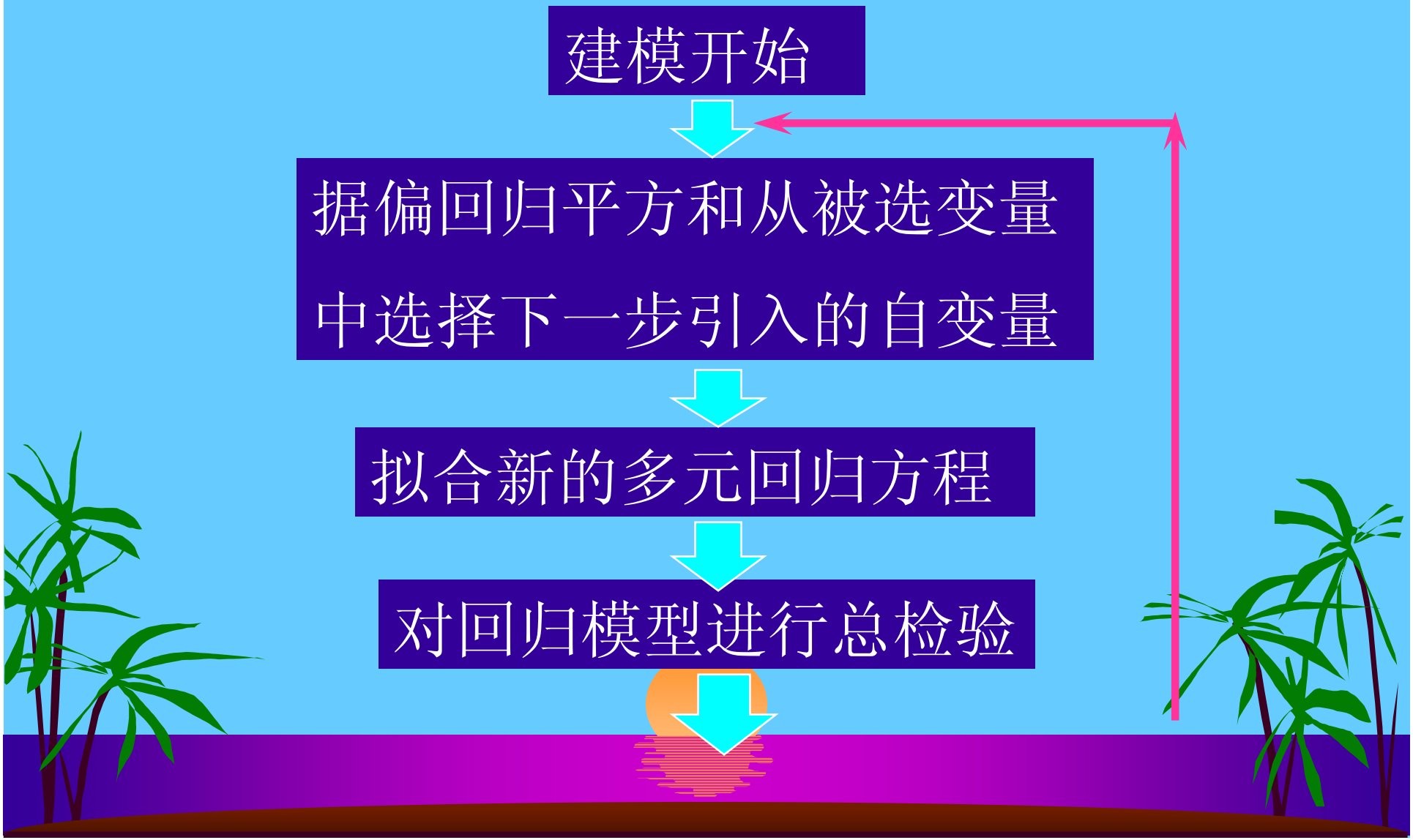
F = 51.96940 p-value = .0000



逐步回归法

是按一定的统计程序，经过多步拟合和检验，从一系列的可供建立回归模型的自变量中，逐步引入回归作用显著的自变量，并从回归模型中逐步趋逐回归作用变得不在显著的自变量，以最终求得“最优”回归模型的技术.

逐步回归法的基本原理



逐步回归法的基本原理

对引入的新变量进行偏检验

对原有的诸自变量进行偏检验

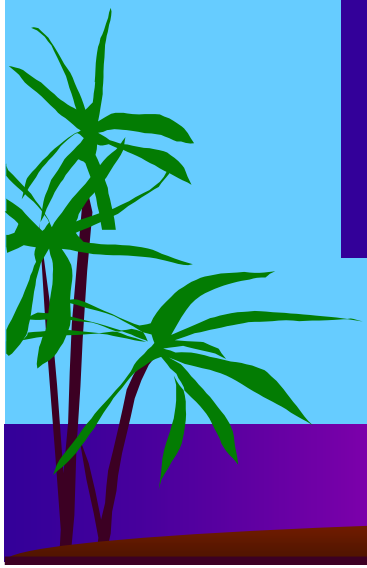
决定：1。新模型是否有显著意义

2。新引入的自变量是否可被接纳

3。原有的自变量是否应被逐出去

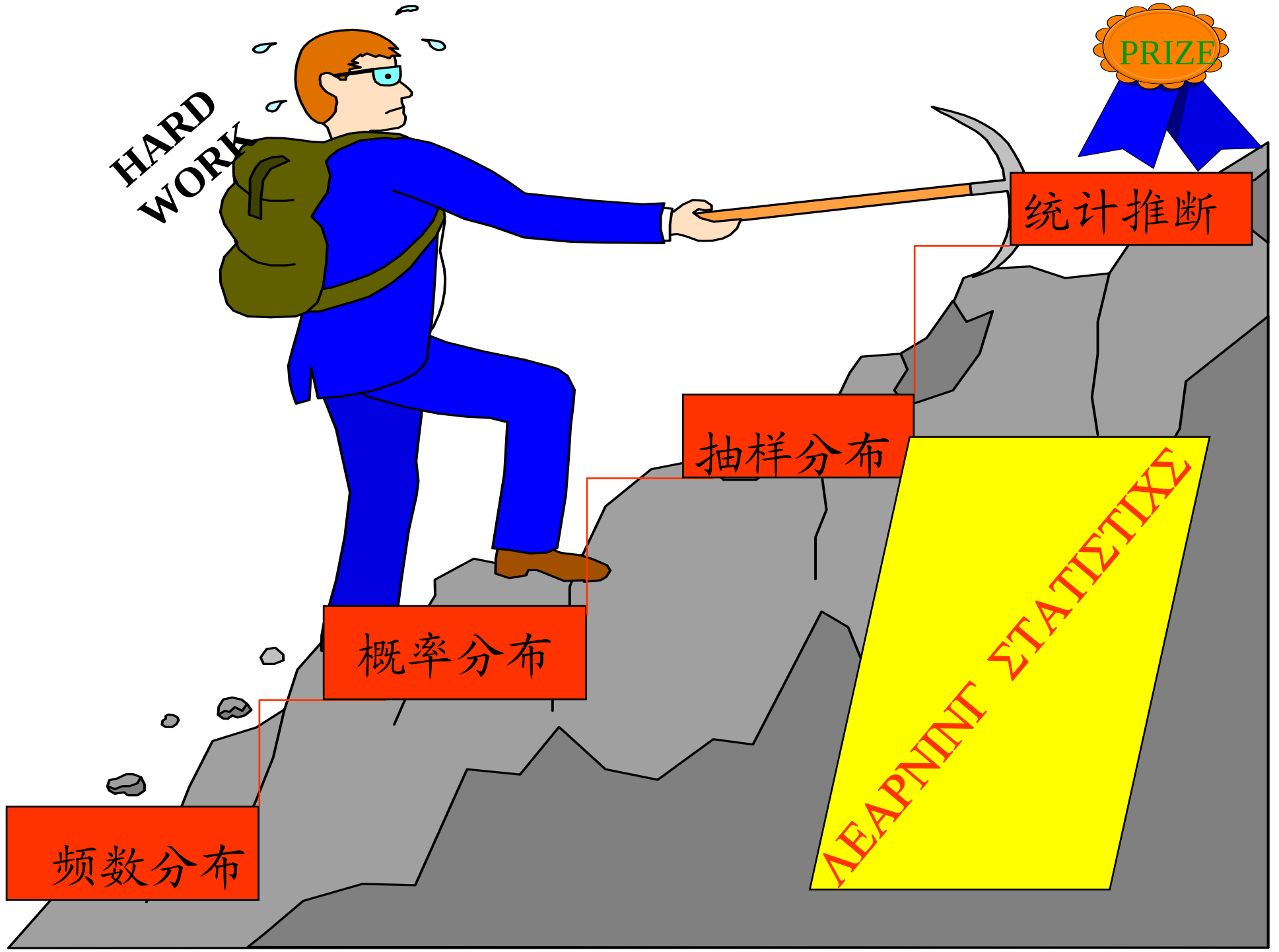
4。是进行下一步回归还是终止

得出最优模型



回归和相关分析中应注意的问题

- 要正确理解和对待变量之间的关系
  - 定量分析之前应进行定性分析
  - 相关关系和因果关系
- 利用回归方程预测时，自变量的取值范围应在样本的取值范围之内
- 利用回归方程预测时，特别注意现在的各项条件是否与建立回归方程时一致


HARD
WORK
PRIZE
统计推断
抽样分布
概率分布
频数分布
ΛΕΑΡΝΙΝΓ ΣΤΑΤΙΣΤΙΧΣ

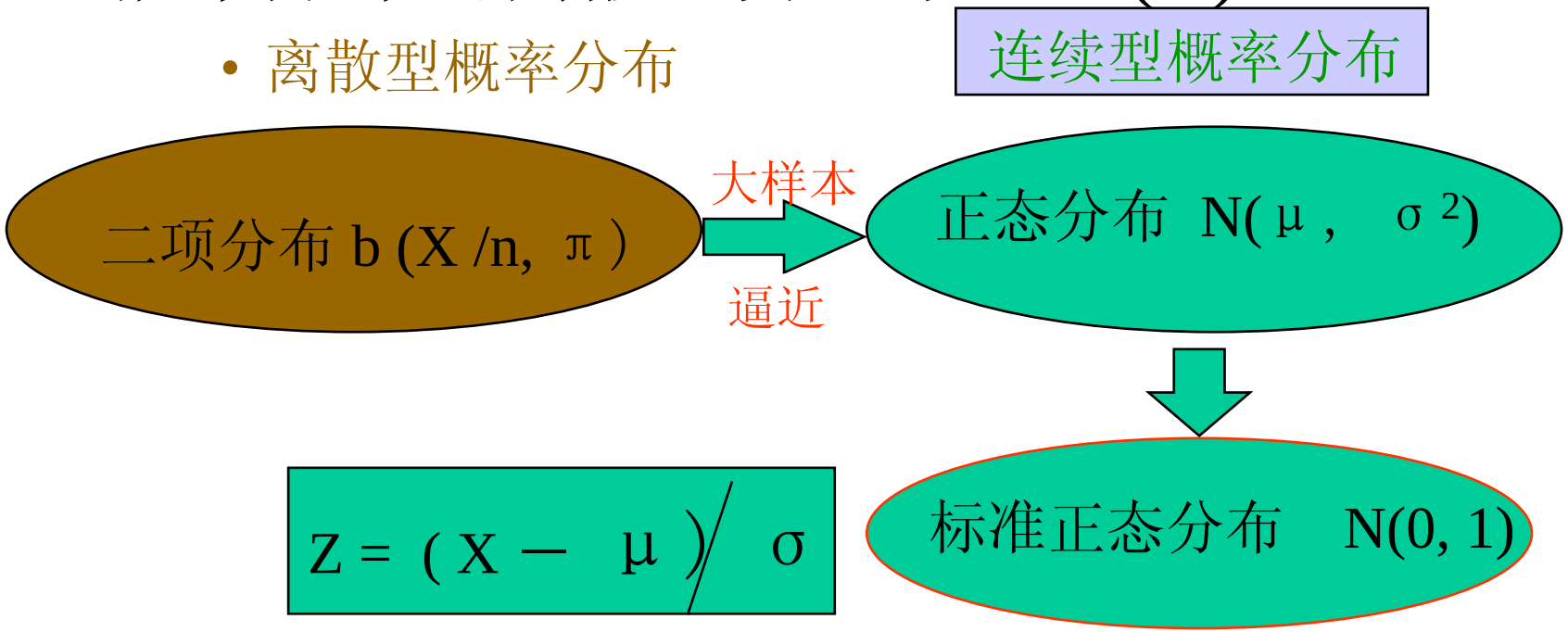
频数分布： 实测数据的整理结果，
变量观察值与出现频次相对应

- 频数分布表： 变量分组、频数、相对频数、累计频数
- 频数分布图： 直方图、频数多边形
- 数据分布特征及其量数：

	集中趋势量数	离散趋势量数
计算量数	算术平均数 $\bar{X}$	方差 s^2，标准差 s
位次量数	中位数 Me 众　数 Mo	全距 $X_{max} - X_{min}$ 四分位距 $Q_3 - Q_1$

概率分布：理论分布。
随机变量取值及其概率相对应。

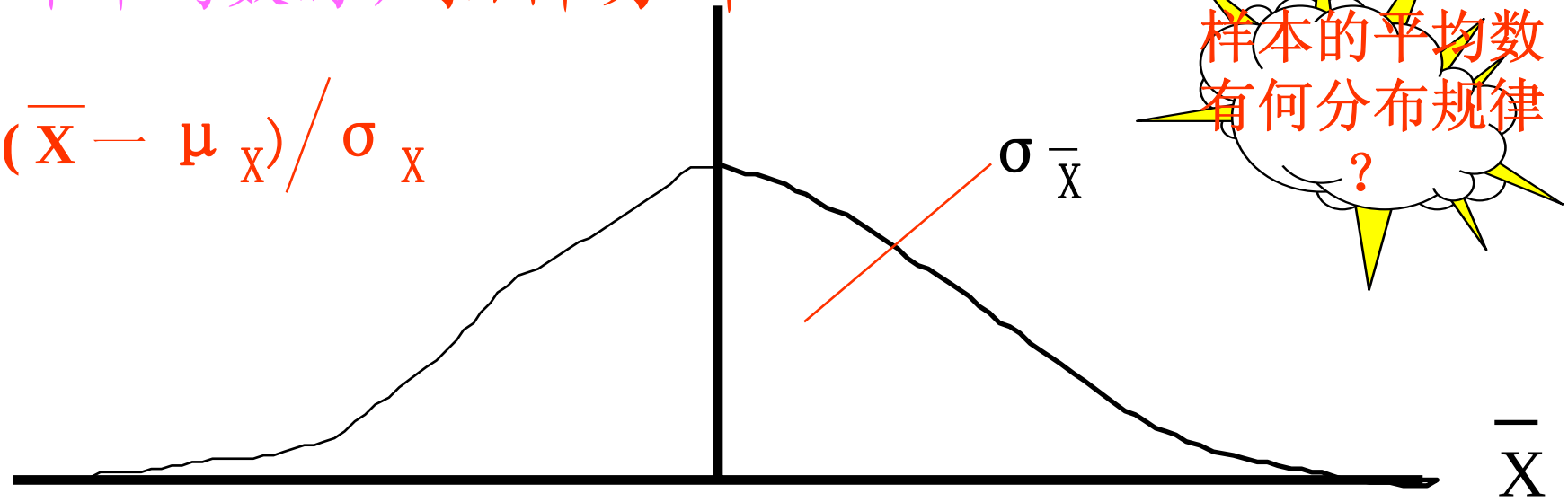
- 概率分布的表述形式：表、图、公式
- 概率分布的集中趋势量数：$E(X) = \mu$
- 概率分布的离散趋势量数?$Var(X) = \sigma^2$
  - 离散型概率分布

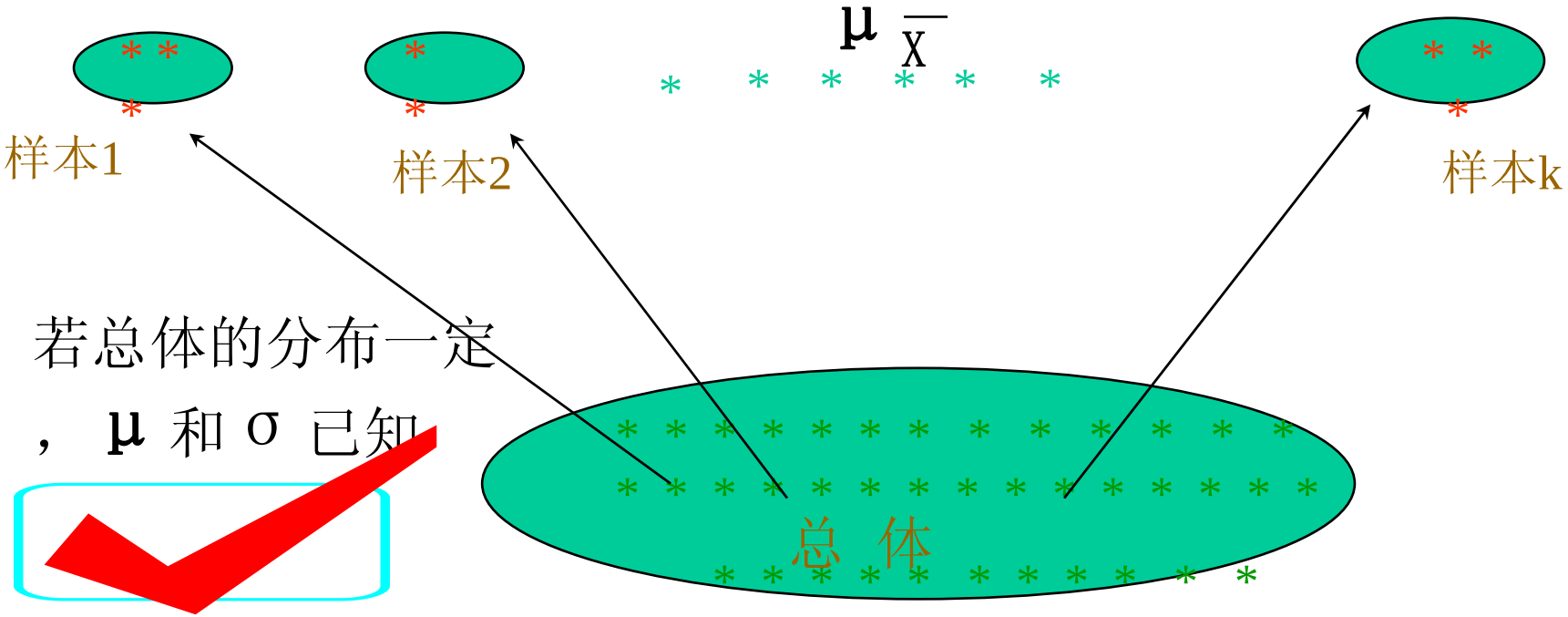


（样本平均数的）抽样分布

$$Z = (\overline{X} - \mu_X) / \sigma_X$$

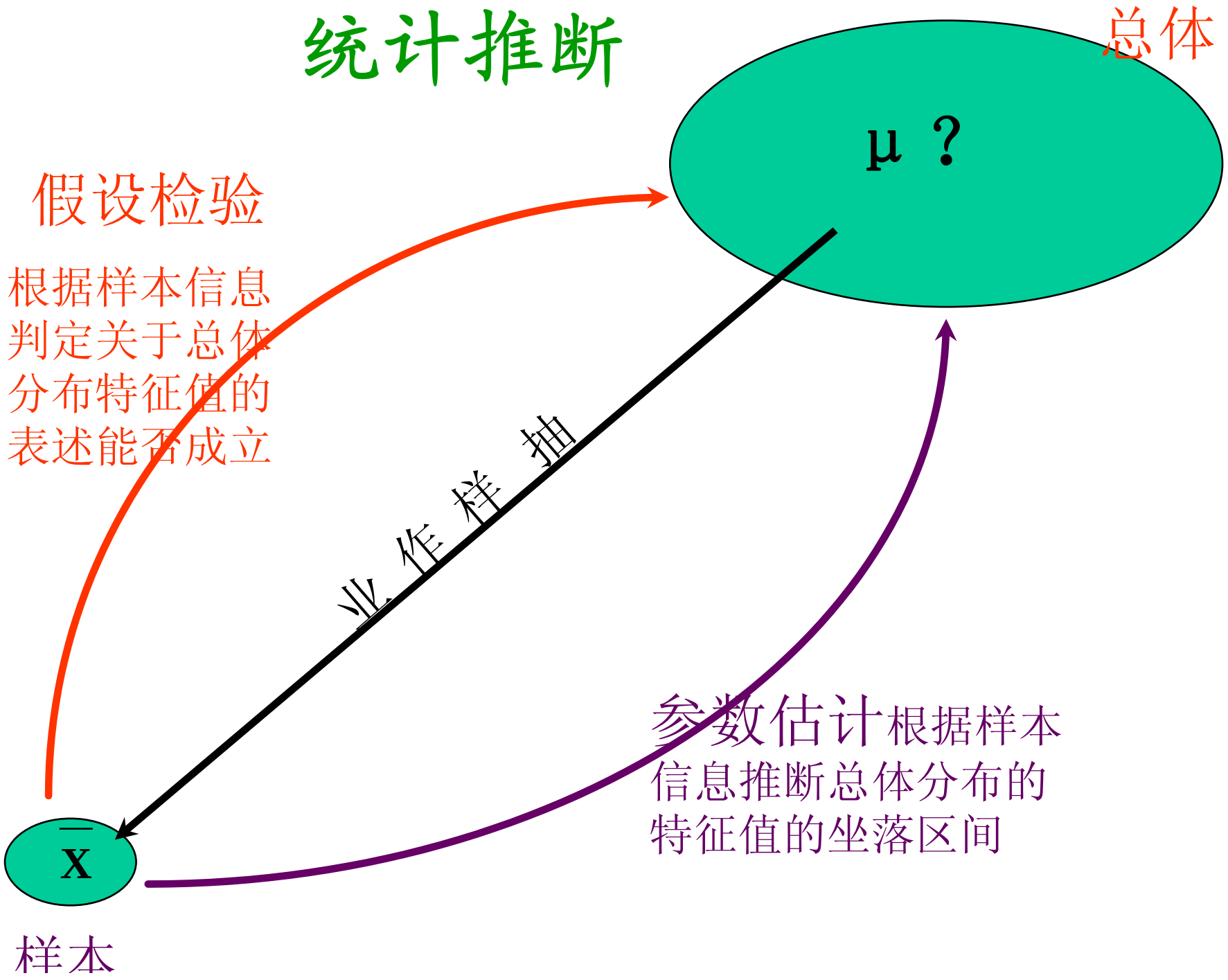
从中抽取的诸样本的平均数有何分布规律？





若总体的分布一定，μ 和 σ 已知

统计推断


总体
μ ？
假设检验
根据样本信息
判定关于总体
分布特征值的
表述能否成立
抽样作业
参数估计根据样本
信息推断总体分布的
特征值的坐落区间
$\bar{x}$
样本