第一章

绪论

主要内容

* 第一节　统计学的基本概念和思想
* 第三节　统计数据的收集方法
* 第三节　统计数据的整理
* 第四节　数据的描述性统计

第一节 统计学基本概念

。

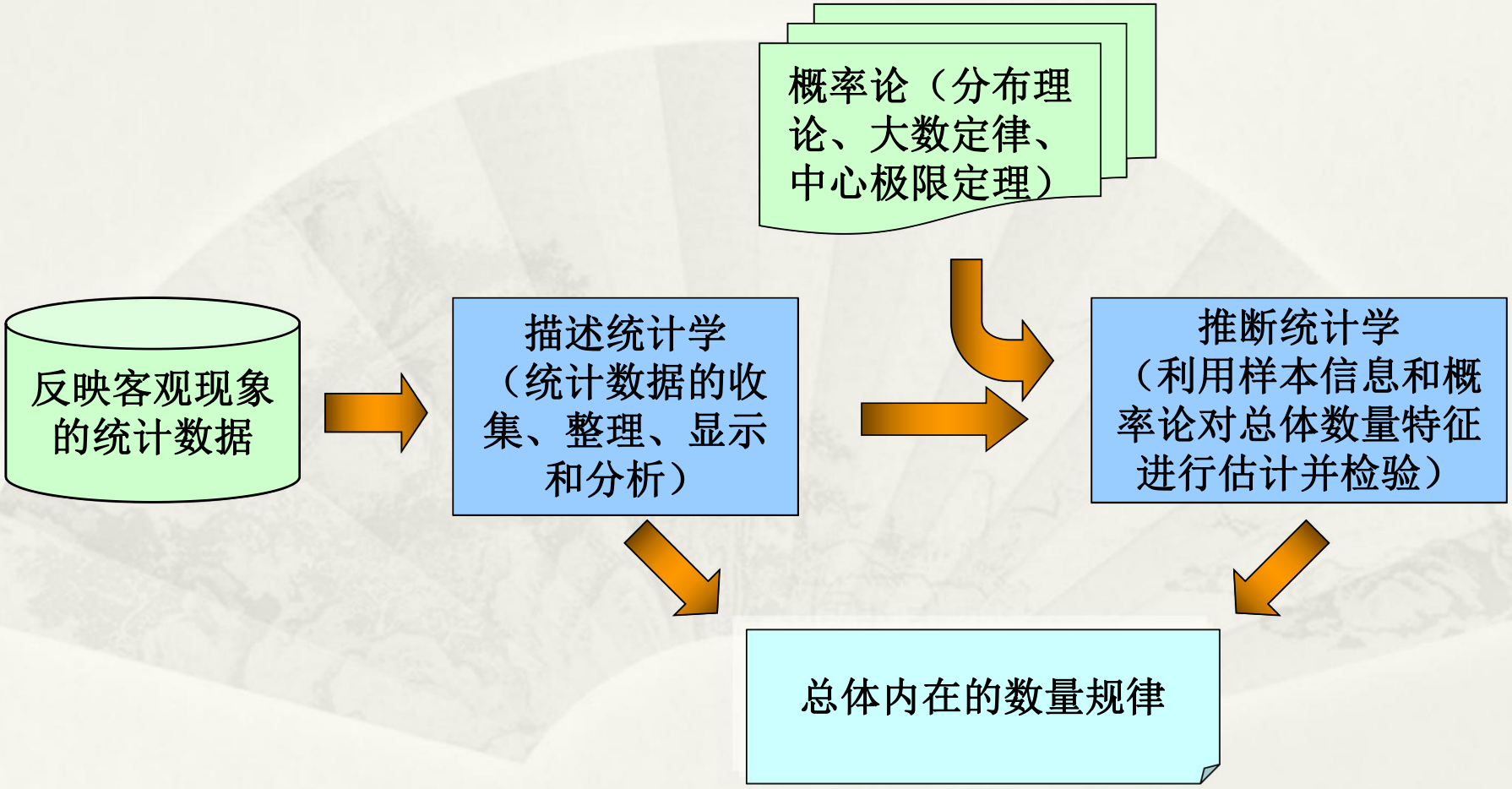
管理统计学

描述统计学

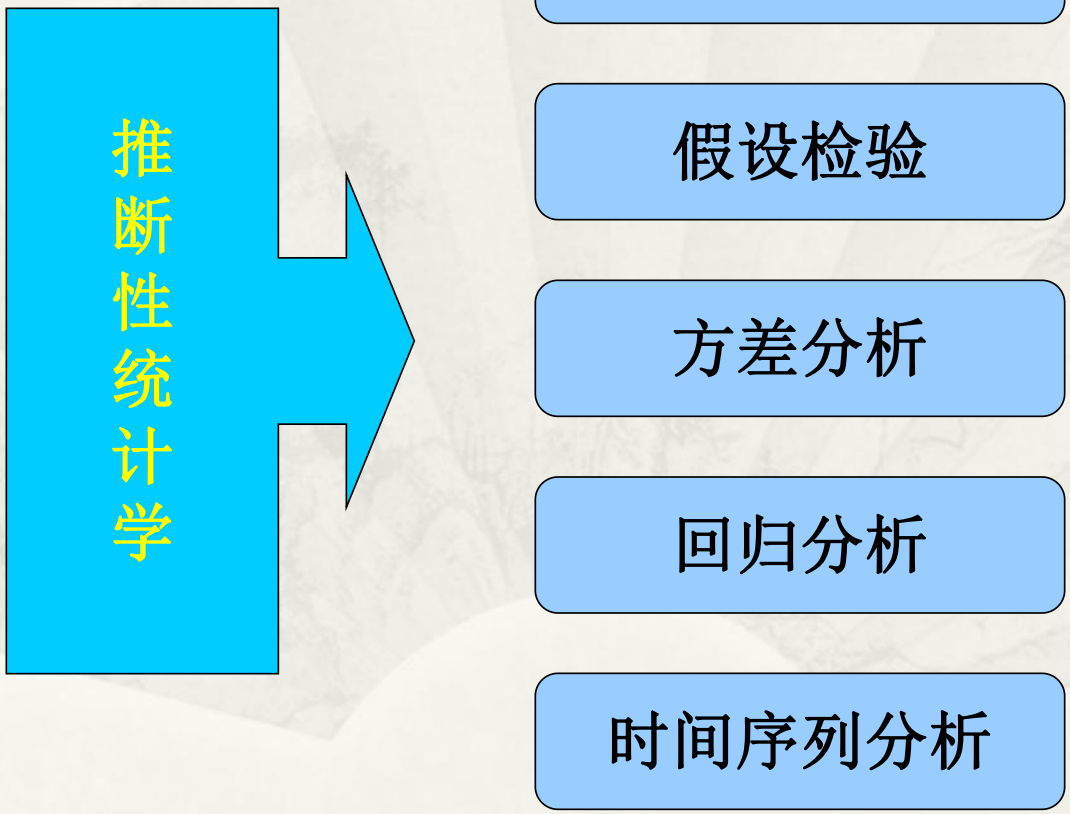
研究如何取得反映客观现象的数据，并通过图表形式对所收集的数据进行加工处理和显示，进而通过综合、概括与分析得出反映客观现象的规律性数量特征

推断统计学

研究如何通过样本数据去推断总体数量特征。是在对样本数据进行描述的基础上，对统计总体的未知数量特征作出以概率形式表述的推断。



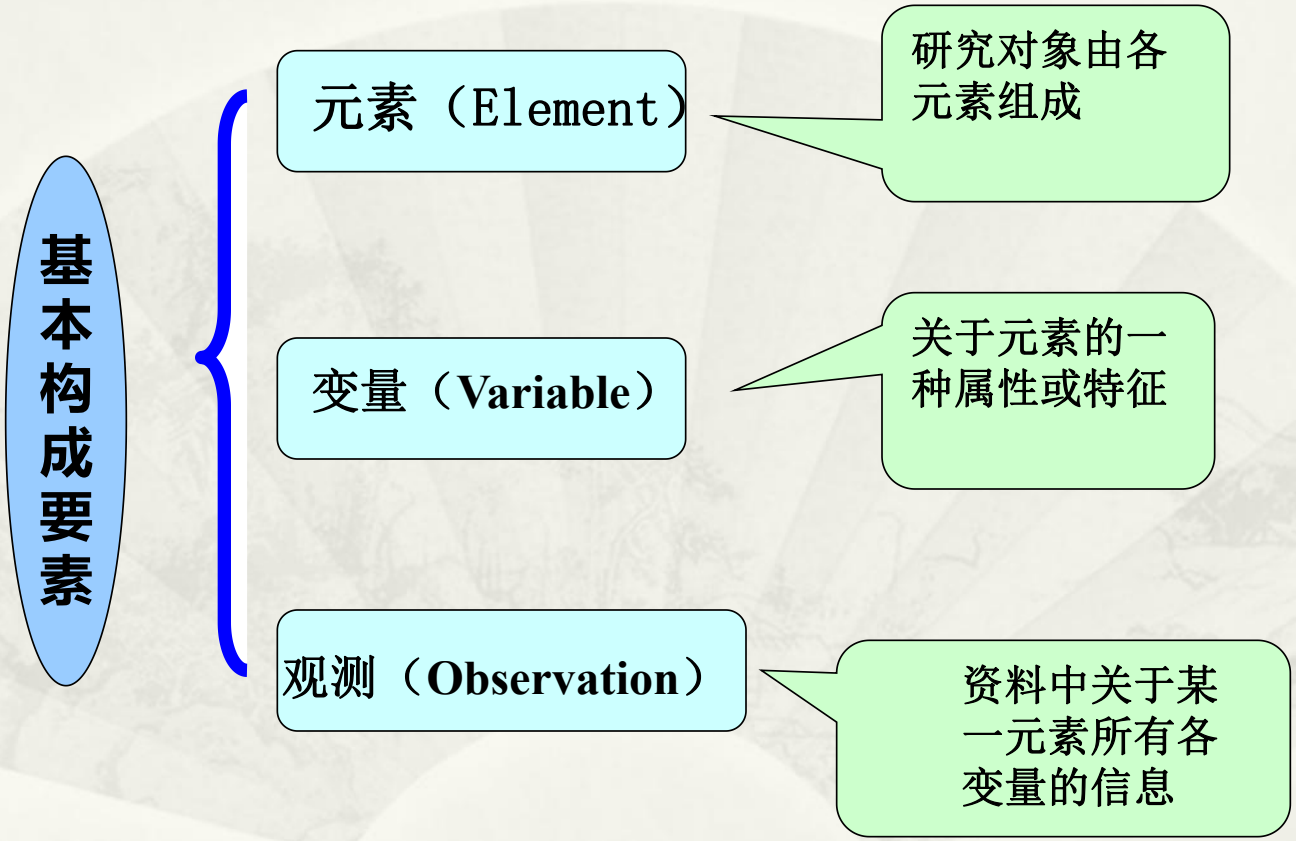
统计学探索客观现象数量规律性的过程


推断性统计学
参数估计
假设检验
方差分析
回归分析
时间序列分析

1、统计资料

定义：

统计资料（Statistical data）是指可用以推导出某项结论的一些事实或数字


基本构成要素
元素（Element）
研究对象由各元素组成
变量（Variable）
关于元素的一种属性或特征
观测（Observation）
资料中关于某一元素所有各变量的信息

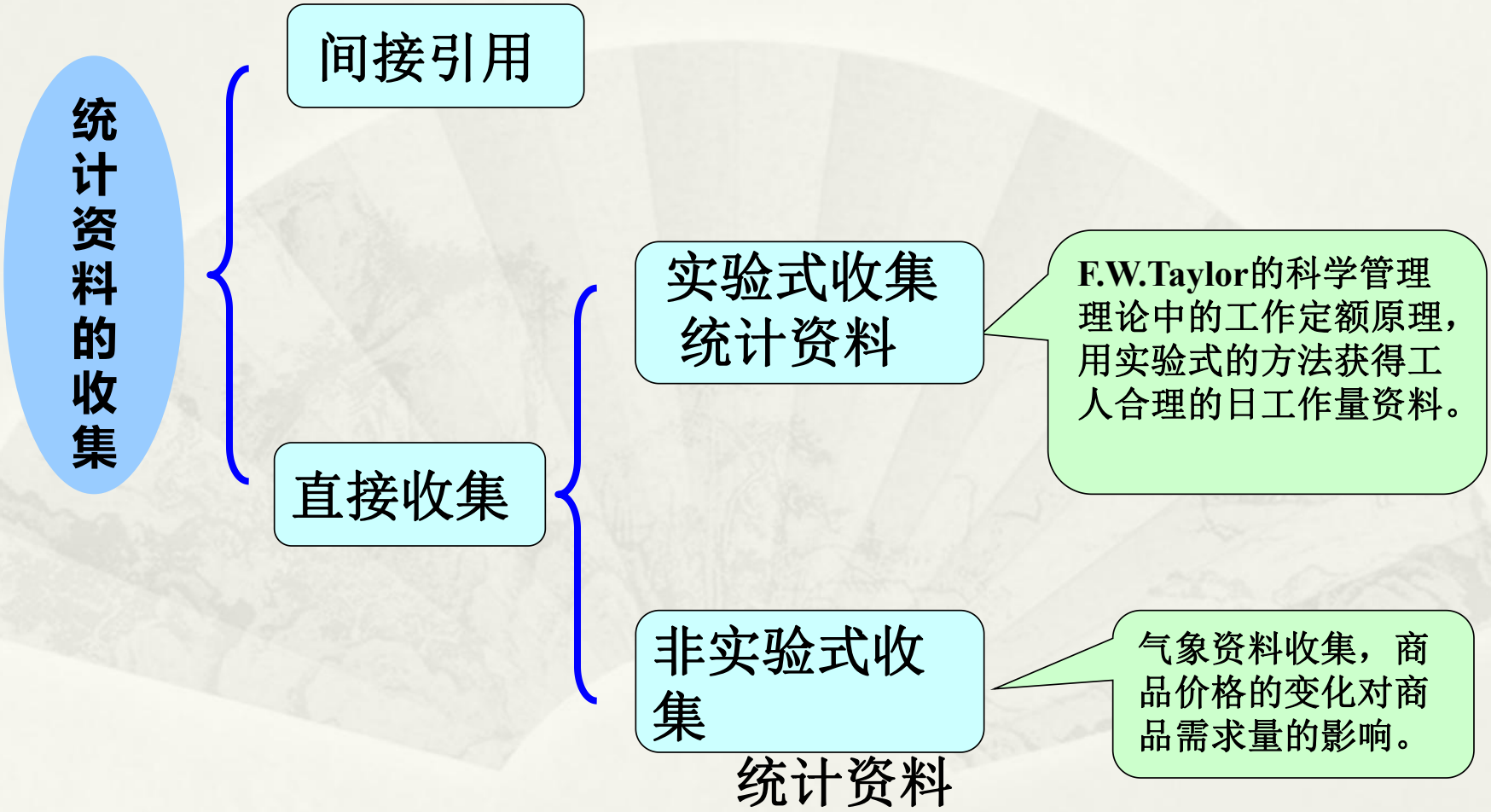
变量（Variable）

- 定量变量（Quantitative variable）
  结果可用数字表示
- 定性变量（Qualitative Variable）
  结果不可用数字表示

表1　员工个人资料表

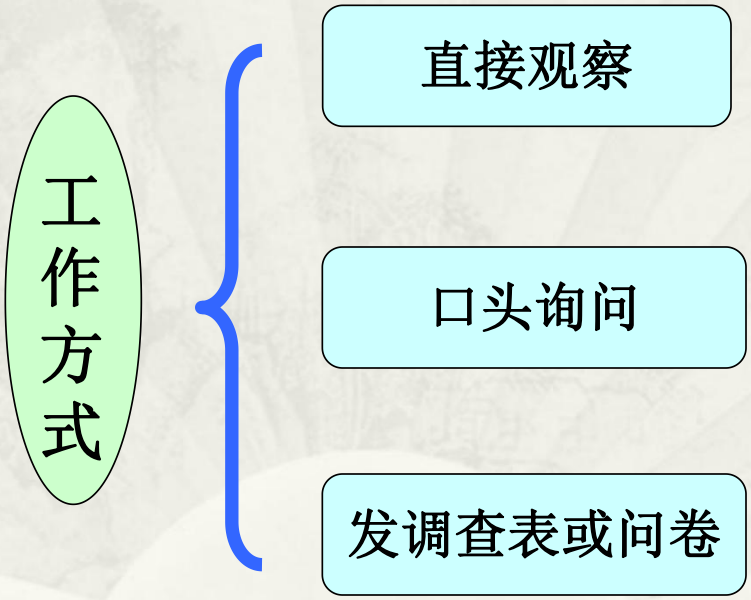
姓名	性别	年龄	身高(m)	体重(kg)	民族	公司服务年限	受教育年限
甲	男	33	1.85	65	汉	3	18
乙	女	25	1.65	55	回	2	16
丙	男	26	1.72	60	满	1	15
丁	女	35	1.60	53	回	4	16
戊	男	32	1。83	68	汉	2	19

2、统计资料的收集


统计资料的收集
间接引用
直接收集
实验式收集统计资料
F.W.Taylor的科学管理理论中的工作定额原理，用实验式的方法获得工人合理的日工作量资料。
非实验式收集统计资料
气象资料收集，商品价格的变化对商品需求量的影响。

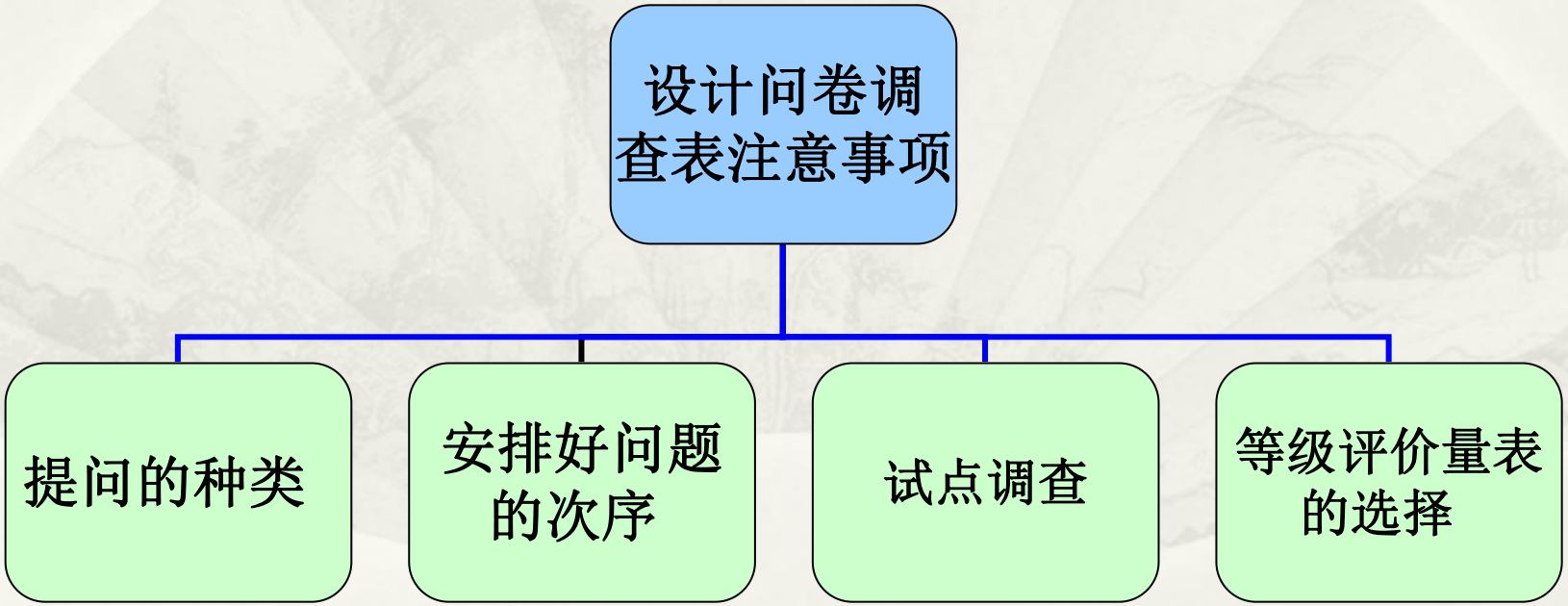
3 、统计调查

直接收集统计资料，无论是实验式的还是非实验式的，都称为统计调查。

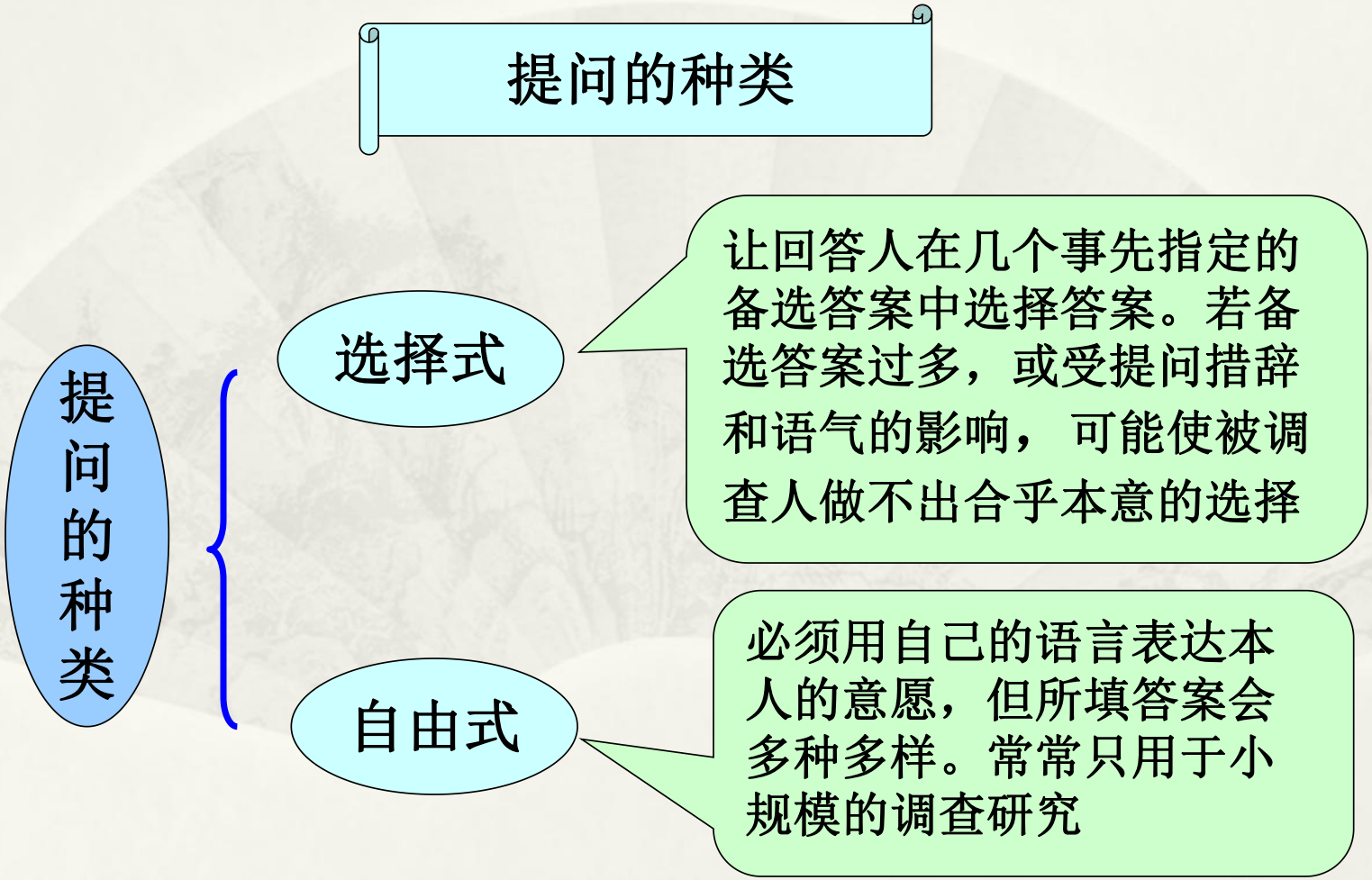


1.3 统计调查

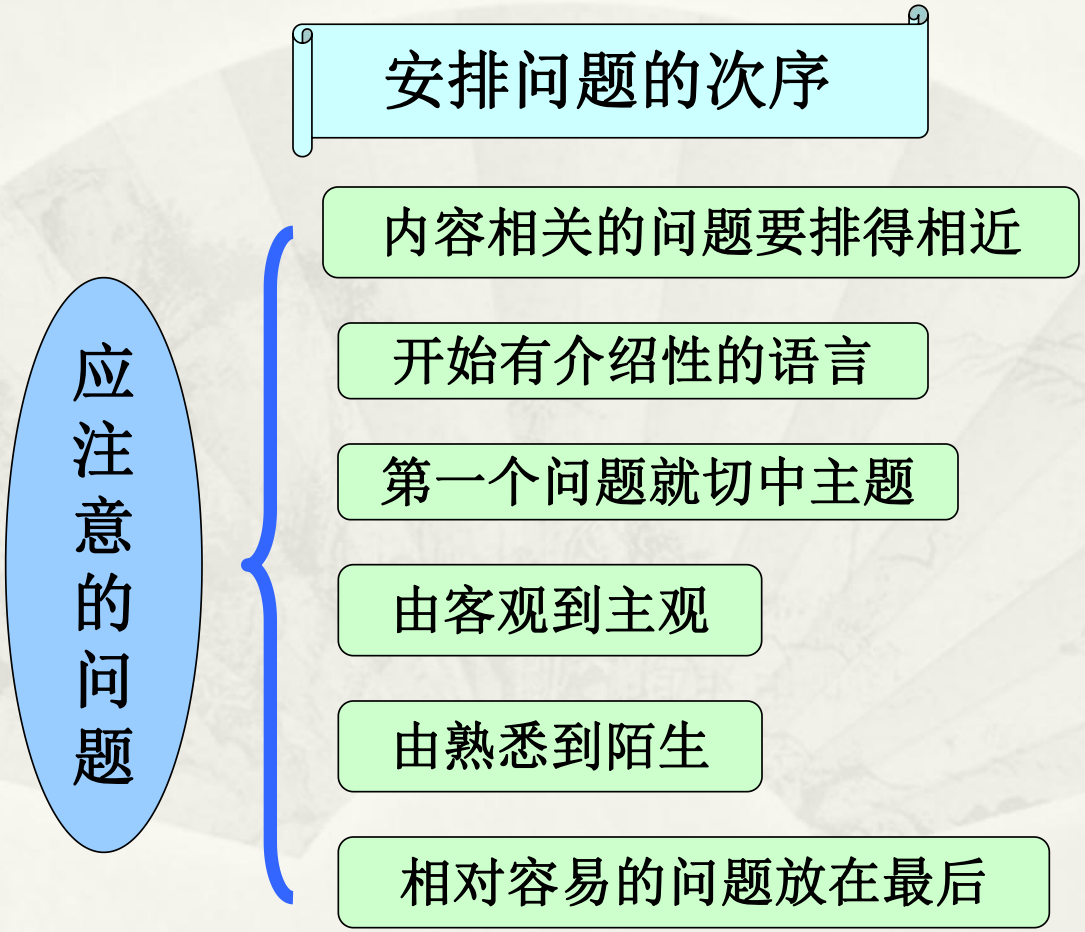
调查表是直接获得统计资料的主要工具，调查表设计的好坏将影响所获资料的可用性与可信性。



统计调查


提问的种类
提问的种类
选择式
自由式
让回答人在几个事先指定的备选答案中选择答案。若备选答案过多，或受提问措辞和语气的影响，可能使被调查人做不出合乎本意的选择
必须用自己的语言表达本人的意愿，但所填答案会多种多样。常常只用于小规模的调查研究

统计调查


安排问题的次序
应注意的问题
内容相关的问题要排得相近
开始有介绍性的语言
第一个问题就切中主题
由客观到主观
由熟悉到陌生
相对容易的问题放在最后

统计调查

案例：一个电话访问的引言和第一个问题

你好，我是XX大学的访问员。我们正在调查居住在学生公寓的人是否对生活条件感到满意。你的名字是从住宿登记簿中随机选取的，我们的调查只会占用您至多四分钟的时间。您可以在任何时候打断我。我现在可以开始访问了吗？

第一个问题是关于您对学生公寓的总体感觉的。

您认为（读选项）：

（1）确实满意

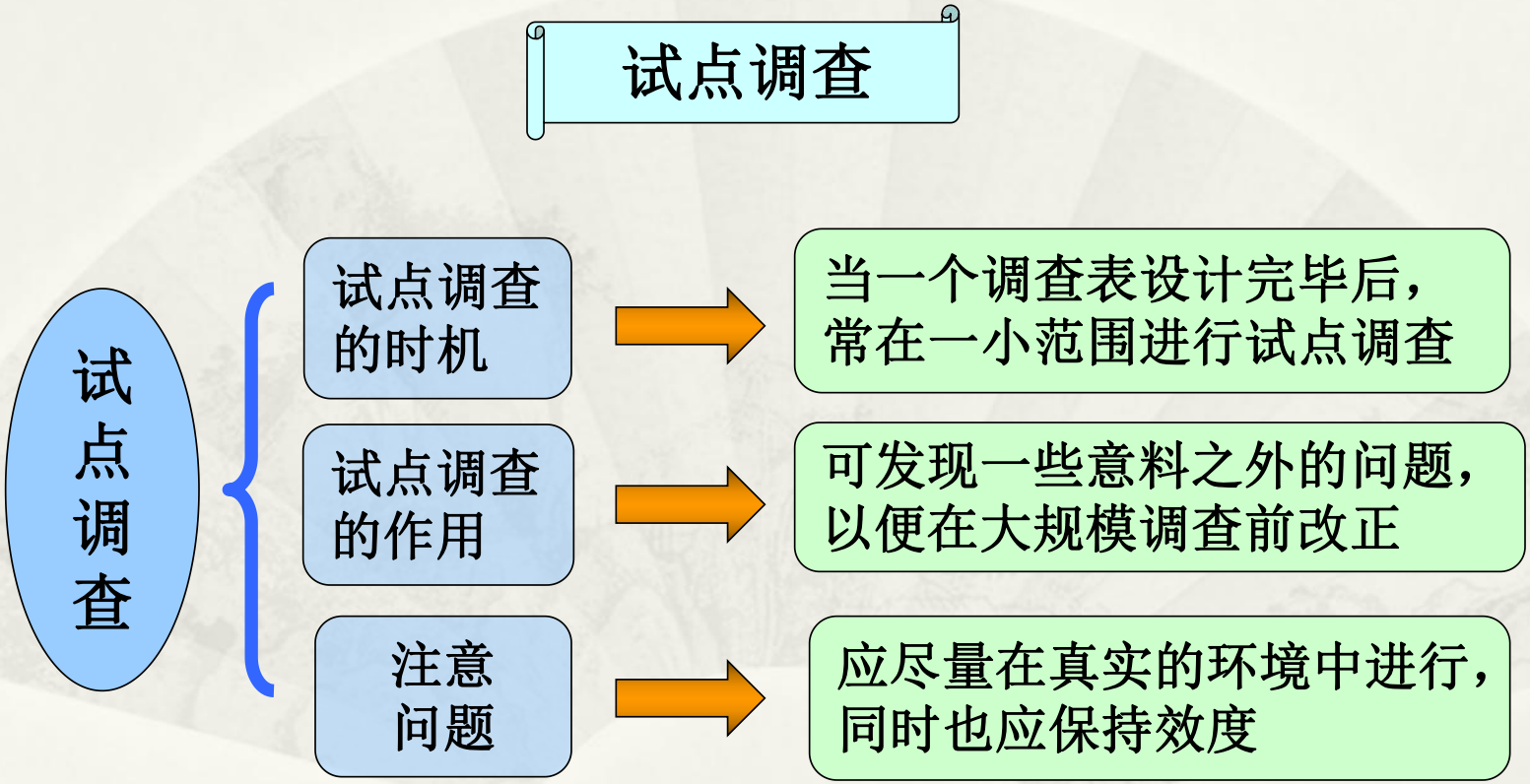
（2）大体满意

（3）大体不满意

（4）确实不满意

（5）（沉默）没想法或者不知道/错误答案

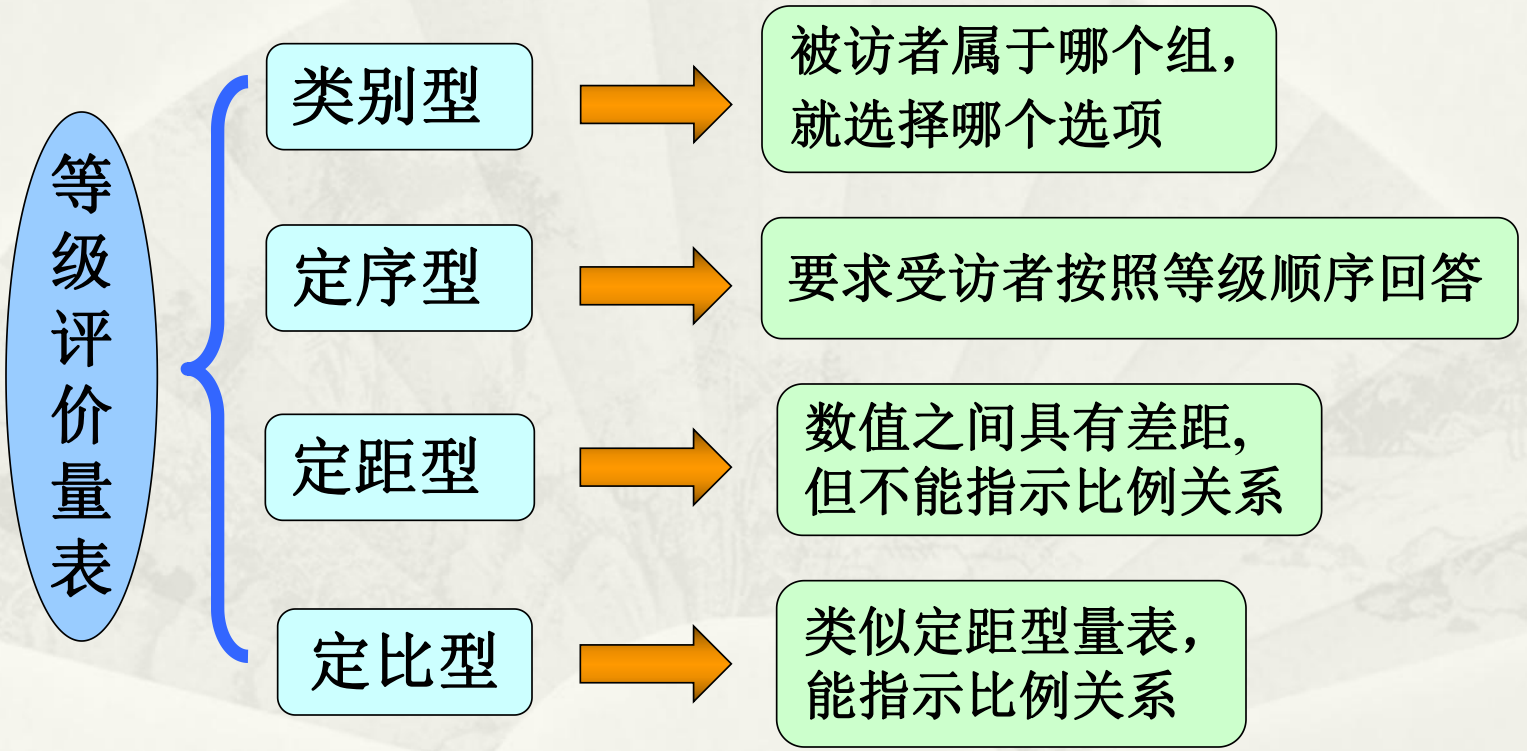
统计调查


试点调查
试点调查
试点调查的时机
当一个调查表设计完毕后，常在一小范围进行试点调查
试点调查的作用
可发现一些意料之外的问题，以便在大规模调查前改正
注意问题
应尽量在真实的环境中进行，同时也应保持效度

统计调查

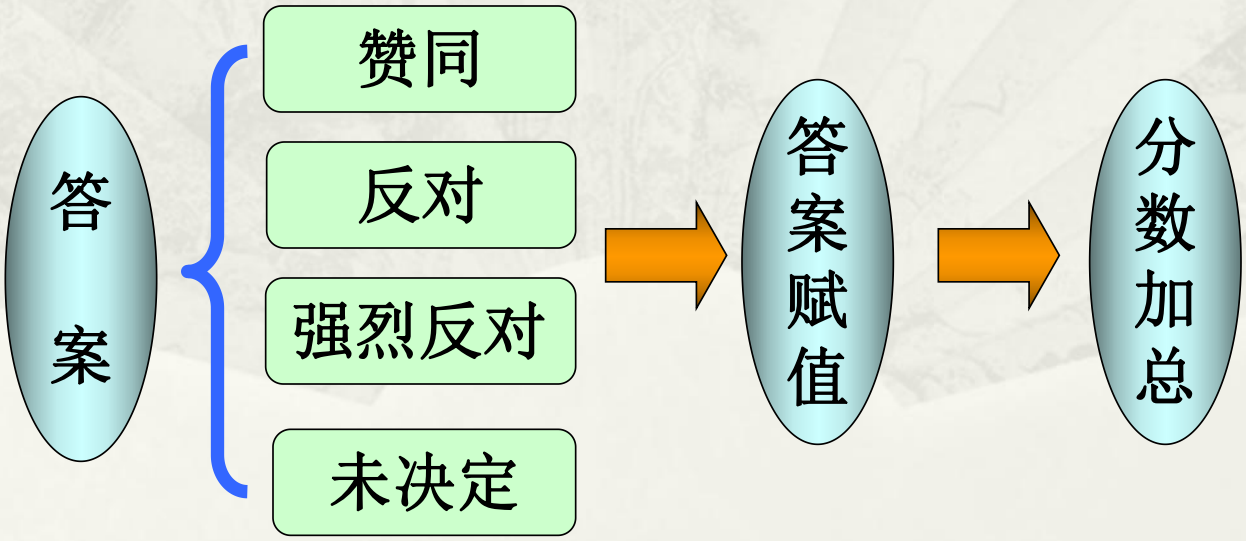
等级评价量表的选择

利用等级评价量表，可以为受访者在一个连续区间的一些点上或者一个类型序列上设定选项，并且为每个级别赋一个量化值。根据实际调查的需要，有四种等级评价量表供选择

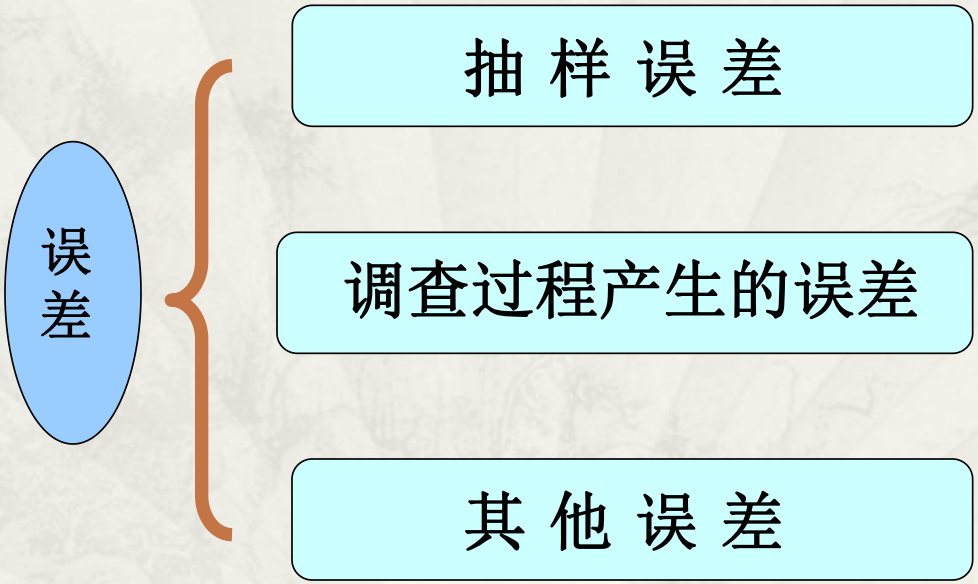

等级评价量表
类别型
被访者属于哪个组，就选择哪个选项
定序型
要求受访者按照等级顺序回答
定距型
数值之间具有差距，但不能指示比例关系
定比型
类似定距型量表，能指示比例关系

李科特量表

李科特量表是一种定距量表，它的基本形式是给出一组陈述，要求调查对象表明他是“强烈赞同”，“赞同”，“反对”，“强烈反对”或“未决定”。最后把各个陈述的分数相加就可以得到总分。

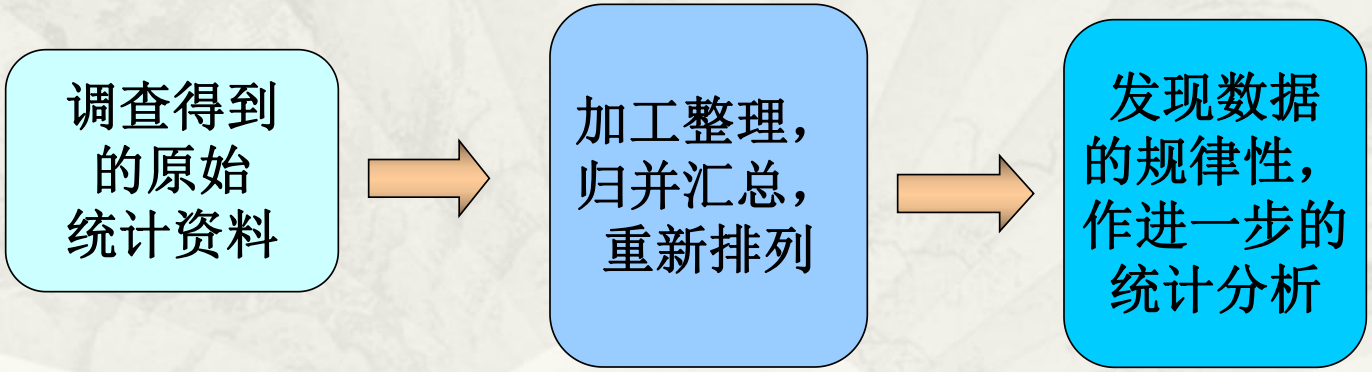


4、统计资料的误差

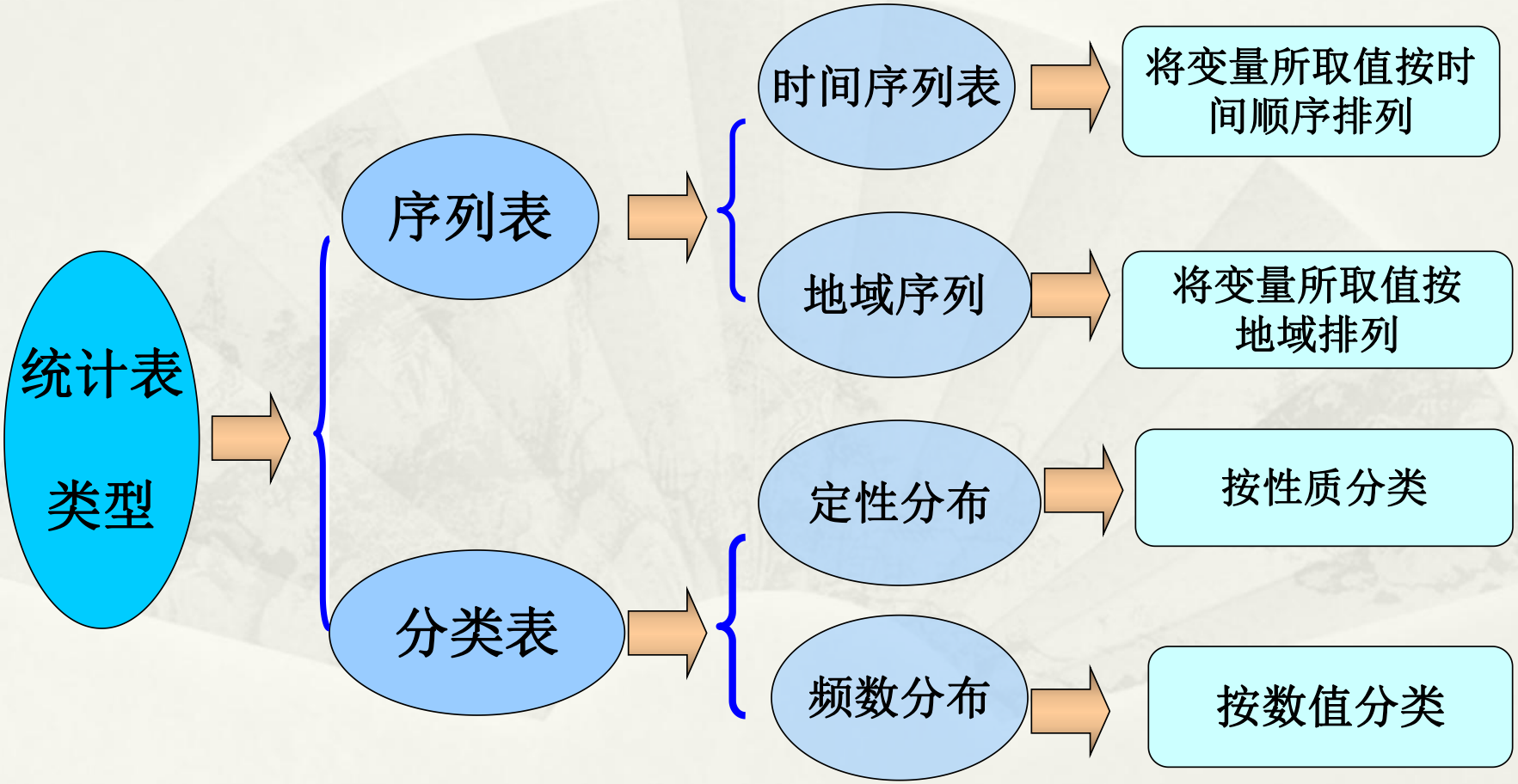


第三节 统计数据的整理

- 调查收集到的原始统计资料常常是大量的。它必须经过加工整理，如分类归并汇总，按时间前后或按数值大小重新排列等，才容易发现数据的规律性，并便于做进一步的统计分析。



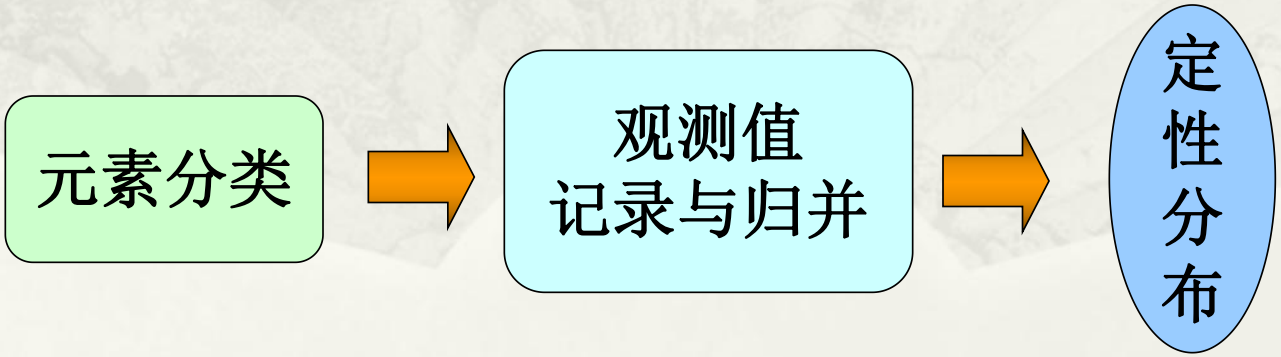
1、统计表


统计表
类型
序列表
分类表
时间序列表
地域序列
定性分布
频数分布
将变量所取值按时间顺序排列
将变量所取值按地域排列
按性质分类
按数值分类

- **定性分布：**

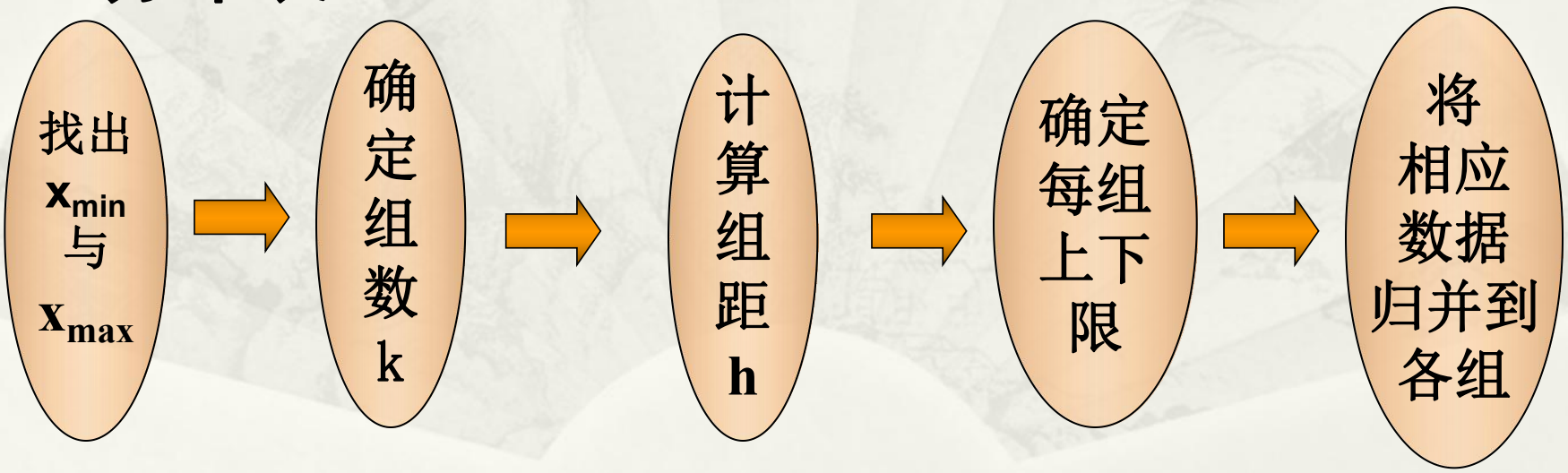
  首先建立一个元素的类别系统，使得各类互相排斥，而且是完备的，使被观测的各元素能既不重复又无遗漏地分到各类中去。

  然后记录分到同类中的元素个数，或将同类中各元素的观测值加以归并，这样得到定性分布。



- **频数分布：**

  按变量所取的值进行分类，于是资料中每个观测值都分到相应类中去。记录各类中观测值出现的次数，制成频数分布表。



X_{min} 最小值,x_{max} 最大值

$$h = \frac{x_{\max} - x_{\min}}{k}$$

表2 某校200个学生高等数学考试成绩

分数	计　　数	人数（f）
40—49	一	1
50—59	正正下	14
60—69	正正正正正正正正正正正	55
70—79	正正正正正正正正正正正下	58
80—89	正正正正正正正正正正下	52
90—99	正正正下	17
100—109	下	3
总数		200

在所属组的记录栏做一记号，按照我国习惯，用写“正”字方法，英文书使用“#”符号

表3　　某校200个学生高等数学考试成绩的频数分布表

分数	人数（f）	分数	人数（f）
40—45	1	76—81	25
46—51	0	82—87	42
52—57	12	88—93	10
58—63	29	94—99	11
64—69	28	100—105	3
70—75	39	总　数	200

累计频数与频率

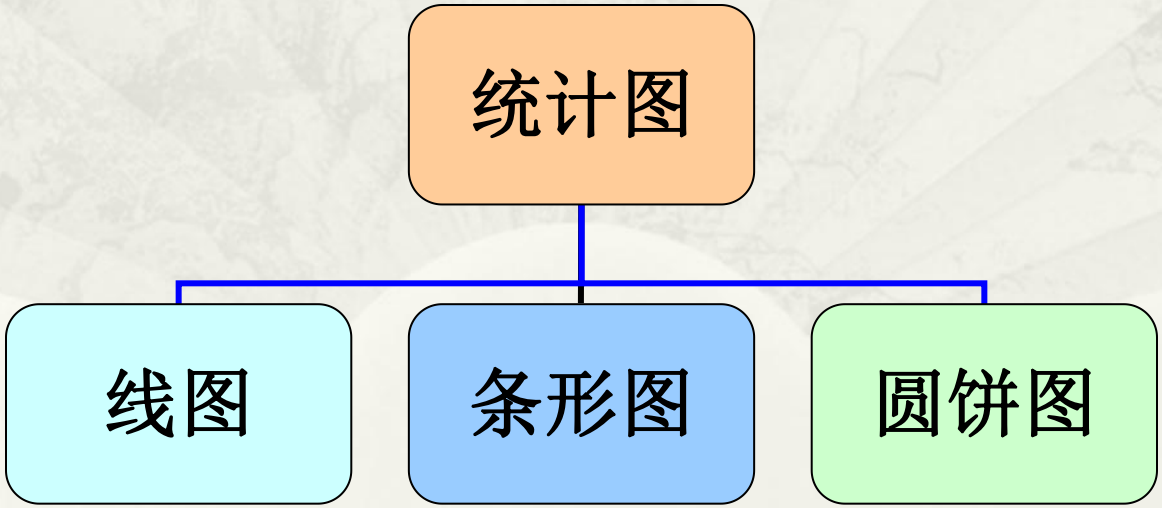
- **累积频数（Cumulative Frequency）：由第一组起至第i组止各频数之和称为第i组的累积频数，记为F_i，即：**

$$F_i = \sum_{k=1}^{i} f_k = F_{i-1} + f_i \qquad (i > 1) \qquad (2\text{-}1)$$

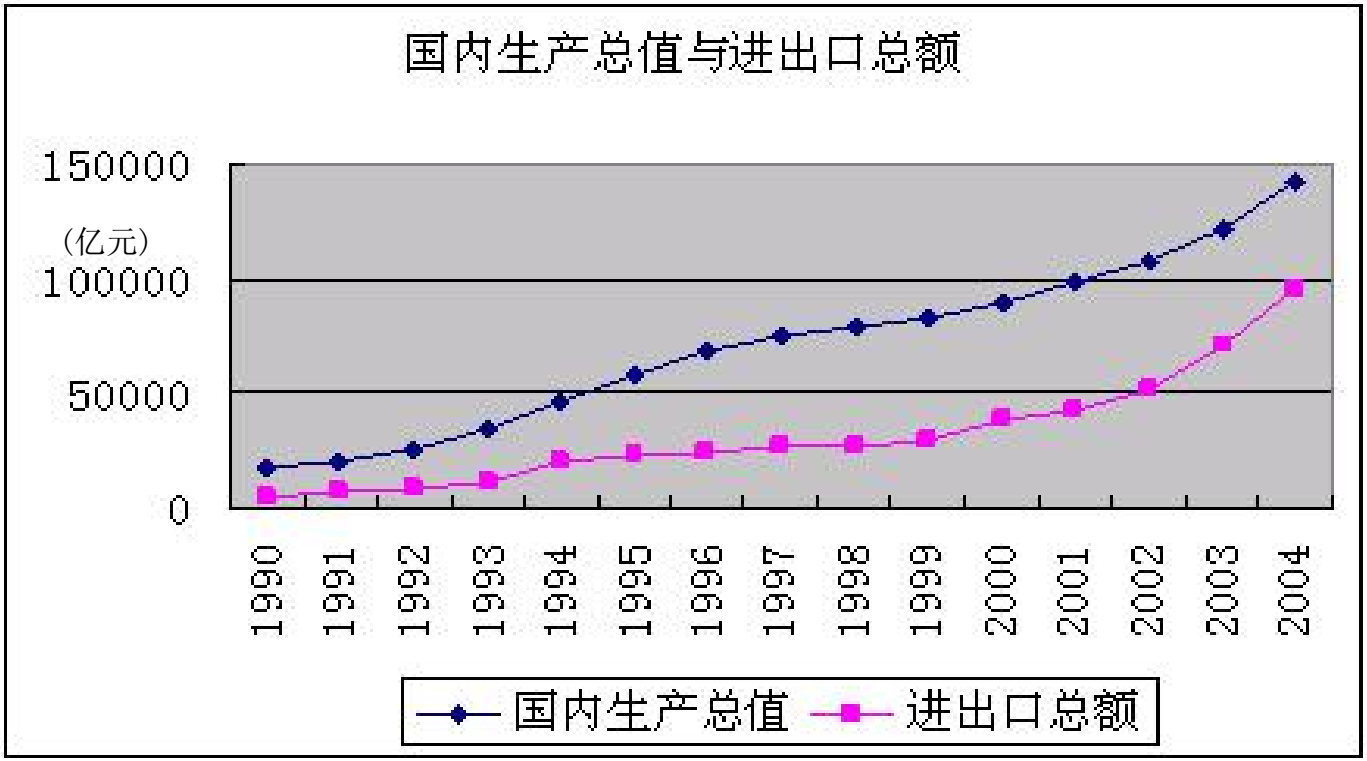
- **频率（Percent Frequency）:就是频数除以总数n：f_i/n，经常以百分数表示。**

2、统计图

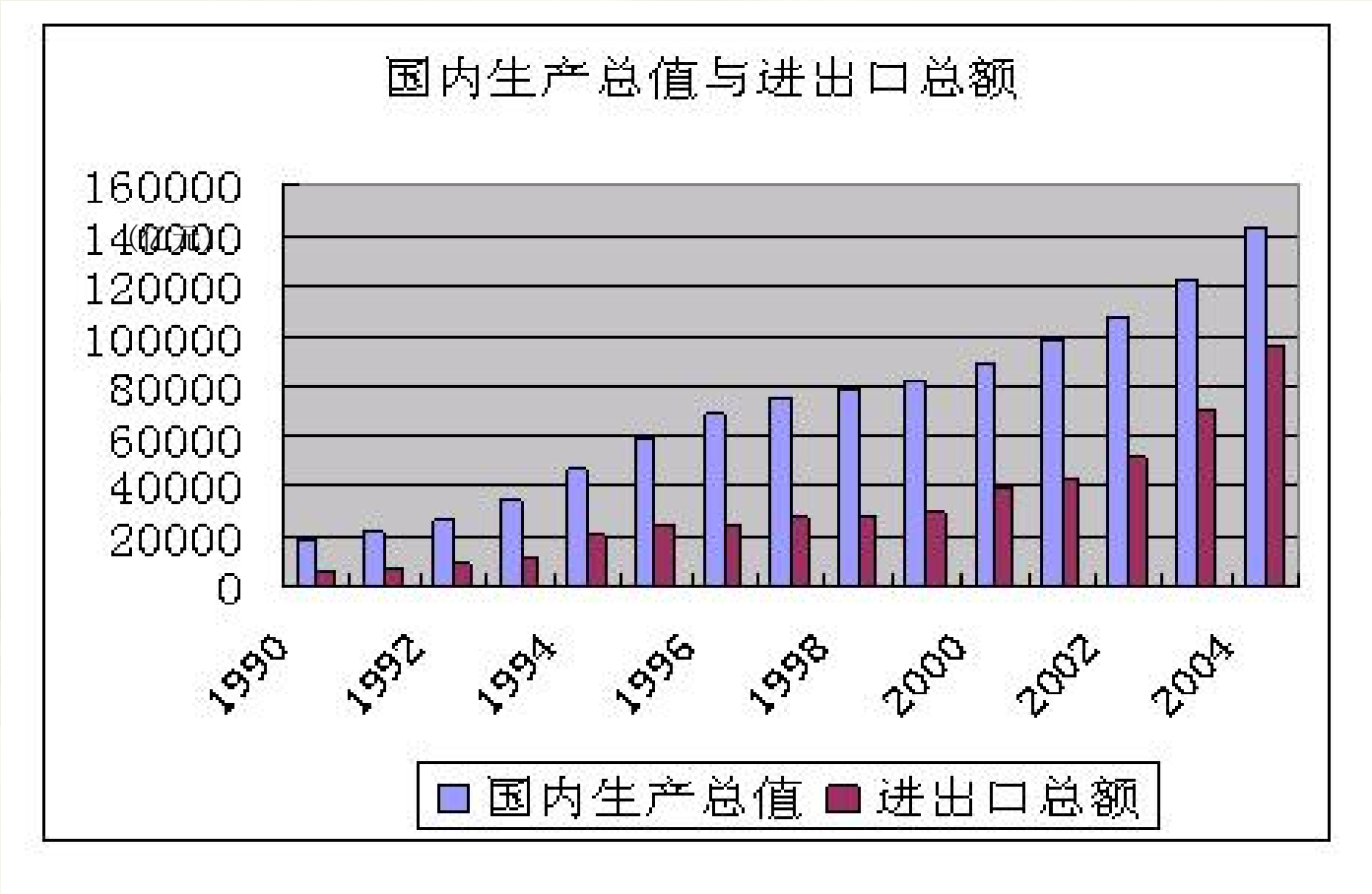
- **统计图：统计资料整理成统计表后，可以比较清晰地展示变量的变化规律。为了使这种规律更有直观性，常采用统计图表示。包括：线图 、条形图 、圆饼图等**



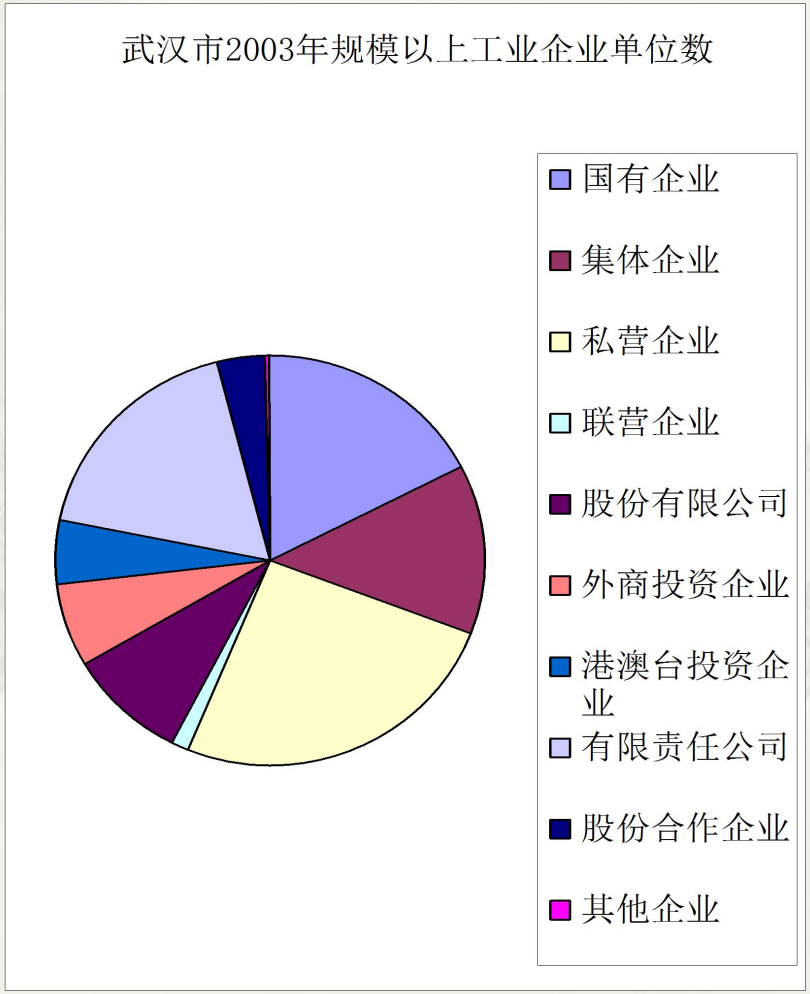
线图（Line graph）



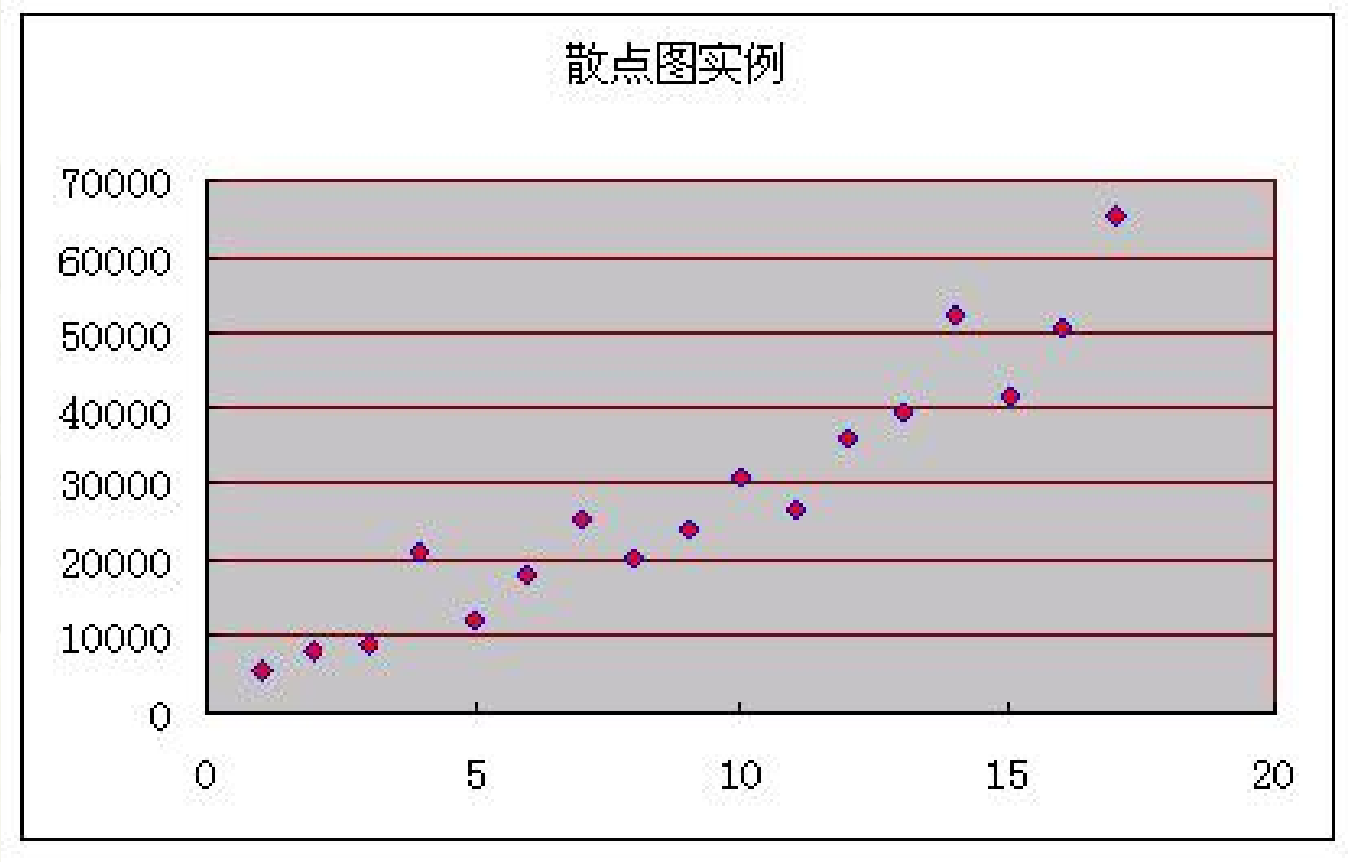
条形图 (Bar chart)



圆饼图（Pie chart）



- 散点图（Scatter Diagram）



3、 双变量的二元分布

- 双变量的统计资料：对每一元素观测两个特征，记录观测结果，就是双变量的统计资料
- 双变量常用（X，Y）形式表示，以区别两个单变量X和Y

[例]

在飞行模拟训练时，用计算机测定并打印出飞行动作的错误，从两方面进行测定：

- **错误发生时的飞行状态，分起飞（T），巡航（C）和着陆(L)三种。**
- **错误发生的原因，分规范理解错误(R)，仪表读数错误(M)和其它原因(O)三种。**

- 测定45次的打印记录如下：

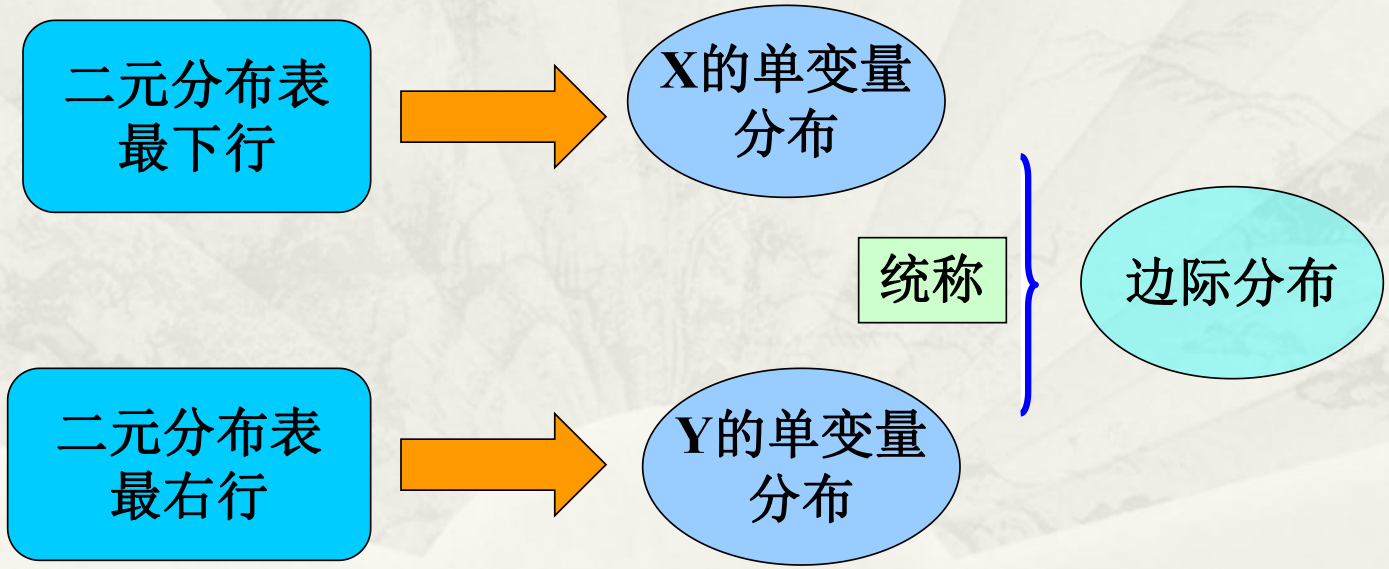
TM	TO	LM	LO	CO	LM	TR	CM	TM
LO	TM	CO	LR	CM	TR	LO	TR	LO
CO	LO	LM	TM	TO	CM	TO	LM	TO
CR	CM	TM	TR	LR	TM	LR	TR	TM
LM	TR	TR	LO	CR	TR	LO	LM	TM

- 根据该记录整理的二元分布表如下：

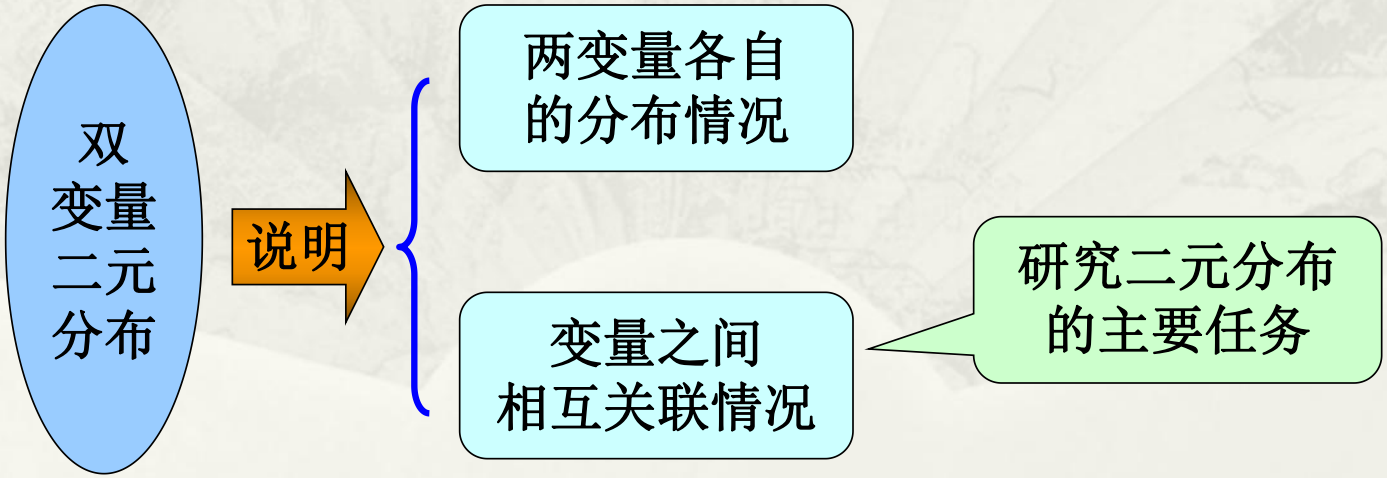
		错误原因			合计
		R	M	O	
飞行状态	T	8	8	4	20
	C	2	4	3	9
	L	3	6	7	16
合计		13	18	14	45

从表中看出，在起飞（T）时容易发生规范理解错误（R）和仪表读数错误（M），而着陆（L）时不太容易发生规范理解错误。

- 边际分布：在二元分布表最下行（合计行）和最右列（合计列）分别是X和Y的单变量分布，称为边际分布。



一个双变量的二元分布绝不同于两个单变量的一元分布，它不仅说明两变量各自的分布情况，而且说明两变量之间（飞行状态与错误原因之间）的相互关联情况。而这种关联情况（即是否存在关联以及关联的性态和程度等）正是研究双变量的二元分布的主要任务。



第四节　数据的描述性统计

一、数据分布的集中趋势

1 、平均数

- 算术平均数（Arithmetic average）
- 几何平均数（Geometric Mean）
- 调和平均数

(1)算术平均数（Arithmetic average）

* **定义：**

一组n个观测值x_1, x_2 ，…，x_n的算术平均数，定义为

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(1)算术平均数（Arithmetic average）

* 如果资料已经分组，组数为k，用x_1, x_2 ，…， x_k 表示各组中点，f_1，f_2…,f_k 表示相应的频数，那么

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

(2)几何平均数（Geometric Mean）

- 在数据为环比类型的问题中，算术平均数是不适用的。例如下表是天津市工业总产值在“十五”期间的逐年增长率，如求该期间平均增长率，算术平均数是不恰当的。几何平均数可以解决这个问题。

(2)几何平均数（Geometric Mean）

- **定义：一组n个数据的几何平均数定义为**

$$G = \sqrt[n]{r_1 r_2 \cdots r_n}$$

在上式中，$r_1, r_2, \cdots, r_5$ 依次为**114.0，119.6，124.1，131.0，120.8**于是几何平均数：

$$\sqrt[5]{114.0 \times 119.6 \times 124.1 \times 131.0 \times 120.8} = 121.8$$

十五期间天津市工业总产值年均增长率为**21.8%**。

(3)调和平均数

- 定义：

一组n个数据的调和平均数H，由下式定义

$$\frac{1}{H}=\frac{1}{n}\left(\frac{1}{R_1}+\frac{1}{R_2}+\cdots+\frac{1}{Rn}\right)$$

在上例中，

$$\frac{1}{H}=\frac{1}{2}\left(\frac{1}{20}+\frac{1}{30}\right)=\frac{1}{24}\ ,H\ =\ 24$$（公里/小时）

2、 众数（Mode）

- 算术平均数表示了集中位置特征，它照顾到每一个值，但它不见得是出现次数最多的值（甚至也可能不是观测值中的一个）。所以有必要研究表示集中位置的其它的特征数。

- 定义：对于有频数分布的变量，它的众数指频数最大的变量的值。

- 对于已分组且等组距的频数分布，根据最大频数，可求得众数所在组。根据众数定义，可知众数不唯一。

3、中位数（Median）

- 算术平均数作为集中位置的特征还有一缺点，就是受观测值中极端值的影响很大，而一组观测值中的极端值常常没有代表性。中位数将避免这种影响。
- 定义：一组n个观测值按数值大小排列，处于中央位置的值称为中位数以 Me 表示。

$$Me=\begin{cases} x_{\frac{n+1}{2}} & \text{，当n为奇数} \\ \frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right) & \text{，当n为偶数} \end{cases}$$

二、数据分布的离散趋势

1、极差（或称全距 Range）R

$$R = x_{\max} - x_{\min}$$

* **定义**

其中x_{max}和x_{min}分别为数据中的极大值和极小值。

2、 平均差（Mean Absolute Deviation）

定义

平均差M.D.是离差的绝对值的平均数，即

$$M.D. = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

对于已分组的频数分布（组数为k）

$$M.D. = \frac{1}{n}\sum_{i=1}^{k} f_i\left|x_i - \bar{x}\right|$$

3、方差（Variance），标准差（Standard Deviation）

* **方差**

总体 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

样本 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

对于已分组的频数分布（组数为k）

总体 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2$

样本 $S^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2$

标准差

总体标准差 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

样本标准差 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

对于已分组的频数分布（组数为k）

总体标准差 $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}$

样本标准差 $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}$

标准差的单位与X的单位相同。

4、变异系数（Coefficient of Variation）

* **定义：　变异系数C**

$$C = \frac{s}{\bar{x}} \times 100\,(\%)$$

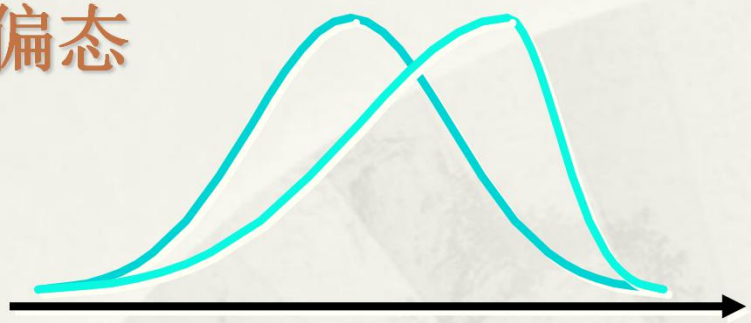
是一个无量纲的量。它适于用在比较有不同算术平均数或有不同量纲的两组数据的情况。例如比较大学生身高与小学生身高，或比较130名大学生身高和体重哪个变化波动范围比较大时，都可用变异系数。

三、数据分布的形状测度

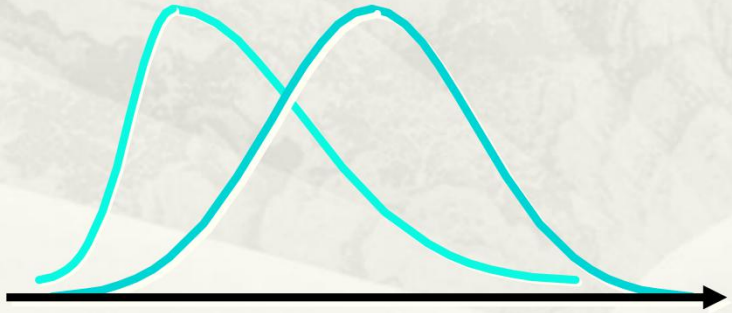
1. 偏态及其测度
2. 峰度及其测度

偏态与峰度分布的形状

偏态

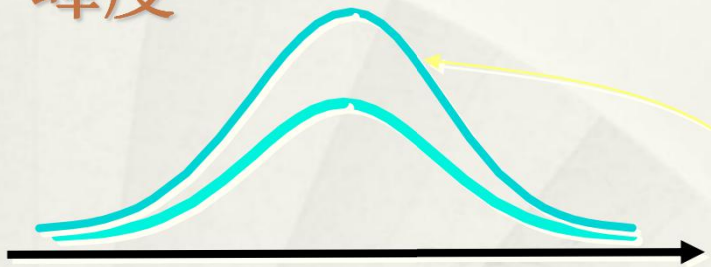


左偏分布

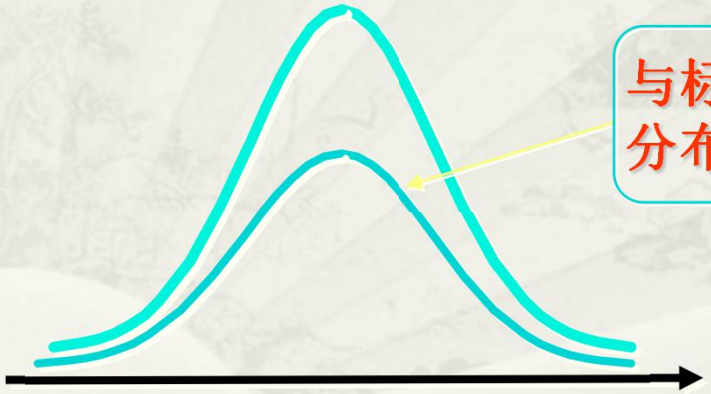


右偏分布

峰度



扁平分布



尖峰分布

与标准正态分布比较！

偏态
(概念要点)

- 1. 数据分布偏斜程度的测度
- 2. 偏态系数=0为对称分布
- 3. 偏态系数> 0为右偏分布
- 4. 偏态系数< 0为左偏分布
- 5. 计算公式为

$$\alpha_3 = \frac{\sum_{i=1}^{K}\left(X_i - \overline{X}\right)^3 F_i}{N\sigma^3}$$

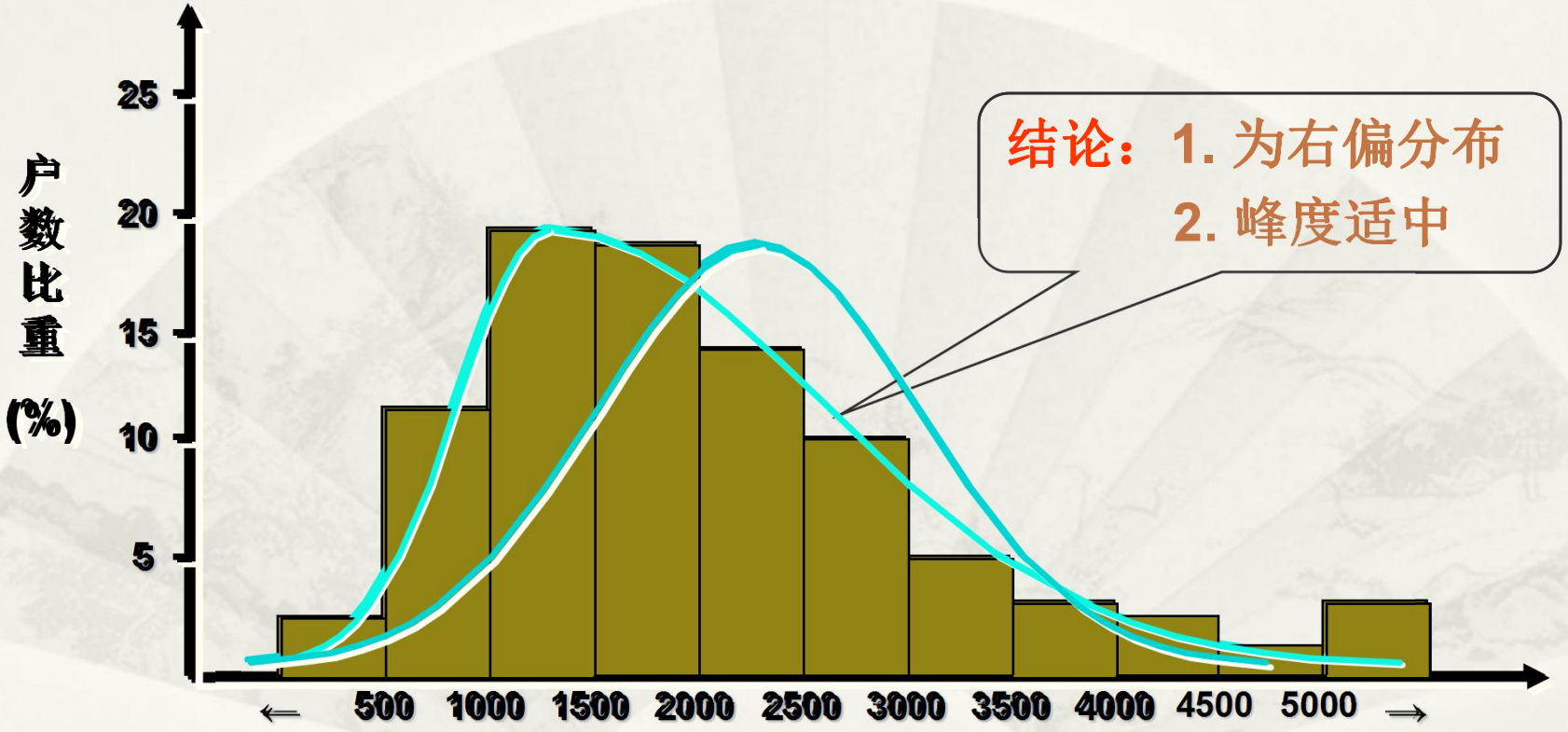
偏态
(实例)

【例】已知1997年我国农村居民家庭按纯收入分组的有关数据如表4。试计算偏态系数

表4　1997年农村居民家庭纯收入数据

按纯收入分组（元）	户数比重（%）
500以下	2.28
500~1000	12.45
1000~1500	20.35
1500~2000	19.52
2000~2500	14.93
2500~3000	10.35
3000~3500	6.56
3500~4000	4.13
4000~4500	2.68
4500~5000	1.81
5000以上	4.94

偏态与峰度
(从直方图上观察)



农村居民家庭村收入数据的直方图

偏态系数
（计算过程）

表5　农村居民家庭纯收入数据偏态及峰度计算表

按纯收入分组（百元）	组中值 X_i	户数比重(%) F_i	$(X_i-\bar{X})F_i^3$	$(X_i-\bar{X})F_i^4$
5以下	2.5	2.28	-154.64	2927.15
5—10	7.5	12.45	-336.46	4686.51
10—15	12.5	20.35	-144.87	1293.53
15—20	17.5	19.52	-11.84	46.52
20—25	22.5	14.93	0.18	0.20
25—30	27.5	10.35	23.16	140.60
30—35	32.5	6.56	89.02	985.49
35—40	37.5	4.13	171.43	2755.00
40—45	42.5	2.68	250.72	5282.94
45—50	47.5	1.81	320.74	8361.98
50以上	52.5	4.94	1481.81	46041.33
合计	—	100	1689.25	72521.25

偏态系数
(计算结果)

根据上表数据计算得

$$\overline{X} \doteq \sum_{i=1}^{K} X_i \bullet \frac{F_i}{\sum_{i=1}^{K} F_i} \doteq 21.429(\text{百元}) \qquad \sigma \doteq \sqrt{\sum_{i=1}^{K} X_i \bullet \frac{F_i}{\sum_{i=1}^{K} F_i}} \doteq 12.089(\text{百元})$$

将计算结果代入公式得

$$\alpha_3 = \frac{\sum_{i=1}^{K} (X_i - \overline{X})^3 F_i}{N\sigma^3} = \frac{\sum_{i=1}^{11} (X_i - 21.429)^3 F_i}{1 \times (12.089)^3} = \frac{1689.25}{1766.7339} = 0.956$$

结论：偏态系数为正值，而且数值较大，说明农村居民家庭纯收入的分布为右偏分布，即收入较少的家庭占据多数，而收入较高的家庭则占少数，而且偏斜的程度较大。

峰度
(概念要点)

* 1. 数据分布扁平程度的测度
* 2. 峰度系数=3扁平程度适中
* 3. 偏态系数<3为扁平分布
* 4. 偏态系数>3为尖峰分布
* 5. 计算公式为

$$\alpha_4 = \frac{\sum_{i=1}^{K}\left(X_i - \overline{X}\right)^4 F_i}{N\sigma^4}$$

峰度系数系数
(实例计算结果)

【例】根据表5中的计算结果，计算农村居民家庭纯收入分布的峰度系数

代入公式得

$$\alpha_4 = \frac{\sum_{i=1}^{K}(X_i - \overline{X})^4 F_i}{N\sigma^4} = \frac{72521.25}{1\times(12.089)^2} = 3.4$$

结论：由于=3.4>3，说明我国农村居民家庭纯收入的分布为尖峰分布，说明低收入家庭占有较大的比重.

第二章

SPSS 基本操作及数据处理

主要内容

第一节　SPSS软件概述

第二节　SPSS 数据文件的建立和管理

第三节　SPSS数据的预处理

第一节　SPSS软件概述

- SPSS的英文缩写：
  - Statistical Package for Social Science
  - Statistical Product and Service Solutions

- SPSS的发展：
  - 60年代：美国斯坦福大学三位研究生研制
  - 70年代：SPSS总部成立于芝加哥，推出SPSSX中小型机版
  - 80年代：SPSS公司（SPSS/PC+微机版1~3）
  - 90年代：SPSS公司（SPSS WINDOWS版5~16）
  - 2009:IBM收购，命名为：IBM SPSS Statistics(多国语言版23版)

SPSS 主要特点

- 操作简便。绝大多数操作是通过菜单、按钮、对话框完成的。
- 无需计算机编程、需记忆大量命令和参数。
- 分析方法丰富、分析结果清晰、直观。
- 可以直接读取其他软件格式的数据文件，如：xls、sas等。
- 最新版本采用分布式分析系统，适应互联网，支持动态收集、分析数据和HTML报告
- 不方便与一般的办公软件直接兼容

SPSS主要窗口：数据编辑器窗口

- 窗口标题：数据编辑器(数据集)
- 功能：对SPSS的数据文件进行录入、 修改、管理等基本操作的窗口。
- 组成：窗口主菜单、工具栏、数据编辑区、状态区
- 特点：
  - SPSS运行过程中自动打开
  - SPSS中各统计分析功能都是针对该窗口中的数据进行的
  - 窗口中的数据文件以.sav存于磁盘上
  - 两个视图：数据视图和变量视图

SPSS主要窗口：数据查看器窗口

- 窗口标题：查看器
- 功能：SPSS统计分析报表及图形的输出的窗口。
- 组成：窗口主菜单、工具栏、结果显示区、状态区
- 特点：
  - 输出窗口可以关闭，窗口内容以.SPV存于磁盘上
  - 两个部分：目录视图和内容视图

SPSS基本运行方式

* 完全窗口菜单方式:
  * 所有分析操作过程都是通过菜单和按钮及对话框方式进行的.

  * 是经常使用的一种运行方式,适用于一般分析和SPSS的初学者.

SPSS基本运行方式

* 程序运行方式:
  * 手工编写SPSS命令程序
  * 一次性提交计算机运行
  * 适用于大规模的分析工作和熟练的SPSS程序员.

* 实现方法:

(1)打开语法窗口并编写和修改SPSS程序

(2)点击语法窗口中的运行菜单项，选择运行方式运行

SPSS 基本运行方式

- 菜单程序混合运行方式:
  - 先通过菜单选择分析过程和参数,不立即提交（确定）执行,而是按粘贴按钮.
  - 计算机自动将用户刚定义的分析过程和参数转换成SPSS的命令,并显示到语法窗口中.
  - 用户可对其进行必要的修改后再提交给计算机执行.
  - 一般适用于熟练的SPSS程序员.

利用SPSS进行数据分析的步骤

- 建立SPSS数据文件
  - 定义数据文件结构
  - 录入修改和编辑待分析数据
- 数据的统计分析
  - 统计分析之前的预处理
  - 统计分析
- 数据和分析结果的保存
- 结果的说明和解释

第二节 SPSS 数据文件的建立和管理

SPSS数据文件

SPSS数据文件是一种有结构的数据文件

一份简单的调查问卷

- 单项选择题
  - 提供几个备选择答案，从其中选择一个答案
  - 一道问题对应一个SPSS变量
    - 变量类型：分类型、定序型、定距型
    - 数据类型：数值
- 多项选择题
  - 提供几个备选择答案，从其中选择多个答案
    - 例：在下列品牌中您信任哪些品牌？
    - 例：根据你的喜好给出以下你经常购物网站的序号
  - 多项选择题需对应多个SPSS变量，以后专门讨论

SPSS数据的结构

- 变量名(Variable name):变量存取的唯一标志。
  - 默认变量名为VARn(如var00001)
- 变量类型(type)与显示宽度(width)
  - 标准数值型(Numeric):默认类型　8.2
    如：　12345678、12345.67、-1234.56
  - 带逗号的数值型(Comma)：从个位开始三位一个逗号8.2
    如：1,234.56
  - 科学计数法(Scientific Notation):表示很大或很小的数据 8.2
    如：1.2E+05
  - 带美元符号(Dollar):表示货币
    格式很多，如：$12.30

SPSS数据的结构

- 变量类型(type)与存储宽度(width)
  - 字符型(String)：存储字符数据 8位
    如：beijing 处理时用双引号扩起来
  - 日期型 (Date):存储日期数据
    格式很多，如：20-AUG-1999
  - 其他:
    - 如：圆点数值型(dot)等

SPSS数据的结构

- 变量名标签(Variable label)
  - 对变量名的一些解释说明，增强分析结果的可视性。可以省略
- 变量值标签(Value label)
  - 对变量所取值的一些解释说明，增强分析结果的可视性。可以省略

SPSS数据的结构

- 变量列格式(Column Format)
  - 对齐方式(Text Alignment)：左对齐(Left):字符型默认；右对齐(Right):数值型默认；居中对齐(Center)
  - 列宽度(Column Width)：默认值为变量的存储宽度
    - 列宽度不影响存储宽度
- 计量尺度(Measurement)
  - Scale:定距;
  - Ordinal:有固有顺序;
  - Nominal:无固有顺序

SPSS数据的结构

- 缺失值(Missing Values)
  - 缺失值：漏填数据；明显错误的数据
- SPSS的用户缺失值：
  - 指定某个特定值为缺失值
  - 一般处理
    - 事先指定：指定某个特定值为用户缺失值
    - 修正：以均值、众数替代等
- SPSS的系统缺失值：
  - 数值型：点（•）；
  - 字符型：空

定义SPSS数据结构

- 操作方法：
  - 利用变量视图

SPSS数据的录入与保存

- 录入时应注意：
  - 黄框单元当前数据单元。
  - 录入带有变量值标签的数据：
    - 手工输入变量值
    - 打开值标签开关：屏幕显示变量值标签，从下拉框中选择。

SPSS数据的录入与保存

- 数据保存格式：
  - (1)*.sav ：SPSS数据文件(默认)。
  - (2)*.xls ： Excel工作表文件。
    - 注意：有些信息会丢失

SPSS数据的编辑

(一)打开数据文件

菜单选项:文件 -> 打开 -> .sav

(二)数据定位

- 按个案号码定位

菜单：编辑->转至个案-> 输入样本号

- 按值定位

光标定位到某列变量上 -> 编辑-> 查找

SPSS数据的编辑

(三)插入和删除一个个案

- 插入：编辑-> 插入个案
- 删除：选定待删行，鼠标右键选择清除

(四)插入和删除一个变量

- 插入：光标定位到某列变量上 -> 编辑 -> 插入变量(插到某列前) 或鼠标右键选择菜单
- 删除：选定列，鼠标右键选择清除

SPSS数据的编辑

(五)数据移动、复制和删除

* 定义源数据块
* 鼠标右键：选择相应菜单项
* 确定目标单元
* 鼠标右键：选择相应菜单项

与其他软件数据共享

- 数据共享
  - xls格式文件的共享
    - 是否有存放变量名的单元
  - 文本数据的读入
    - 利用文本向导读入数据
  - 数据库文件的共享
    - 利用ODBC共享数据

SPSS数据文件的合并

- 目的:
  - 将两个SPSS数据文件合并到一个数据文件中。
- 文件合并的方式:
  - 纵向合并
  - 横向合并

SPSS数据文件的合并

(一)纵向数据合并

(1)含义:

- 将磁盘或其他数据编辑器窗口中的SPSS数据追加到当前数据编辑器窗口中的数据文件中。

(2)前提:

- 两个SPSS数据文件应可以合并的内容，且最好有相同的变量名和变量类型。

(3)菜单选项:

数据 -> 合并文件 -> 添加个案

SPSS数据文件的合并

(二)横向数据合并

(1)含义:

将磁盘或其他数据编辑器窗口中的SPSS数据中的若干个变量增加到当前数据编辑器窗口中的数据文件中。

(2)前提:

a.两个数据文件必须有一个共同的变量名为关键字段---合并的依据;

b. 两个数据文件应事先按关键字段升序排序。

SPSS数据文件的合并

(二)横向数据合并

(3)菜单选项:

数据 -> 合并文件 -> 添加变量

(4)选项说明:

* 以关键字作为合并标志。
* 合并后的文件的数据由两个文件共同提供。
* 以当前数据编辑器中的数据为基础添加。
* 以磁盘文件或其他编辑器窗口中的数据为基础添加。

第三节　SPSS数据的预处理

数据排序

- 目标：排序在数据分析中的作用？
  - 快速找到可能的离群点
- 手段：将所有个案按照用户指定的某一个或多个变量的变量值的升序或降序重新排列
- 菜单选项：

  数据 -> 排序个案
- 注意：

  (1)排序的次序：升序、降序。

  (2)多重排序,选择变量名的次序很关键。

变量计算

* 目的：产生新变量或对原变量进行必要的转换

(如:预测问题　产生比率数据 偏态数据的正态处理　时间序列的平稳处理等)

(1)含义:根据用户给出的SPSS算术表达式，对所有或部分样本数据进行加工。

(2)菜单选项：

转换-> 计算变量；　如果按钮

(3)SPSS算术表达式：

* 由算术运算符(+ 、-、 *、 /、 **)、SPSS函数以及SPSS变量名组成的式子。

变量计算

(4) SPSS函数

- 算术函数
- 统计函数
- 分布函数
- 逻辑函数
- 字符串函数
- 缺失值函数
- 日期时间函数
- 其他函数

- Abs() sqrt() rnd() trunc() mod()
- mean() sd() sum() cfvar() max() min()
- normal() uniform() rv.() cdf.() idf.()
- range() any()
- index() length() lower() lpad() ltrim() substr()
- missing() sysmis()

变量计算

(5)SPSS条件表达式:由SPSS关系运算符、逻辑运算符、SPSS函数以及SPSS变量名组成的式子。

- 关系运算符: > (大于)、<(小于)、=(等于)、~=(不等于)、>=(大于等于)、<=(小于等于)
  - 如: nl>32、sr<=700
- 逻辑运算符:&(AND):并且、|(OR):或者、~(NOT):非
  - 如: (nl>32) and (sr<=700)
  - 如: (nl=32) | (sr<>700)
  - 如: not xb=1

数据分组

- 目标:更好地了解连续型变量的分布特点
- 手段：组距分组
  - 指定按哪个变量分组;定义分组区间(不重不漏);指定存放分组结果的组标志变量
- SPSS的区间
  - 狭义区间：
    - 职工工资的分组（850以下，851至900,901至950,951至1000,1000以上）
  - 广义区间：
    - 用户缺失值的定义；变量类别的重新调整

数据分组

•性格打分（内向、一般、外向）

1、与生人交往会“自来熟”

（1）从不　（2）偶尔　（3）有时　（4）经常

2、与不熟悉的异性交往，会脸红

（1）从不　（2）偶尔　（3）有时　（4）经常

3、 在公众场合下你会大声发表自己的意见

（1）从不　（2）偶尔　（3）有时　（4）经常

•极为内向：3分；较为内向：6分；较为外向:9分；极为外向：12分

个案选取

- 目标：个案选取的意义？
- 手段：从现有数据中选出部分数据
  - 按条件选取；随机选取；选取指定区间中的样本

- 例：对住房调查数据
  - 挑出本市户口的样本
  - 随机挑出70%的样本

- 注意：以后的操作都针对选出的数据进行

计数

- 目标：
  - 例：学生成绩整体状况的分析
  - 例：住房满意程度的粗略分析
- 手段：对所有或部分个案，计算若干个变量中有几个变量的值落在指定的区域内，并将结果存入新变量中
- 例：
  - 学生成绩得优门次的整体状况分析
  - 住房满意程度的粗略分析

分类汇总

- 目标：分析各分组下样本的统计特征
- 手段：
  - 按指定的分组变量值对样本分组
  - 分别计算各组中汇总变量的基本统计量
- 例：对比男女职工的平均年龄和平均工资

性别	年龄	奖金
男	40	1000
女	35	550
男	20	200

原始数据

性别_1	年龄_1	奖金_1
男	30	600
女	35	550

按性别变量汇总数据

分类汇总

* 菜单选项:
  数据 -> 分类汇总

* 说明:
  * 多重分组时，变量名的选择顺序。
  * 生成的新文件名默认为:aggr.sav。可修改。
  * 生成的新变量名默认为原变量名后加_1。可修改
  * 可以在新文件中存贮个分组个案数.

指定加权变量

- 目标：
  - 例：蔬菜的平均价格、男足打分
- 手段：指定某一变量为加权变量
  - 例：蔬菜的平均价格
- 菜单选项：

  数据 -> 加权个案
- 说明：
  - 如果取消加权变量应重新定义

第三章

SPSS基本统计分析

主要内容

* 第一节　频数分析
* 第二节　计算描述统计量
* 第三节　列联分析
* 第四节　多选项分析

第一节 频数分析

- 目的：粗略把握变量值的分布状况。
  - 例：研究被调查者的特征（如:性别,年龄,收入）
  - 研究被调查者对某个问题的总体看法（如:教学方式,选修课程）

- 采用的方法
  - 计算频分布表：包括频数、累计频数、百分比、累计百分比
  - 绘制统计图形：条形图、饼图

频数分析

* 基本操作步骤

(1)菜单选项:分析->描述统计->频率
(2)选择几个待分析的变量到变量框.
(3)图表选项，选择所需要的图形

频数分析

- 频数分析中的其他分析
  - 计算分位数:适用于定距数据
    - 数据按升序排序后，找到若干个分位点上的变量值
    - 计算四分位数：25%(Q_L)、50%(中位数)、75%(Q_U)
  - 分位数的应用：在排除极端值影响的条件下，通过计算分位数差，比较两组样本数据的离散程度
    - 例:(Q_L=50,Q_U=80)和(Q_L=70,Q_U=75)的比较

频数分析

- 例：对于住房调查数据
  - 住房满意程度的分析
    - 条形图和饼图的编辑
  - 购房意向的分析
    - 有效百分比
  - 现住房面积的分位数分析
  - 住房满意程度与现住房面积的分位数对比
    - 数据拆分处理

与频数分析相关的图形

* 以制作条形图为例

绘制简单条图（单式条图）

绘制复式条图

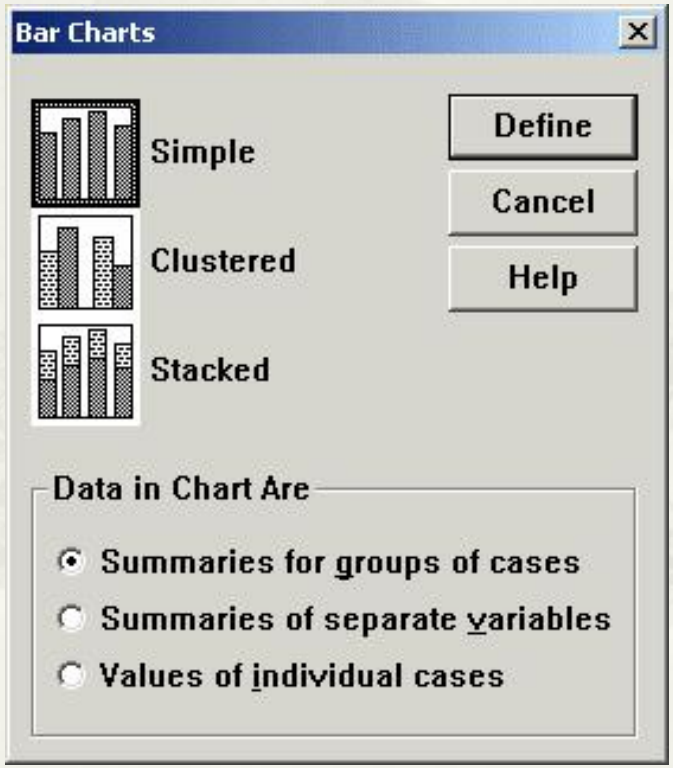
绘制堆积条图（分段条图）

定义统计图中数据的表达类型：

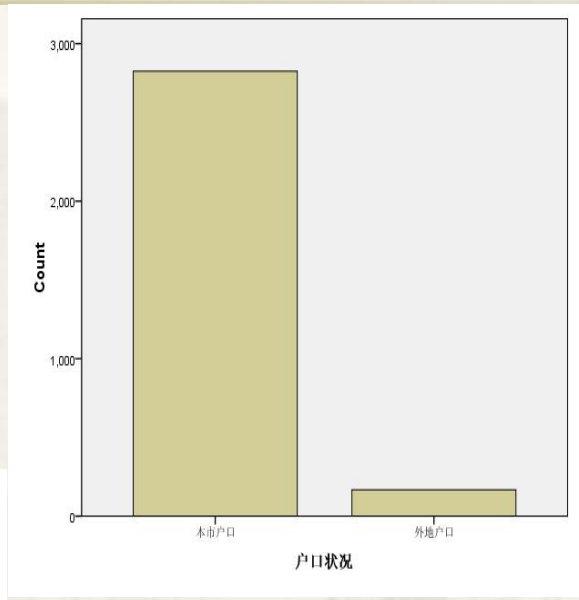
同一变量若干条记录的分组汇总

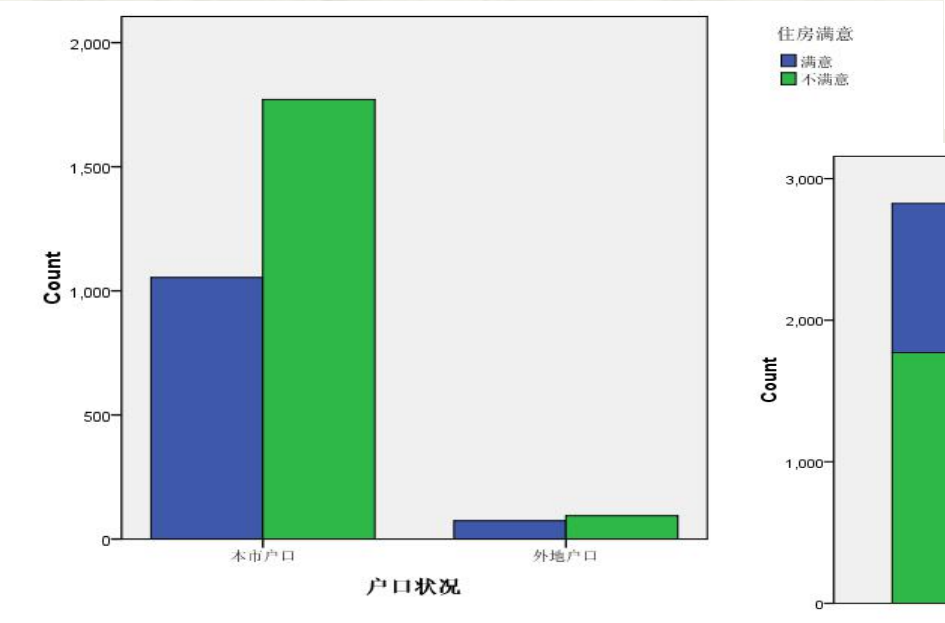
条图反映了不同变量的汇总

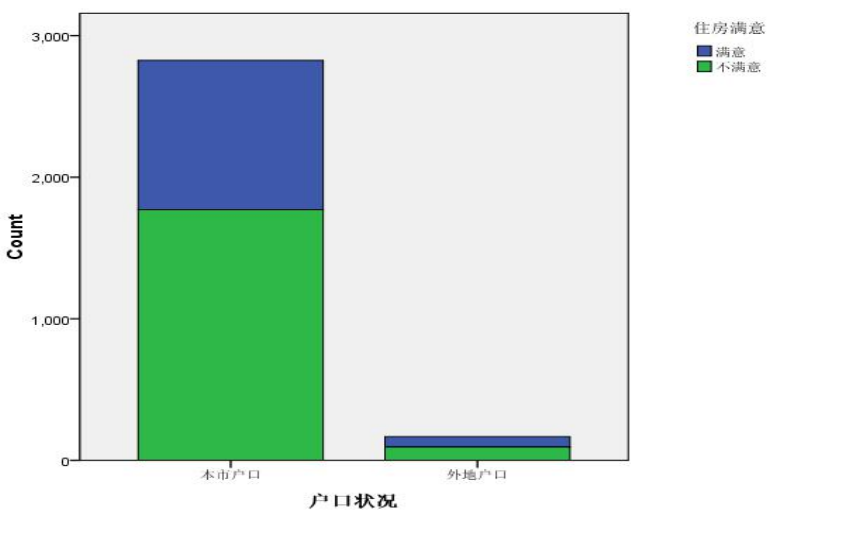
条图反映了个体观察值



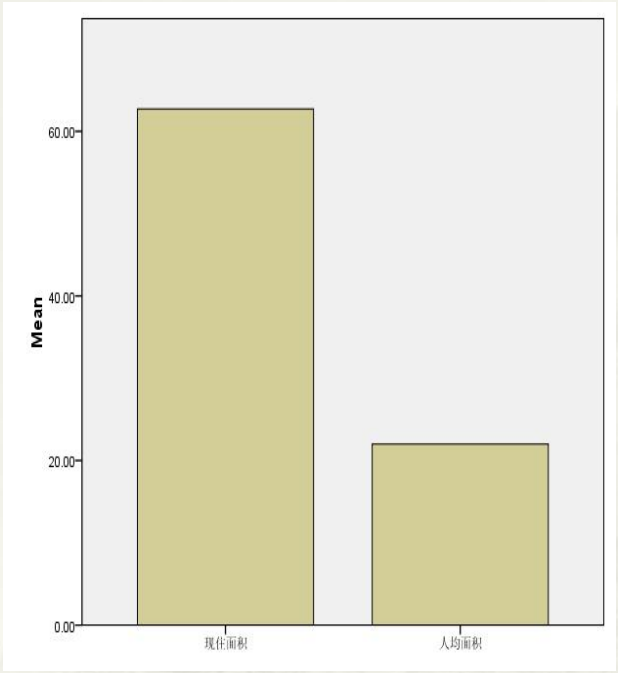
* 第一种模式：用于变量在各组下的频数对比

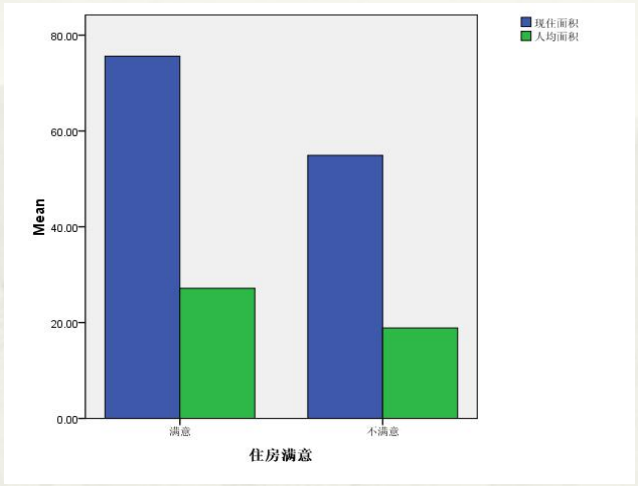


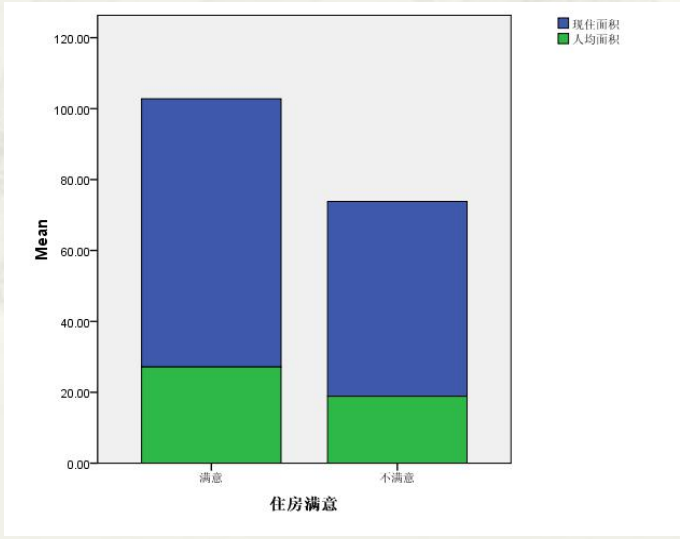




* 第二种模式：用于多个变量基本描述统计量的对比

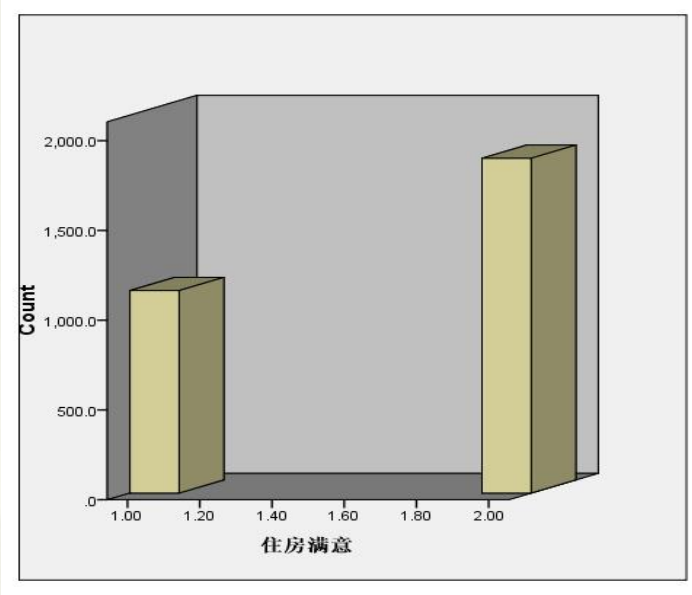


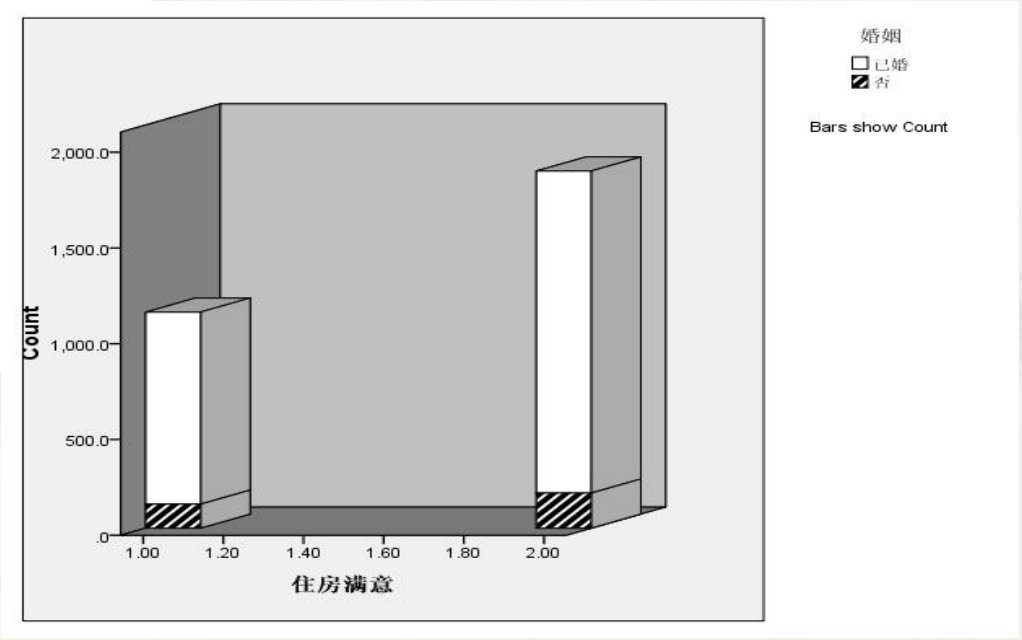




与频数分析相关的图形

* 交互作图：以制作条形图为例





第二节 计算描述统计量

- 目的：精确把握变量的总体分布状况，了解数据的集中趋势、离散趋势、对称程度、陡峭程度。

- 基本方法：
  - 计算基本描述统计量

- 描述集中趋势的统计量
  - 均值:表示某变量所有变量值集中趋势或平均水平的统计量。
    - 适用于定距数据。
    - 特点：利用了全部数据，易受极端值的影响。
- 描述离散程度的统计量
  - 标准差:表示某变量的所有变量值离散程度的统计量。
    - SPSS中计算的是样本标准差
  - 极差:最大值—最小值

* 描述对称程度的统计量
  * 偏度(skewness):描述某变量分布形态的偏斜程度和方向的统计量.
    * 偏度为0表示对称;
    * 大于0表示正偏差大(右偏
    * 小于0表示负偏差大(左偏)

* 描述陡峭程度的统计量
  * 峰度(kurtosis):描述某变量所有变量值分布形态陡缓程度的统计量。
    * 峭度为0表示与标准正态分布峭度相同。
    * 大于0表示比标准正态分布陡，尖峰。
    * 小于0表示比标准正态分布缓，平峰。

- 其他统计量
  - 均值标准误差(means of S.E)
    - 中心极限定理认为：样本均值~$N(u,\sigma^2/n)$
    - 反映样本均值与总体真值间的平均离散程度
    - 样本数越大，样本均值的离散程度越小，对真值的估计越准确

* 基本操作步骤

(1)菜单选项:分析->描述统计->描述

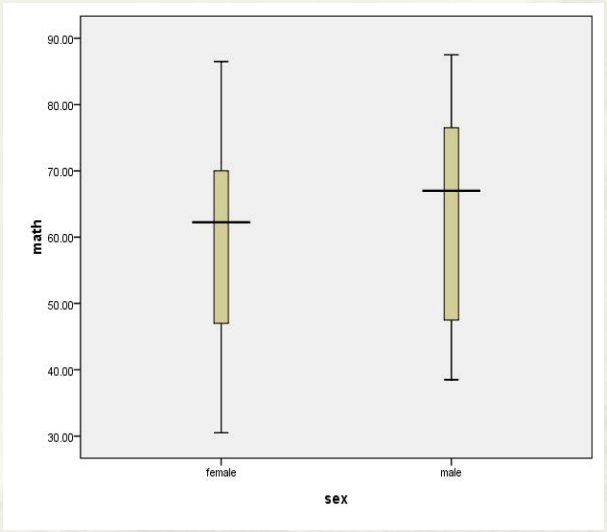
(2)选择将参加计算的数值型变量名到变量框

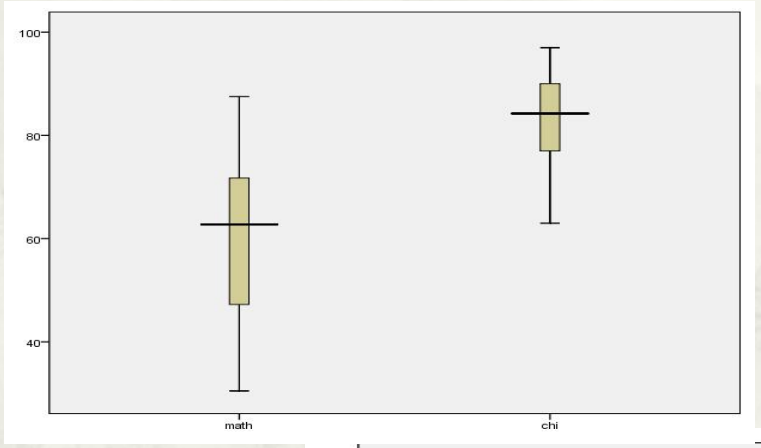
* 其他功能
  * 数据标准化处理
    * 新变量的均值为0,标准差为1;
    * 小于0表示在平均水平下,大于0反之.
    * 正态分布的数据标准化后呈标准正态分布
      * 3σ准则：（68.2%，95.4%，99.7%）
    * 将变量作标准化后,结果存入名为“Z+原变量名”的新变量中.

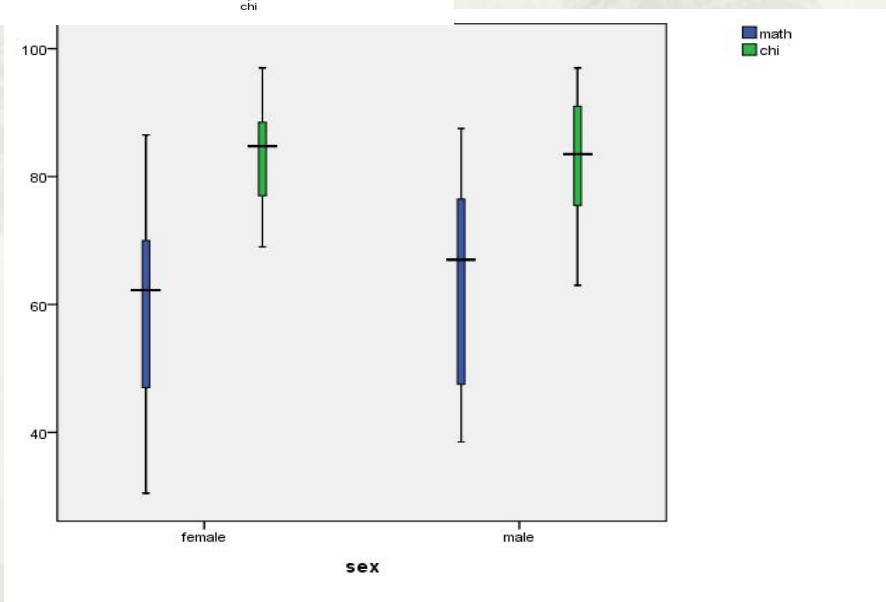
* 例：对于学生成绩
  * 分别对比男女生的数学成绩的基本描述统计量
  * Z分数的计算

描述连续变量分布的图形

* 箱线图：以四分位差的1.5倍为标准剔除极端值

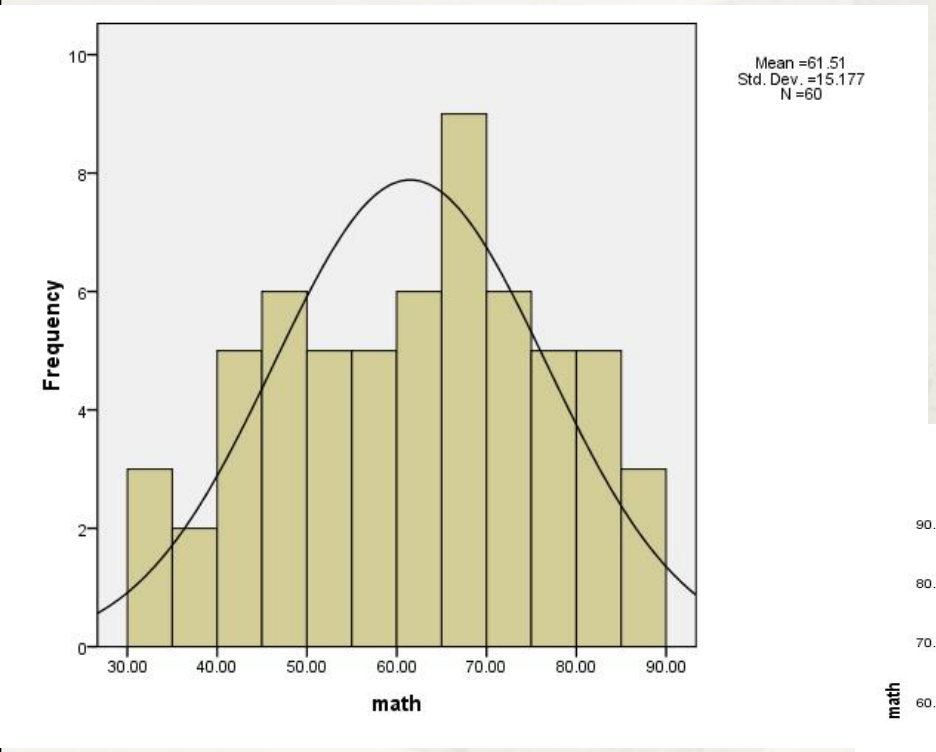


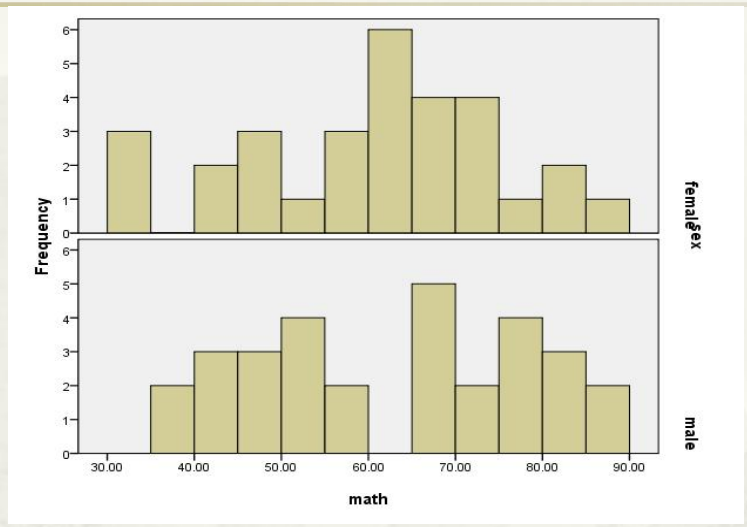


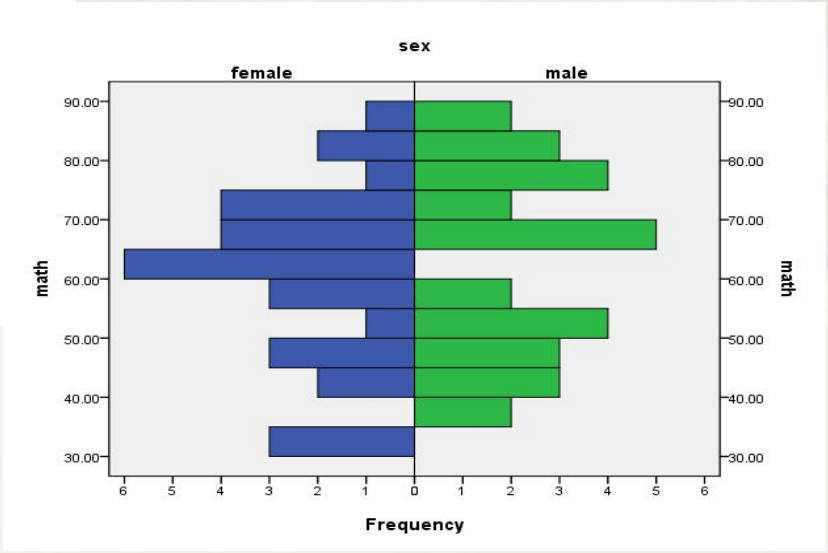


描述连续变量分布的图形

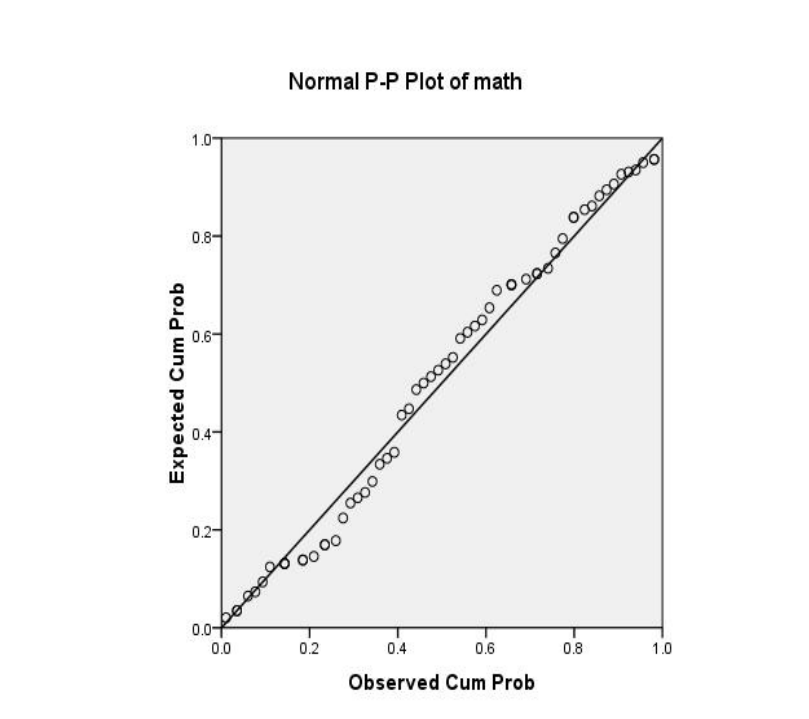
* 直方图和金字塔图

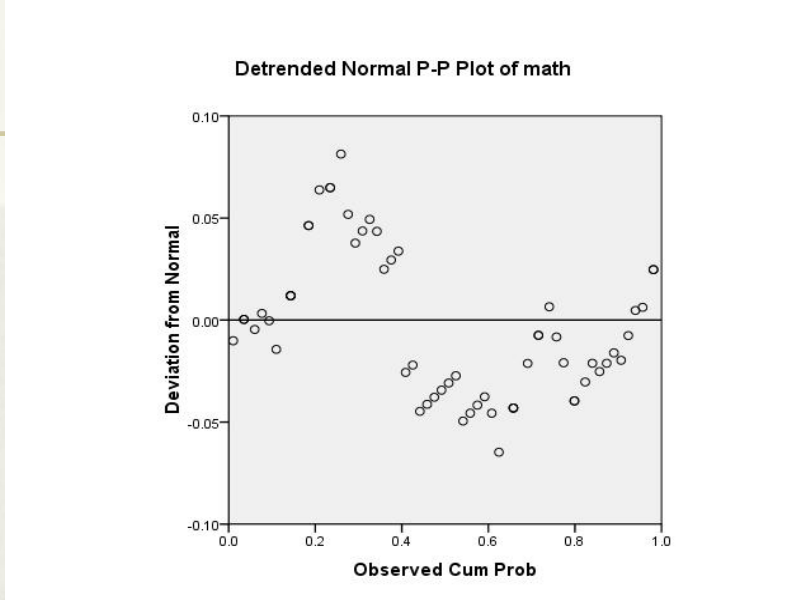


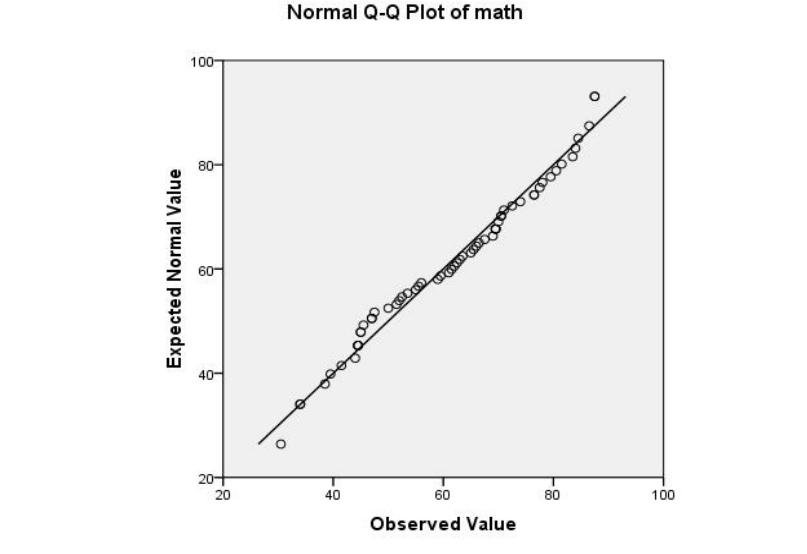




* Q-Q图和P-P图：累计分布函数(CDF)和概率密度函数(PDF)函数的应用







交叉分组下的频数分析

- 目的：了解不同变量在不同水平下的数据分布
  - 例：学习成绩与性别有关联吗？（两变量）
  - 例：职业、性别、爱逛商店有关联吗？（三变量）

- 分析的主要步骤
  - 产生交叉列联表
  - 分析列联表中变量间的关系

第三节 列联表

* 列联表中的元素：

职称	收入		
	高(人)	中(人)	低(人)
高工			
工程师			
助工			
技术员			
合计			

产生交叉列联表

* 基本操作步骤

(1) 菜单选项：分析->描述统计-> 交叉表

(2) 选择一个变量作为行变量到行框.

(3) 选择一个变量作为列变量到列框.

(4) 可选一个或多个变量作为控制变量到层框.

控制变量的层次设置:同层为水平数加;不同层为水平数积.

(5) 是否显示复式条形图

产生交叉列联表

- 进一步计算
  - 单元格选项:选择在频数分析表中输出各种百分比.
    - 行百分比；列百分比;总百分比

列联表

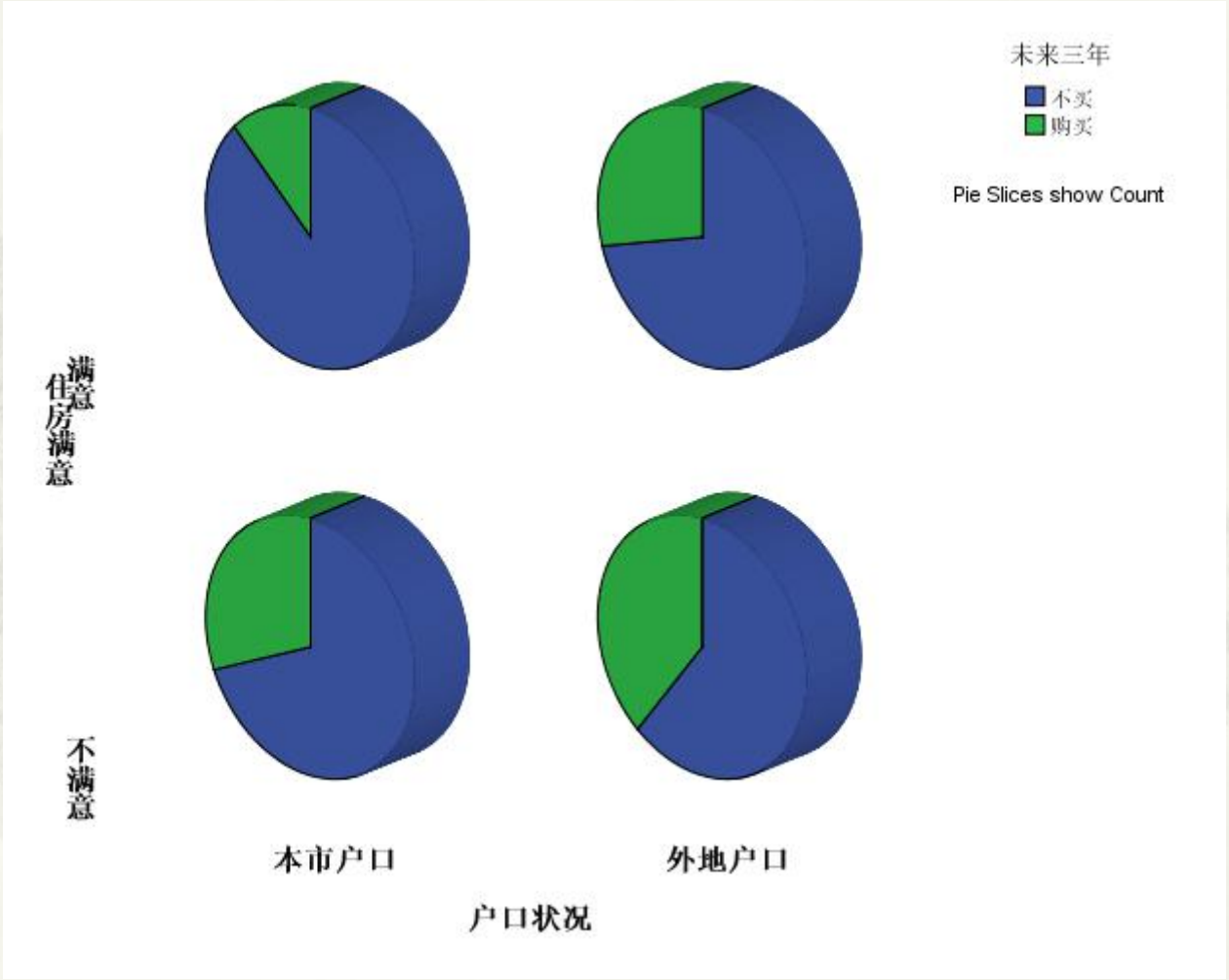
* 例：对于住房调查数据，满意程度与购房计划

住房满意 * 未来三年 Crosstabulation

			未来三年		
			不买	购买	Total
住房满意	满意	Count	958	157	1115
		% within 住房满意	85.9%	14.1%	100.0%
		% within 未来三年	44.3%	21.8%	38.7%
		% of Total	33.3%	5.5%	38.7%
	不满意	Count	1203	562	1765
		% within 住房满意	68.2%	31.8%	100.0%
		% within 未来三年	55.7%	78.2%	61.3%
		% of Total	41.8%	19.5%	61.3%
Total		Count	2161	719	2880
		% within 住房满意	75.0%	25.0%	100.0%
		% within 未来三年	100.0%	100.0%	100.0%
		% of Total	75.0%	25.0%	100.0%

与列联表相关的图形

* 交互绘图



列联表中行列变量间的关系

- 目的：通过列联表分析，检验行列变量之间是否独立
- 方法：卡方检验（分类变量相关性的检验）

年龄与工资收入交叉列联表

	低	中	高
青	400	0	0
中	0	500	0
老	0	0	600

	低	中	高
青	0	0	500
中	0	600	0
老	400	0	0

列联表中行列变量间的关系

- 卡方检验基本步骤

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

(1)H_0:行列变量独立

(2)构造卡方统计量:从(r-1)*(c-1)个自由度的卡方分布

- 期望分布反映的是H_0成立情况下的分布特征

(3)计算卡方的观测值,得到概率P值

(4)比较显著性水平和概率P值。小于等于则拒绝H_0,否则不能拒绝

	优	良	中	及格	总数
男	10	5	5	3	23
女	8	12	4	1	25
总数	18	17	9	4	48
	37.5	35.4	18.8	8.3	100

列联表中行列变量间的关系

- 例：对于住房调查数据，分析满意程度与购房计划的关系
- 例：不同行业的人职业选择标准是否存在差异？

–	制造业	服务业
物质报酬	105	45
稳定性	40	35

2乘2的列联表进行yates连续性校正：

$$\chi^2_{\text{yates}} = \sum_{i=1}^{n} \frac{(|f_i^o - f_i^e| - 0.5)^2}{f_i^e}$$

列联表中行列变量间的关系

- 卡方检验的要求：
  - 一般要求列联表中期望频数小于5的格子数不超过20%，否则会夸大卡方值，容易得出拒绝结论，可以合并单元格。
  - 卡方值会受样本数的影响

列联表中行列变量间的关系

- 行列变量相关性的其他测度指标
  - phi系数：适用于2×2列联表

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \frac{A_{11}A_{22} - A_{12}A_{21}}{\sqrt{R_1 R_2 C_1 C_2}}$$

    - 行列变量独立时(期望频数)：有：$\varphi = 0$

$$\frac{A_{11}}{C_1} = \frac{A_{12}}{C_2}, \frac{A_{21}}{C_1} = \frac{A_{22}}{C_2}$$

    - 行列变量完全相关时：$A_{12} = A_{21} = 0$ 有：$\varphi = 1$

A11	A12	R1
A21	A22	R2
C1	C2	

    - 越接近于1，相关性越强。越接近0，相关性越弱

列联表中行列变量间的关系

- 行列变量相关性的其他测度指标
  - 列联C系数(contingency coefficient)：
    - [0,1)；取值受到行列数的影响（见EXCEL）

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

  - V系数：[0,1]

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

  - 值越大表示行列变量的相关性越大

第四节 多选项分析

- 多选项分析是针对多选项问题的
- SPSS多选项问题的处理思路：
  - 将一个问题定义成几个变量。分别用几个变量描述问题的几个可能被选择的答案
- 具体策略：采用不同的编码方式
  - 多选项二分法(multiple dichotomize method)
    - 将每个答案作为一个变量，每个变量只有两个取值(0或1)
  - 多选项分类法(multiple category method)
    - 预先指定多选项问题被选择的最多答案数
    - 每个答案建立一个变量,取值为多选项问题的备选答案

多选项分析

- 多选项分析的基本思路
  - 定义多选项变量集
  - 多选项频数分析
  - 多选项交叉分组下的频数分析

多选项分析

- 定义多选项变量集
  - 目的:将已分解的变量定义为一个集合,便于进行多选项分析
  - 菜单选项:分析->多重响应->定义变量集
  - 从原变量中选取被分解的变量(数值型)到集合中的变量框
  - 指定被分解的变量是按多选项二分法分解还是按多选项分类法分解的
  - 为变量集命名。系统自动在名字前加字符$.

多选项分析

- 多选项频数分析
  - 菜单选项:分析->多重响应->频率
- 多选项交叉分析下的频数分析
  - 菜单选项:分析->多重响应->交叉表

多选项分析

- 例：对保险数据，分析购买养老保险的原因
- 例：三城市受访者对几种常见饮料的喜好情况：
  - 采用二分法组织数据
  - 受访人群中最受欢迎的饮料是哪种？
  - 男、女喜爱的饮料有无差异？
  - 三个城市的人群对饮料的喜好有无差异？

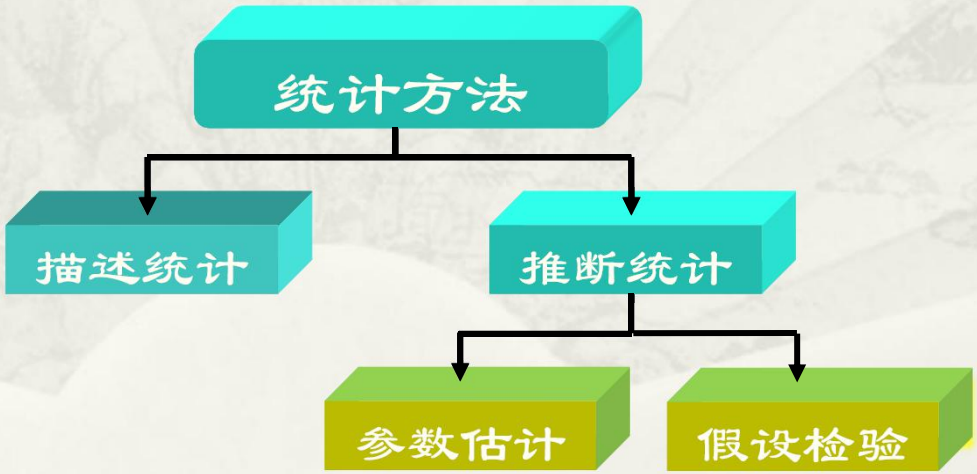
第四章

参数检验及方差分析

主要内容

* 第一节　假设检验概述
* 第二节　单样本t检验
* 第三节　独立样本t检验
* 第四节　配对样本t检验
* 第五节　单因素方差分析

- 统计学的范畴：推论统计
  - 根据样本数据推断总体的分布或均值方差等总体统计参数
  - 方法：
    - 参数检验
    - 非参数检验



第一节 假设检验概述

- 假设检验是一种根据样本数据推断总体的分布或均值、方差等总体统计参数的方法。
- 根据样本来推断总体的原因：
  - 总体数据不可能全部收集到。如：质量检测问题
  - 收集到总体全部数据要耗费大量的人力和财力
- 假设检验包括：
  - 参数检验
  - 非参数检验

假设检验的基本步骤

- 提出基本假设H_0
- 构造服从某种理论分布的检验统计量
- 利用样本数据和基本假设计算检验统计量的观测值，并得到概率P值（检验统计量在特定极端区域取值在H_0成立时的概率）
- 如果概率P值小于用户给定的显著性水平a，则拒绝H_0；否则，不拒绝H_0

假设检验的基本原理

- 基本信念：利用小概率原理进行反证明。小概率事件在一次实验中不可能发生。
- 例如：对大学男生平均身高进行推断
  - H_0：平均身高为173
  - 样本平均身高为178，由于存在抽样误差，不能直接拒绝H_0。而需要考虑：在H_0成立的条件下，一次抽样得到平均身高为178的可能性有多大。如果可能性较大，是个大概率事件（与α相比较），则认为H_0正确。否则，如果可能性较小，是个小概率事件，但确实发生了，则只能认为H_0不正确。
  - 概率P值即为观测结果或更极端现象在零假设成立时出现的概率

SPSS中的参数检验方法

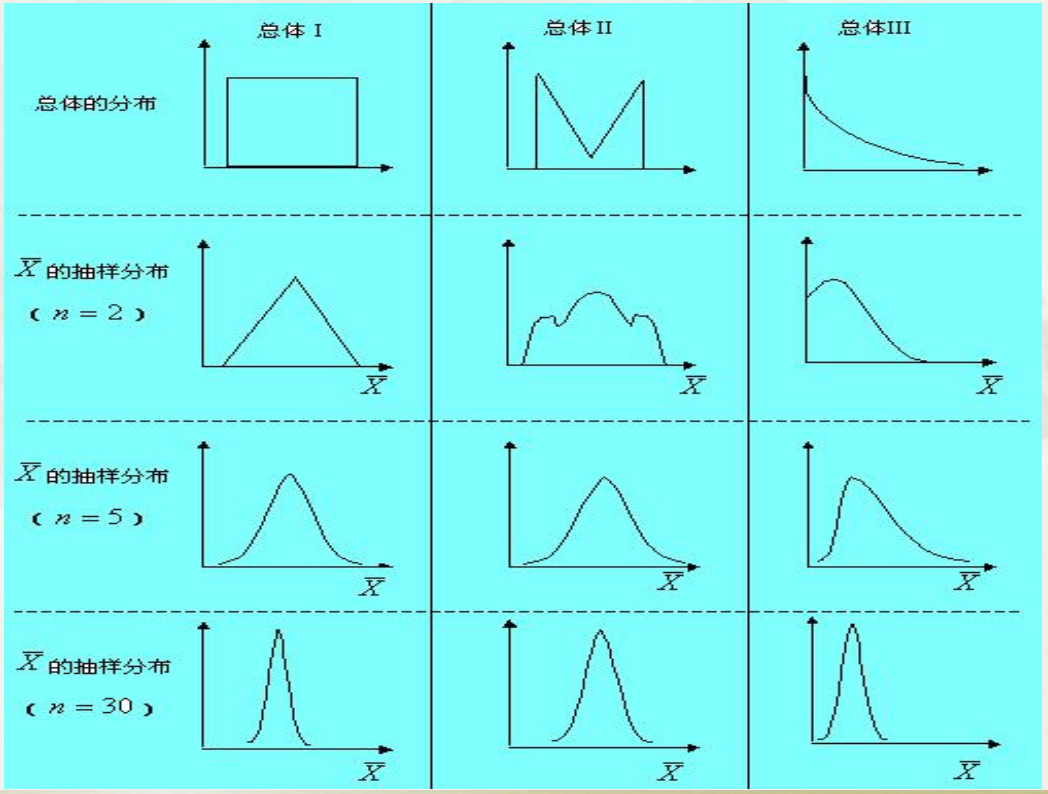
* 单样本t检验
* 两独立样本t检验
* 两配对样本t检验

第二节 单样本t检验

- 目的：对某个总体的均值与指定的检验值之间是否存在显著差异进行检验
  - 例：大学毕业生的月平均工资与3500元是否有显著差异
- 手段：利用单个样本的均值对总体均值进行检验
- 理论依据：样本均值的抽样分布
  - 抽样分布：样本统计量的概率分布
  - 结果来自容量相同的所有可能样本
  - 提供了有关样本统计量的概率信息，是推断的理论基础，是抽样推断科学性的重要依据

- 当总体服从正态分布$N\sim(\mu,\sigma^2)$时，来自该总体的所有容量为n的样本的均值 $\bar{X}$也服从正态分布， $\bar{X}$的数学期望为μ，方差为σ^2/n。即 $\bar{X}\sim N(\mu,\sigma^2/n)$

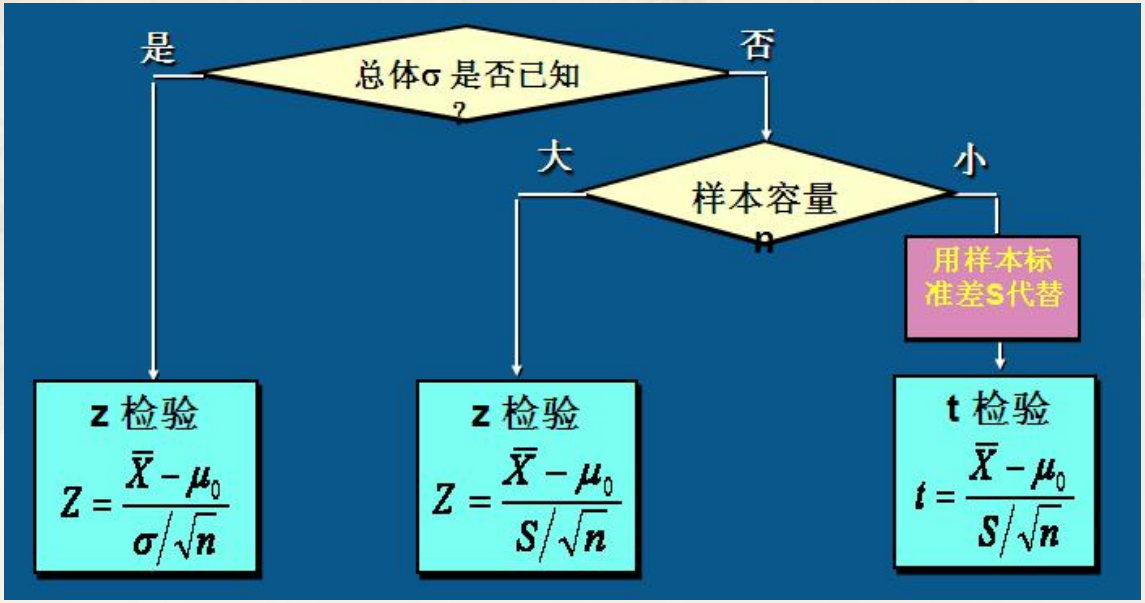
- 设从均值为μ，方差为σ^2的一个任意总体中抽取容量为n的样本，当n充分大时，样本均值的抽样分布近似服从均值为μ、方差为σ^2/n的正态分布



* 基本步骤：
  * $H_0: u=u_0$，总体均值与检验值之间不存在显著差异
  * 选择检验统计量
  * 计算t统计量的观测值和概率P值
  * 结论：$P \leqslant \alpha$，拒绝H_0，认为总体均值与检验值之间有显著差异. $P > \alpha$，不能拒绝H_0

•例：职工平均工资的检验

•注意：SPSS给出的双侧检验的概率P值



单样本t检验

* 基本操作步骤

(1)菜单选项:分析->比较均值->单样本T检验

(2)指定检验值：在检验值框中输入原假设值

单样本t检验

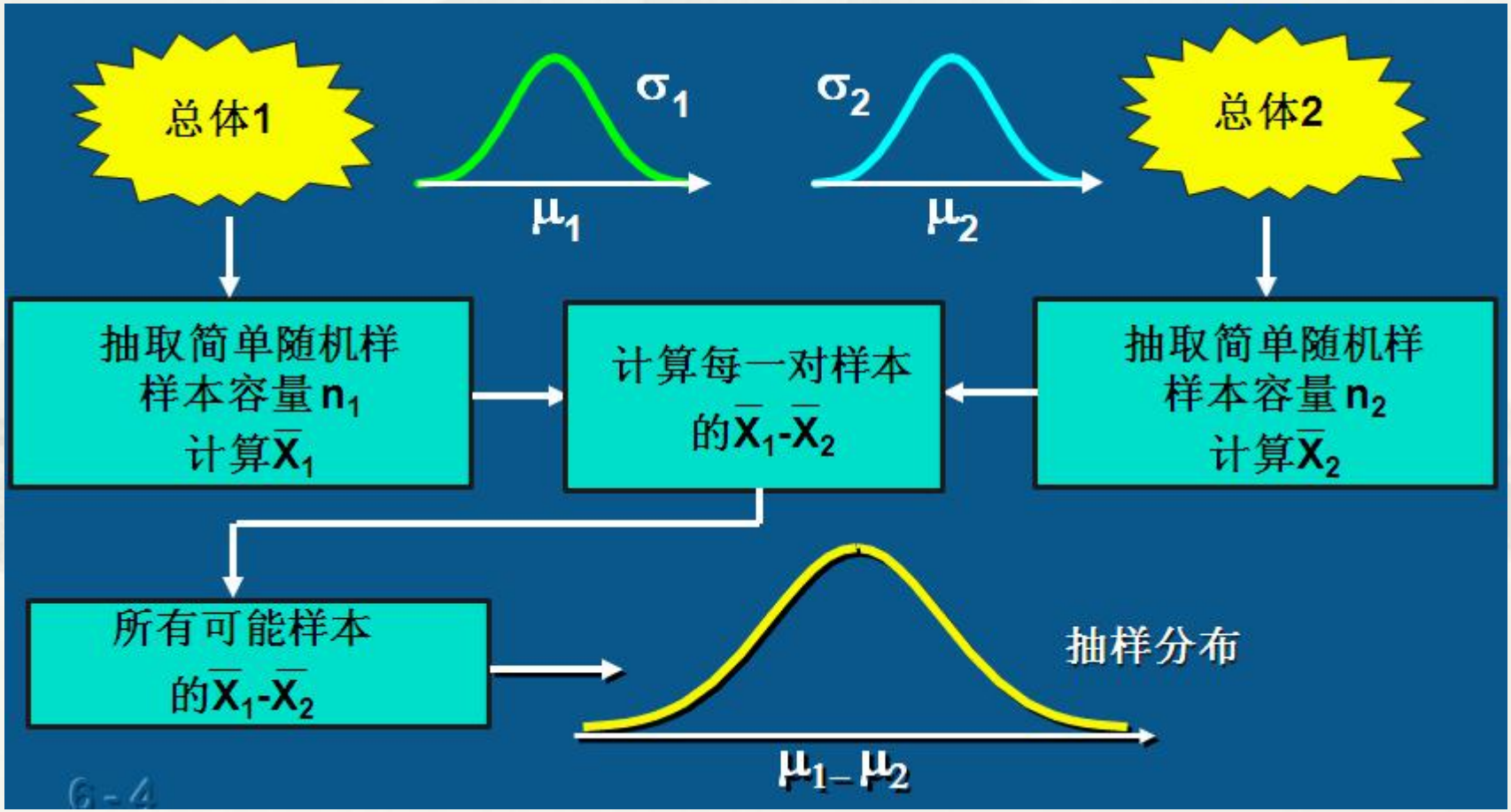
- SPSS中的选项
  - 置信区间:指定输出$\mu-\mu_0$的置信区间.默认值为95%.
  - 缺失值的处理策略
    - 当涉及缺失值变量的计算时剔除包含缺失值的样本
    - 剔除所有含缺失值的个案后再计算

第三节　两独立样本t检验

- 目的：对两总体的均值是否有显著差异进行推断
  - 例：男女生的月平均工资是否存在显著差异
- 手段：利用两个独立样本的均值差对两总体的均值差进行检验
  - 独立样本：抽取一个样本对抽取另一个没有影响
- 理论依据：两独立样本均值差的抽样分布

两独立样本t检验

* 理论依据：两独立样本均值差的抽样分布



- 理论依据：两独立样本均值差的抽样分布
  - 两总体方差已知：

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

  - 两总体方差未知且相等：

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sim t(n_1+n_2-2)$$

$$S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

  - 两总体方差未知且相等：

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}\sim t(f)$$

$$f=\frac{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{S_2^2}{n_2})^2}{n_2-1}}$$

* 基本步骤：
  * H_0: $u_1-u_2=0$, 两总体均值不存在显著差异
  * 选择检验统计量
  * 计算t统计量的观测值和概率P值
    * SPSS给出方差齐性和异方差下的两个检验结果
  * 首先判断方差是否齐性；然后对t检验做决策
* SPSS方差齐性F检验：Levene F检验
  * H_0: 两总体方差无显著差异.
  * 方法：计算各观测与所属组均值之差的绝对值；对绝对离差进行单因素方差分析.

两独立样本t检验

- 结论：
  - 首先，如果F检验的P≤α，则拒绝F检验的H_0，认为方差不齐性；其次看方差不齐行的t检验概率. 如果≤α，则拒绝t检验的H_0，认为两总体均值有显著差异；如果>α，则不拒绝t检验的H_0.
  - 首先，如果F检验的P >α，则不能拒绝F检验的H_0，认为方差齐性；其次看方差齐行的t检验概率. 其余同上

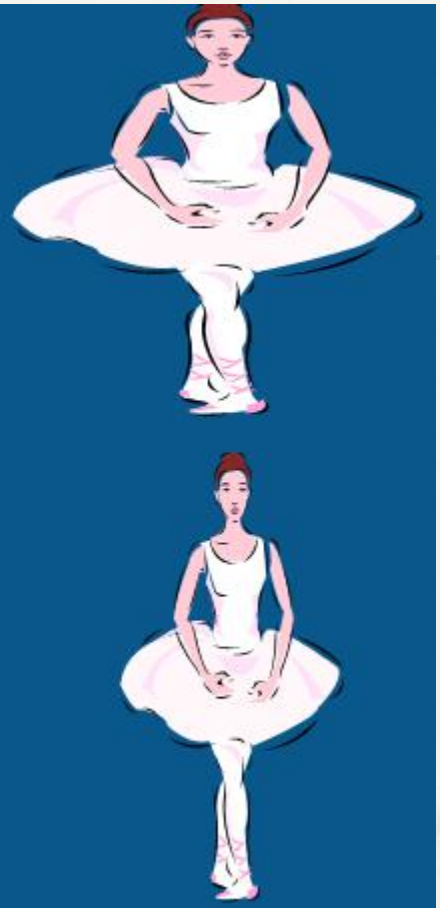
两独立样本t检验

- 基本操作步骤
  (1)菜单选项:分析->比较均值->独立样本T检验
  (2)选择若干变量作为检验变量到检验变量框
  (3)选择代表不同总体的变量作为分组变量到分组变量框
  (4)定义分组变量的分组情况:
    - 定义分组变量的分组标志值分别是什么
    - 若分组变量为连续变量.输入一个数字,将大于等于该值的分成一组,小于该值的分成另一组.

第四节　两配对样本t检验

- 目的：对两总体的均值是否有显著差异进行推断
  - 例：研究某减肥产品的减肥效果，对比减肥前与减肥后的体重总体
- 手段：利用两配对样本的均值差对两总体的均值差进行检验
  - 配对样本：抽取一个样本对抽取另一个有影响
- 理论依据：均值差的抽样分布

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



实质：先求出每对测量值的差值;然后检验差值样本的均值是否与0有显著差异.

两配对样本t检验

- 基本步骤：
  - H0：差值样本的均值u_0=0.
  - 构造统计量：同单样本均值检验
  - 如果差值的均值与0有显著差异，认为两总体均值存在显著差异；否则，与0无显著差异，则认为两总体均值不存在显著差异

$$t = \frac{\overline{D}}{S/\sqrt{n}}$$

两配对样本t检验

- 基本操作步骤

  (1) 菜单选项:分析->比较均值->配对样T检验

  (2) 选择一对或若干对配对变量作为检测变量到成对变量框.

- 例：研究减肥茶的减肥效果
  - 相关系数的作用

第五节　单因素方差分析

多个总体的均值检验

- 目的：对多个总体的均值是否有显著差异进行推断
  - 例：不同专业大学生月平均收入是否存在显著差异
- 手段：利用两独立样本的均值差对两总体的均值差进行逐对检验（多次采用两独立样本的t检验）
- 问题：犯第一类错误的概率明显增大
  - 例:K个总体做两两t检验需作$N=k!\div(2!\times(k-2)!)$次。若α为0.05，则每次不犯α错的概率为0.95。N次检验均不犯α错的概率为0.95^N，犯α错的概率为$1-0.95^N$，远远大于设定的0.05
- 解决方法：方差分析

方差分析概述

- 目的：试验设计中最优方案的设计
  - 例：不同品种的亩产量分析
  - 例：为获得最佳的产品销售量研究：哪些因素是影响销售量的主要因素；哪些因素的那种情况更利于提高销售量；哪些因素的组合更利于提高销售量
- 特点：从分析数据的差异入手，分析哪些因素是影响数据差异的众多因素中的主要因素
- 相关概念：
  - 观测变量；控制变量及水平；随机因素

研究对象：来自观测变量多个总体的多个独立样本

观测变量

控制因素

月收入	专业
Xxx,xxx,xxx,xxx Xxx,xxx,xxx,xxx	专业 1
Xxx,xxx,xxx Xxx,xxx,xxx,xxx	专业 2
Xxx,xxx,xxx,xxx Xxx,xxx	专业 3

三个水平

* 核心思路：从数据差异角度看：
  * 观测变量的差异=控制因素造成+随机因素造成
  * 当控制因素对结果有显著影响时，和随机因素共同作用必然使观测变量产生显著变动；反之，观测变量的变动较小，将归结为随机性造成的(指抽样误差)

2000 2000 2000	专业1
3000 3000 3000	专业2
4000 4000 4000	专业3

2100 2110 2200	专业1
2000 2100 2050	专业2
2100 2150 2100	专业3

2500 3110 3200	专业1
3000 3500 2800	专业2
2900 3400 3900	专业3

方差分析概述

- 类型：
  - 单因素方差分析:只考虑一个控制因素的影响
  - 多因素方差分析:考虑两个以上的控制因素和它们的交互作用对观测变量的影响
  - 协方差分析：在尽量排除其他因素的影响下，分析单个或多个控制因素对观测变量的影响(引入协变量)
  - 研究一个数值型变量和多个分类型变量之间的关系

单因素方差分析

- 目的：检验某个控制因素的改变是否会给观察变量带来显著影响
  - 例：考察不同肥料对某农作物亩产量是否有显著差异；
  - 例：考察不同温度下某化工产品的获得率
  - 例：考察妇女生育率在不同地区是否有显著差异
  - 例：考察不同学历是否对工资收入产生显著影响

单因素方差分析

- 基本思路
  - 入手点:检验控制变量的不同水平下，各总体的分布是否存在显著差异，进而判断控制变量是否对观测变量产生了显著影响.
  - 前提:各组样本独立；不同水平下各总体服从方差相等的正态分布.
  - H_0:不同水平下,各总体均值无显著差异.即:不同水平下控制因素的影响不显著

单因素方差分析

- 检验统计量：总变差=组间差异+组内差异
  - SST= SSA+SSE(设:k个水平，每个水平有n_i个数据)

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

  - 考察平均的组间差异与平均的组内差异的比值：

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

方差齐性检验

- 各水平下的方差齐性检验
  - SPSS方差齐性F检验：Levene F检验
    - H_0:两总体方差无显著差异.
    - 方法：计算各观测与所属组均值之差的绝对值；对绝对离差进行单因素方差分析.

单因素方差分析中的多重比较

- 目的：若各总体均值存在差异,F检验不能说明哪个水平造成了观察变量的显著差异
  - 对每个水平的均值逐对进行比较检验
- 几种常用的多重比较方法
  - LSD(Least significant Difference)最小显著性差异法

$$t=\frac{\bar{x}_i-\bar{x}_j}{\sqrt{MSE(\frac{1}{n_i}+\frac{1}{n_j})}}\sim t(n-k)$$其中n为总样本数

  - 特点:利用全部样本数据;在一定程度上克服了放大犯一类错误的问题

多因素方差分析

- 基本思路：认为观测变量的变动是由各控制变量独立作用、它们的交互作用、以及随机因素造成的
- 以两个控制变量为例：

$$x_{ijk} = \mu + a_i + b_i + (ab)_{ij} + \varepsilon_{ijk}$$

SST=SSA+SSB+SSAB+SSE

(main effects) (N-way 交互) (Residual)

(explained)

其中：SSAB表示两个控制变量交互影响带来的变差

* 基本思路：SST=SSA+SSB+SSAB+SSE

  A有p个水平，B有q个水平，每组有r个样本

$$SST=\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(x_{ijk}-\bar{x})^2 \qquad SSB=pr\sum_{j=1}^{q}(\bar{x}_j^B-\bar{x})^2$$

$$SSA=qr\sum_{i=1}^{p}(\bar{x}_i^A-\bar{x})^2 \qquad SSE=\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(x_{ijk}-\bar{x}_{ij})^2$$

$$SSAB=SST-SSA-SSB-SSE$$

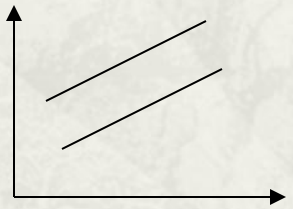
$$F_A=\frac{SSA/(p-1)}{SSE/pq(r-1)} \quad F_B=\frac{SSB/(q-1)}{SSE/pq(r-1)} \quad F_{AB}=\frac{SSAB/(p-1)(q-1)}{SSE/pq(r-1)}$$

多因素方差分析说明

- 多因素方差分析中因素的划分：
  - 固定效应因素：该因素的所有可能水平在样本中都出现。
    - 如:性别；糖尿病有无:糖尿病,糖耐量异常,正常人--固定效应模型
  - 随机效应因素：无法对所有水平值进行准确控制和观测，研究的水平值是随机挑选出的
    - 如:城市规模,教育水平等-- 随机效应模型
  - 混合效应模型

- 交互效应：两个或多个控制变量各水平搭配对观测变量的影响。若一个因素所产生的效应在另一个因素的不同水平下有明显差异，则称这两因素存在交互效应
  - 直观上：饮食习惯、适量运动对减肥的作用；排球对的二传手和主攻手对赢球的作用
  - 交互作用的图形观察：

	A1	A2
B1	2	5
B2	7	10



当A从A1变化到A2时，观测变量值均增加且幅度相同，与B1或B2无关；同理B

	A1	A2
B1	2	5
B2	7	3

A对观测变量值的影响与B取什么水平有关

协方差分析

- 目的：将无法或很难控制的因素作为协变量，在排除协变量影响下分析控制变量对观测变量的影响.
- 例：

体重增量	饲料	初始体重
xxx,xxx,xxx xxx,xxx	1	xxx,xxx,xxx xxx,xxx
xxx,xxx, xxx,xxx,xxx	2	xxx,xxx, xxx,xxx,xxx
xxx,xxx,xxx xxx,xxx,xxx	3	xxx,xxx,xxx xxx,xxx,xxx

	A	B	C	D
1	月工资收入（元） y	工作年限（年） x_1	性别 x_2	x_2
2	2900	2	男	1
3	3000	6	女	0
4	4800	8	男	1
5	1800	3	女	0
6	2900	2	男	1
7	4900	7	男	1
8	4200	9	女	0
9	4800	8	女	0
10	4400	4	男	1
11	4500	6	男	1

协方差分析

- 基本思路：
  - 观测变量总变差：协变量、控制变量、交互作用、随机因素
  - 用线性回归的方法找出观测变量与协变量之间的数量关系，求得在假定协变量相等情况下的修正的观测变量值，然后再进行方差分析
- H0:协变量对观测变量没有显著影响；在剔除协变量影响的条件下，控制变量各水平下观测变量的总体均值无显著差异.

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij}$$

协方差分析

- 对协变量的要求：
  - 协变量是数值型的；协变量与观测变量的线性关系在各水平均成立，且斜率大致相同
  - 协方差分析是界于方差分析和回归分析之间的一种分析方法（定距型变量、分类变量）
- 检验统计量：
  - F= MSA/MSE　　F=MSB/MSE
  - F= MSAB/MSE　　F= MSZ/ MSE

协方差分析

- 例：不同饲料是否对小猪体重的增加产生显著差异
  - 一般单因素方差分析
  - 注意：R^2值不很高
  - 存在的问题：初始体重有显著差异(单因素方差分析)
    - 第一种饲料猪的初始体重最低，第三种饲料猪的初始体重最高
    - 如果初始体重对猪的催肥有显著影响，则它与饲料的效应就会混杂
    - 观测每种饲料下初始体重与增重的关系：散点图可见，线性关系，且斜率大致相同，可考虑采用协方差分析

* 例：采用协方差分析
  * 注意：R^2值有提高，变差分析发生变化

Tests of Between-Subjects Effects

Dependent Variable: 喂养后体重增加

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1317.583[a]	2	658.792	11.172	.000
Intercept	204057.042	1	204057.042	3460.339	.000
SL	1317.583	2	658.792	11.172	.000
Error	1238.375	21	58.970		
Total	206613.000	24			
Corrected Total	2555.958	23			

a. R Squared = .515 (Adjusted R Squared = .469)

Tests of Between-Subjects Effects

Dependent Variable: 喂养后体重增加

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2328.344[a]	3	776.115	68.196	.000
Intercept	980.448	1	980.448	86.150	.000
WYQ	1010.760	1	1010.760	88.813	.000
SL	707.219	2	353.609	31.071	.000
Error	227.615	20	11.381		
Total	206613.000	24			
Corrected Total	2555.958	23			

a. R Squared = .911 (Adjusted R Squared = .898)

- 变差的分解：
  - SSA的分解：SSA是各水平均值与总均值差的平方和。为排除协变量影响，应从总体上将协变量作用扣除后再计算SSA
  - 利用所有数据计算回归方程：$wyh = 63.333 + 1.5wyq$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	63.333	4.861	
	喂养前体重	1.500	.243	.796

a. Dependent Variable: 喂养后体重增加

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1621.125	1	1621.125	38.151	.000[a]
	Residual	934.833	22	42.492		
	Total	2555.958	23			

a. Predictors: (Constant), 喂养前体重

b. Dependent Variable: 喂养后体重增加

  - 从体重增量的总变差中扣除回归平方和后的剩余平方和为喂养前体重不能解释的变差(934.834)，为剔排除协变量影响后的观测变量的总变差
  - 饲料可解释的变差：SST-SSE=SSA

$$934.834 - 227.615 = 707.219$$

- 变差的分解：
  - SSE的分解：SSE是各观测值与各组均值差的平方和。为排除协变量影响，应在各组内部将协变量的作用扣除后再计算SSE：
  - （1）分别建立三个水平下的回归方程：

$$wyh = 33.516 + 3.508wyq$$

$$wyh = 54.570 + 2.332wyq$$

$$wyh = 43.141 + 2.118wyq$$

  - 分别计算三个水平下协变量和观测变量的差积S_{zy}以及协变量离差平方和S_{zz}(利用已知的SSR和B计算)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\hat{y} - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} = \frac{S_{zy}}{S_{zz}}$$

S_{zy} =SSR/B S_{zz}=S_{zy}/B

* 变差的分解：SSE的分解

  * (2)计算共同的B*

$$S_{zy}^{1} = 387.627 \div 3.508 = 110.5$$

$$S_{zz}^{1} = 110.5 \div 3.508 = 31.5$$

$$S_{zy}^{2} = 151.507 \div 2.332 = 65$$

$$S_{zz}^{2} = 65 \div 2.332 = 27.87$$

$$S_{zy}^{3} = 519.602 \div 2.118 = 245.3$$

$$S_{zz}^{3} = 245.3 \div 2.118 = 115.8$$

$$S_{zy}^{*} = S_{zy}^{1} + S_{zy}^{2} + S_{zy}^{3} = 110.5 + 65 + 245.3 = 420.8$$

$$S_{zz}^{*} = S_{zz}^{1} + S_{zz}^{2} + S_{zz}^{3} = 31.5 + 27.87 + 115.8 = 175.17$$

$$\beta^{*} = 420.8 \div 175.17 = 2.4$$

  * (3)计算各水平下具有共同斜率的三个回归方程：

$wyh = 48.75 + 2.4wyq$　$wyh = 53.30 + 2.4wyq$　$wyh = 35.975 + 2.4wyq$

  * SSE为：对剔除协变量影响后的残差的组内离差平方和(227.615)

ANOVA

Res

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	48.000	2	24.000	2.214	.134
Within Groups	227.615	21	10.839		
Total	275.615	23			

- 变差的分解：协变量可解释的总变差(1010.76)：

$$S_{zy}^{*} = S_{zy}^{1} + S_{zy}^{2} + S_{zy}^{3} = 110.5 + 65 + 245.3 = 420.8$$

$$S_{zz}^{*} = S_{zz}^{1} + S_{zz}^{2} + S_{zz}^{3} = 31.5 + 27.87 + 115.8 = 175.17$$

$$\beta^{*} = 420.8 \div 175.17 = 2.4$$

$$\beta^{*} \times S_{zy}^{*} = 2.4 \times 420.8 = 1010.8$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\hat{y} - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} = \frac{S_{zy}}{S_{zz}}$$

S_{zy} =SSR/B S_{zz}=S_{zy}/B

* 各水平均值的对比：
  * 利用均值进行比较

Descriptive Statistics

Dependent Variable: 喂养后体重增加

饲料种类	Mean	Std. Deviation	N
1.00	81.7500	8.34523	8
2.00	98.0000	5.12696	8
3.00	96.8750	8.99901	8
Total	92.2083	10.54176	24

* 利用修正的均值进行比较：

$$\bar{y}_i^* = \bar{y}_i - \beta^*(\bar{z}_i - \bar{z})$$

  * 修正是将各水平下本组协变量的效益从本组观测变量中剔除
  * 计算三种饲料下增重的修正均值分别约为：94.95、99.05、82.175。第一种饲料比第三种饲料平均多增重12.793，第二种比第三种平均多增重17.336，第二种比第一种平均多重4.52

第五章

相关与回归分析

主要内容

* 第一节　相关分析概述
* 第二节　偏相关分析
* 第三节　简单线性回归分析
* 第四节　多元线性回归分析

第一节 相关分析概述

(一)相关关系

(1)函数关系:事物间的一种一一对应的确定性关系.即:当一个变量x取一定值时,另一变量y可以依确定的关系取一个确定的值

如:销售额与销售量;圆面积和圆半径

(2)统计关系:事物间的关系不是确定性的.即:当一个变量x取一定值时,另一变量y的取值可能有几个.一个变量的值不能由另一个变量唯一确定

* 如:收入和消费;身高的遗传.

相关分析概述

- 统计关系的常见类型：
  - 线性相关：正线性相关、负线性相关
  - 非线性相关

  - 统计关系不象函数关系那样直接,但却普遍存在,且有强有弱.如何测度?

- 相关分析的研究对象:统计关系
- 相关分析旨在测度变量间线性关系的强弱程度

相关分析

(一)目的

通过样本数据,研究两变量间线性相关程度的强弱.

(二)基本方法

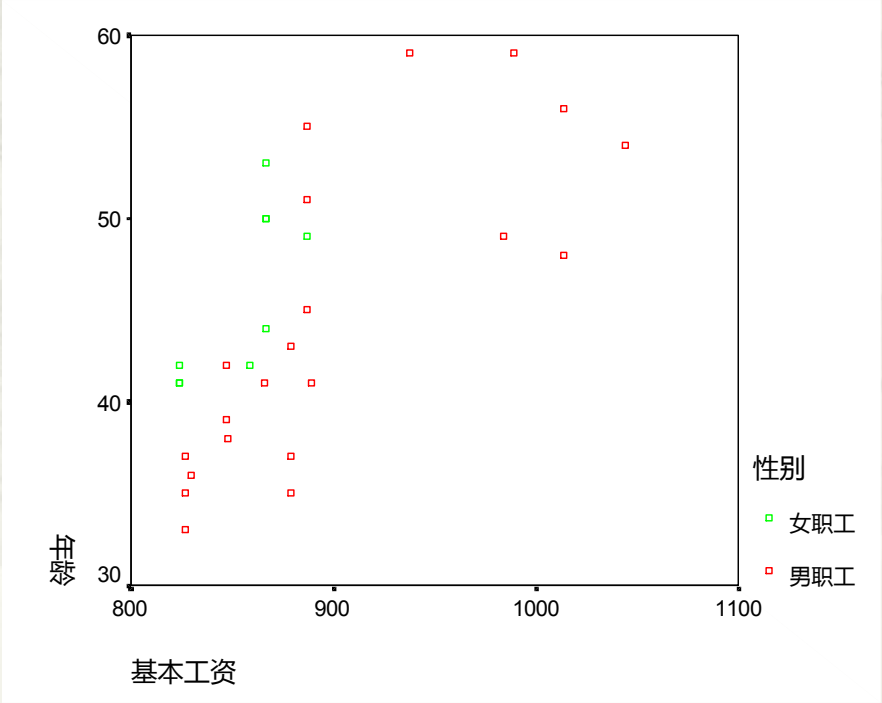
绘制散点图、计算相关系数

绘制散点图

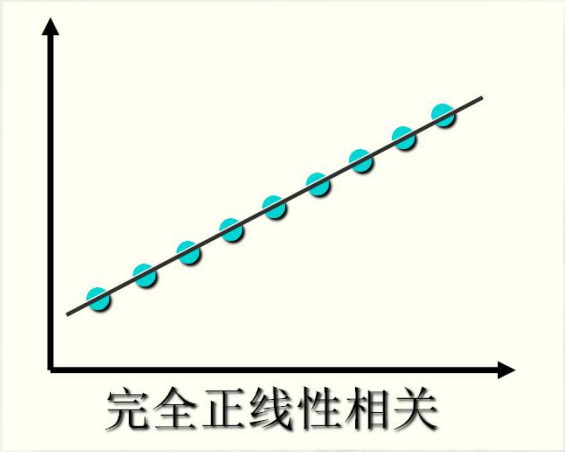
(一)散点图

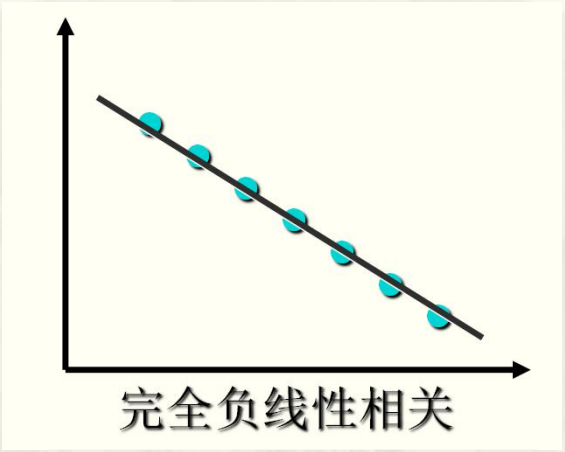
将数据以点的形式绘制在直角平面上.比较直观,可以用来发现变量间的关系和可能的趋势.

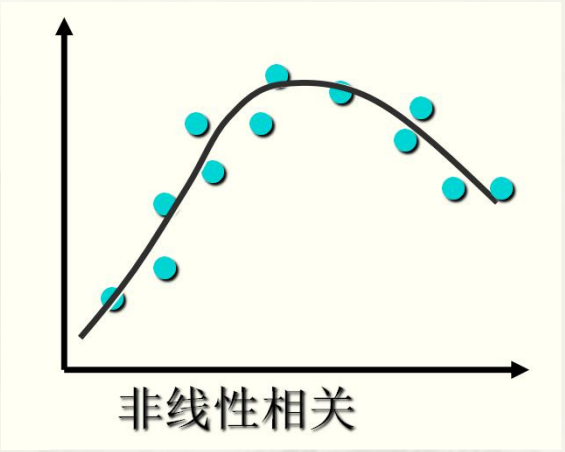
正相关趋势

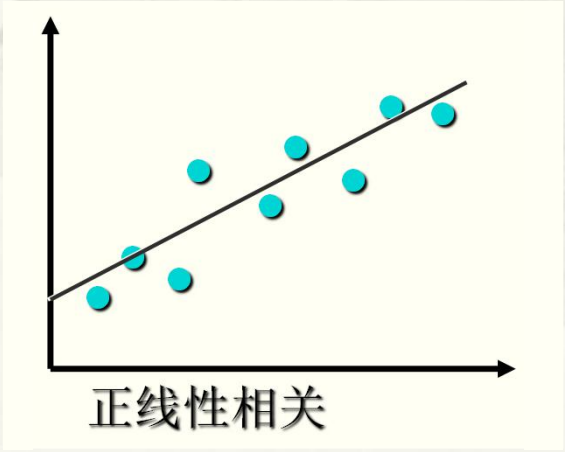


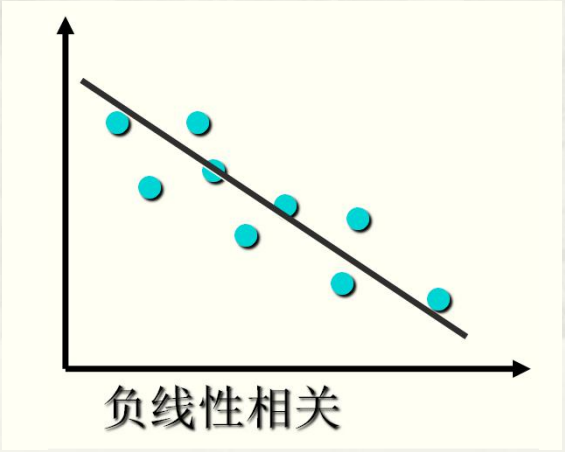
绘制散点图

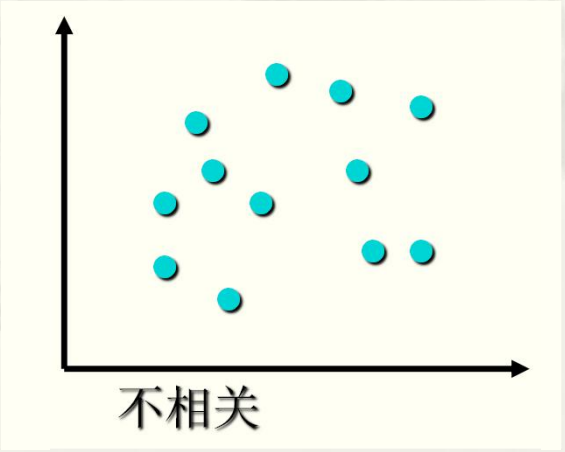












计算相关系数

(一)相关系数

(1)作用:

* 以精确的相关系数(r)体现两个变量间的线性关系程度.
* r:[-1,+1]; r=1:完全正相关; r=-1:完全负相关; r=0:无线性相关; |r|>0.8:强相关; |r|<0.3:弱相关

计算相关系数

(2)说明:

* 相关系数只是较好地度量了两变量间的线性相关程度,不能描述非线性关系.
* 如:x和y的取值为:(-1,-1) (-1,1) (1,-1) (1,1), r=0 但 $x^2+y^2=2$
* 数据中存在极端值时不好
* 如:(1,1)(2,2)(3,3),(4,4),(5,5),(6,1), r=0.33,但总体上表现出x=y,应结合散点图分析

计算相关系数

(3)种类:

- 简单线性相关系数(Pearson):针对定距数据.

  (如:身高和体重)

$$r=\frac{\sum_{1=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$$

计算相关系数

- Spearman相关系数:用来度量定序变量间的线性相关关系(如:不同年龄段与不同收入段,职称和受教育年份)
  - 利用秩(数据的排序次序).认为:如果x与y相关,则相应的秩U_i、V_i也具有同步性.
  - 首先得到两变量中各数据的秩(U_i、V_i),并计算D_i^2统计量.
    - 若两变量存在强正相关性,则Di2应较小,秩序相关系数较大.若两变量存在强负相关性,则Di2应较大,秩序相关系数为负,绝对值较大
  - 计算Spearman相关系数,与简单相关系数形式完全相同.

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (U_i - V_i)^2$$

$$R = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)}$$

计算相关系数

* Kendall相关系数:度量定序变量间的线性相关关系
  * 首先计算一致对数目(U)和非一致对数目(V)

  如：对x和y求秩后为： x: 2 4 3 5 1

  y： 3 4 1 5 2

  x的秩按自然顺序排序后： x: 1 2 3 4 5

  y: 2 3 1 4 5

  一致对U:(2,3) (2,4)(2,5)(3,4)(3,5)(1,4)(1,5)(4,5)； 非一致对V:(2,1)(3,1)
  * 然后计算Kendall相关系数
  * 若两变量存在强相关性，则秩相关系数较大；若两变量存在强负关联性，则秩关，秩相关系数为负，绝对值较大

$$T = (U - V) - \frac{2}{n(n-1)}$$

计算相关系数

(二)相关系数检验

* 应对两变量来自的总体是否相关进行统计推断.
  * 原因:抽样的随机性、样本容量小等

(1)H_0:两总体零相关

(2)构造统计量

简单相关系数

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Spearman系数,大样本下,近似正态分布

$$Z = R\sqrt{n-1}$$

kendall系数,大样本下,近似正态分布

$$Z = \frac{3T\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

计算相关系数

(二)相关系数检验

(3)计算统计量的值,并得到对应的相伴概率p

(4)结论:

- 如果p<=a,则拒绝H_0,两总体存在线性相关;
- 如果p>a,不能拒绝H_0.

计算相关系数

(三)基本操作步骤

(1)菜单选项:分析->相关->双变量

(2)选择计算相关系数的变量到变量框.

(3)选择相关系数.

(4)显著性检验

* 输出双尾检验概率P
* 输出单尾检验概率P

第二节 偏相关分析

(一)偏相关系数

(1)含义：

在控制了其他变量的影响下计算两变量的相关系数

* 虚假相关.
* 研究商品的需求量和价格、消费者收入之间的关系.因为:需求量和价格之间的相关关系包含了消费者收入对商品需求量的影响；收入对价格也产生影响，并通过价格变动传递到对商品需求量的影响中。

偏相关分析

(2)计算方法：

$$r_{y1.2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}$$

偏相关分析

(二)基本操作步骤

(1)菜单选项:分析->相关->偏相关

(2)选择将参加计算的变量到变量框.

(3)选择控制变量到控制框。

(4)选项:

* 零阶相关系数:输出简单相关系数矩阵

第三节 简单线性回归分析

(一)回归分析理解

(1)“回归”的含义：galton研究父亲身高和儿子身高的关系时的独特发现.

(2)回归线的获得方式一:局部平均

- 回归曲线上的点给出了相应于每一个x(父亲)值的y(儿子)平均数的估计

(3)回归线的获得方式二:拟和函数

* 使数据拟和于某条曲线;
* 通过若干参数描述该曲线;
* 利用已知数据在一定的统计准则下找出参数的估计值(得到回归曲线的近似);

回归分析概述

(二) 回归分析的基本步骤

* (1) 确定自变量和因变量(父亲身高关于儿子身高的回归与儿子身高关于父亲身高的回归是不同的).
* (2) 从样本数据出发确定变量之间的数学关系式,并对回归方程的各个参数进行估计.
* (3) 对回归方程进行各种统计检验.
* (4) 利用回归方程进行预测.

回归分析概述

(三)参数估计的准则

* 目标:观察值与回归线上的预测值之间的距离总和达到最小
* 最小二乘法(利用最小二乘法拟和的回归直线与样本数据点在垂直方向上的偏离程度最低)

一元线性回归分析

(一)一元回归方程:

- $y=\beta_0+\beta_1 x$
- β_0为常数项；β_1为y对x回归系数，即:x每变动一个单位所引起的y的平均变动

(二)一元回归分析的步骤

* 利用样本数据建立回归方程
* 回归方程的拟和优度检验
* 回归方程的显著性检验(t检验和F检验)
* 残差分析
* 预测

一元线性回归方程的检验

(一)拟和优度检验:

(1)目的:检验样本观察点聚集在回归直线周围的密集程度，评价回归方程对样本数据点的拟和程度

(2)思路:

- 因为：因变量取值的变化受两个因素的影响
  - 自变量不同取值的影响；其他因素的影响
- 于是：因变量总变差=自变量引起的+其他因素引起的
- 即:因变量总变差=回归方程可解释的+不可解释的
- 可证明:因变量总离差平方和=回归平方和+剩余平方和

一元线性回归方程的检验

- (3)统计量：判定系数 $R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$
  - R^2=SSR/SST=1-SSE/SST.
  - R^2体现了回归方程所能解释的因变量变差的比例；1-R^2则体现了因变量总变差中，回归方程所无法解释的比例。
  - R^2越接近于1，则说明回归平方和占了因变量总变差平方和的绝大部分比例，因变量的变差主要由自变量的不同取值造成，回归方程能够较好拟合样本数据点
    - 在一元回归中$R^2=r^2$；因此，从这个意义上讲，判定系数能够比较好地反映回归直线对样本数据的代表程度和线性相关性。

一元线性回归方程的检验

(二)回归方程的显著性检验：F检验

(1)目的:检验自变量与因变量之间的线性关系是否显著,是否可用线性模型来表示.

(2)H_0: β=0 即:回归系数与0无显著差异

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/k}{\sum(y_i-\hat{y}_i)^2/(n-k-1)}$$

(3)利用F检验,构造F统计量:

* F=平均的回归平方和/平均的剩余平方和~F(1,n-1-1)
* 如果F值较大，则说明自变量造成的因变量的线性变动远大于随机因素对因变量的影响,自变量于因变量之间的线性关系较显著

(4)计算F统计量的值和相伴概率p

(5)判断

* p<=a:拒绝H_0,即:回归系数与0有显著差异，自变量与因变量之间存在显著的线性关系。反之，不能拒绝H_0

一元线性回归方程的检验

(三)回归系数的显著性检验:t检验

(1)目的:检验自变量对因变量的线性影响是否显著.

(2)H_0: β=0 即:回归系数与0无显著差异

(3)利用t检验,构造t统计量:

$$t_i = \frac{\beta_i}{S_{\beta i}} \qquad S_{\beta i} = \sqrt{\frac{S^2_y}{\sum (x_i - \bar{x}_i)^2}}$$

- 其中:S_y是回归方程标准误差(Standard Error)的估计值，由均方误差开方后得到，反映了回归方程无法解释样本数据点的程度或偏离样本数据点的程度
- 如果回归系数的标准误较小，必然得到一个相对较大的t值，表明该自变量x解释因变量线性变化的能力较强。

(4)计算t统计量的值和相伴概率p

(5)判断

一元线性回归方程的检验

(四)t检验与F检验的关系

- 一元回归中,F检验与t检验一致,即：$F=t^2$,两种检验可以相互替代

(五)F统计量和R^2值的关系

$$F=\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

- 如果回归方程的拟合优度高，F统计量就越显著。F统计量越显著，回归方程的拟合优度就会越高。

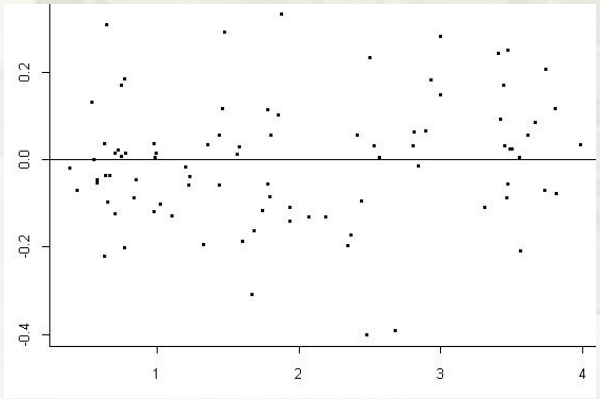
线性回归方程的残差分析

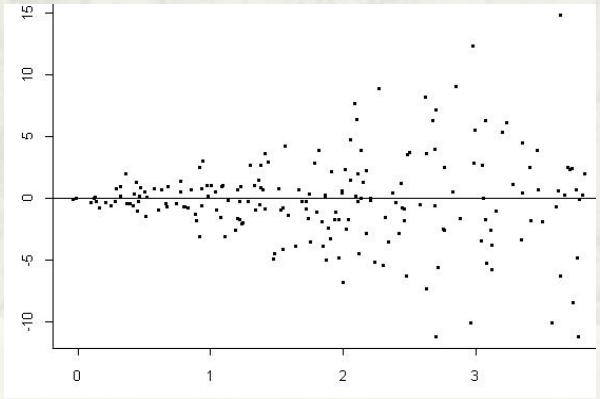
(一)残差序列的正态性检验:

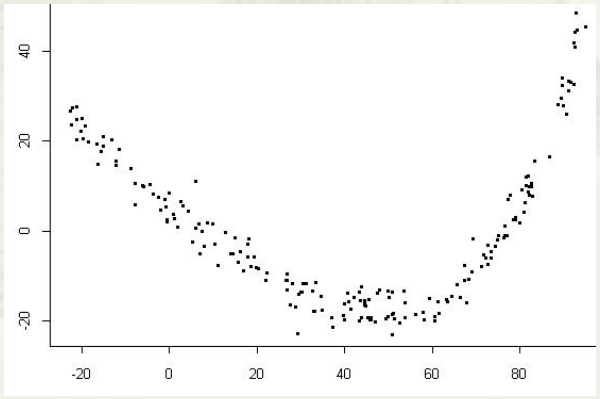
- 绘制标准化残差的直方图或累计概率图

(二)残差序列的随机性检验

- 绘制残差和预测值的散点图,应随机分布在经过零的一条直线上下







线性回归方程的残差分析

(三)残差序列独立性检验:

- 残差序列是否存在后期值与前期值相关的现象,利用D.W(Durbin-Watson)检验
- d-w=0:残差序列存在完全正自相关;d-w=4:残差序列存在完全负自相关;0<d-w<2:残差序列存在某种程度的正自相关;2<d-w<4:残差序列存在某种程度的负自相关;d-w=2:残差序列不存在自相关.
- 残差序列不存在自相关,可以认为回归方程基本概括了因变量的变化;否则,认为可能一些与因变量相关的因素没有引入回归方程或回归模型不合适或滞后性周期性的影响.

线性回归方程的残差分析

(四)异常值(casewise或outliers)诊断

* 利用标准化残差不仅可以知道观察值比预测值大或小,并且还知道在绝对值上它比大多数残差是大还是小.一般标准化残差的绝对值大于3,则可认为对应的样本点为奇异值
* 异常值并不总表现出上述特征.当剔除某观察值后，回归方程的标准差显著减小,也可以判定该观察值为异常值

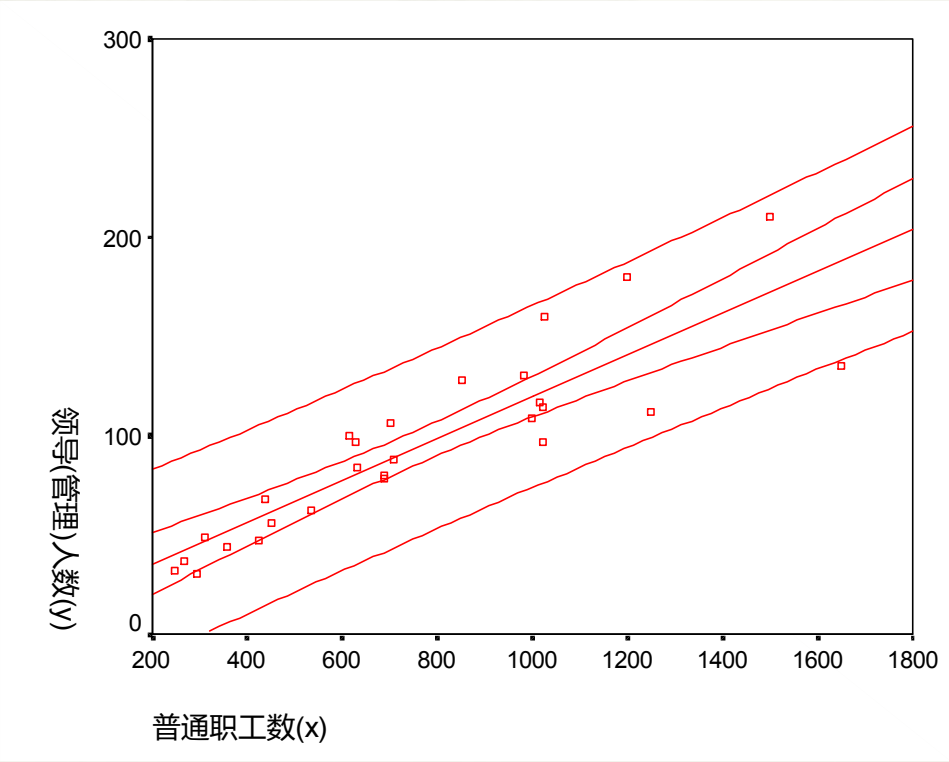
线性回归方程的预测

(一)点估计

y_0

(二)区间估计

x_0为x_i的均值时,预测区间最小,精度最高.x_0越远离均值,预测区间越大,精度越低.



第四节 多元线性回归分析

(一)多元线性回归方程

多元回归方程：$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

- β_1、β_2、β_k为偏回归系数。
- β_1表示在其他自变量保持不变的情况下，自变量x_1变动一个单位所引起的因变量y的平均变动

(二)多元线性回归分析的主要问题

* 回归方程的检验
* 自变量筛选
* 多重共线性问题

多元线性回归方程的检验

(一)拟和优度检验:

$$\overline{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST}$$

$$\overline{R}^2 = 1 - \frac{\text{均方误差}}{\text{因变量的样本方差}}$$

(1)判定系数R^2:

- R是y和x_i的复相关系数(或观察值与预测值的相关系数),测定了因变量y与所有自变量全体之间线性相关程度

(2)调整的R^2

* 考虑的是平均的剩余平方和,克服了因自变量增加而造成R^2也增大的弱点
* 在某个自变量引入回归方程后，如果该自变量是理想的且对因变量变差的解释说明是有意义的，那么必然使得均方误差减少，从而使调整的R^2得到提高；反之，如果某个自变量对因变量的解释说明没有意义，那么引入它不会造成均方误差减少，从而调整的R^2也不会提高。

多元线性回归方程的检验

(二)回归方程的显著性检验：

(1)目的:检验所有自变量与因变量之间的线性关系是否显著，是否可用线性模型来表示.

(2)H_0: $\beta_1=\beta_2=\cdots=\beta_k=0$ 即:所有回归系数同时与0无显著差异

(3)利用F检验,构造F统计量:

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/k}{\sum(y_i-\hat{y}_i)^2/(n-k-1)}$$

* F=平均的回归平方和/平均的剩余平方和
* 如果F值较大，则说明自变量造成的因变量的线性变动大于随机因素对因变量的影响,自变量于因变量之间的线性关系较显著

(4)计算F统计量的值和相伴概率p

(5)判断： p<=a:拒绝H_0,即:所有回归系数与0有显著差异，自变量与因变量之间存在显著的线性关系。反之，不能拒绝H_0

多元线性回归方程的检验

(三)回归系数的显著性检验

(1)目的:检验每个自变量对因变量的线性影响是否显著.

(2)H_0: $\beta_i=0$ 即:第i个回归系数与0无显著差异

(3)利用t检验,构造t统计量:

$$t_i = \frac{\beta_i}{S_{\beta i}} \qquad S_{\beta i} = \sqrt{\frac{S^2_y}{\sum (x_i - \bar{x}_i)^2}}$$

(4)逐个计算t统计量的值和相伴概率p

(5)判断

多元线性回归方程的检验

(四)t统计量与F统计量

* 一元回归中,F检验与t检验一致,即：$F=t^2$,可以相互替代
* 在多元回归中，F检验与t检验不能相互替代
* $F_{change}=t_i^2$ $F_{change}=\frac{R_{ch}^2(n-k-1)}{1-R^2}$ $R_{ch}^2=R^2-R_i^2$
* 从F_{change}角度上讲，如果由于某个自变量x_i的引入，使得F_{change}是显著的(通过观察F_{change}的相伴概率值)，那么就可以认为该自变量对方程的贡献是显著的，它应保留在回归方程中，起到与回归系数t检验同等的作用。

自变量筛选

(一)自变量筛选的目的

* 多元回归分析引入多个自变量. 如果引入的自变量个数较少,则不能很好的说明因变量的变化;

* 并非自变量引入越多越好.原因:
  * 有些自变量可能对因变量的解释没有贡献
  * 自变量间可能存在较强的线性关系,即:多重共线性. 因而不能全部引入回归方程.

自变量筛选

(二)自变量向前筛选法(forward):

* 即:自变量不断进入回归方程的过程.
* 首先,选择与因变量具有最高相关系数的自变量进入方程,并进行各种检验;
* 其次,在剩余的自变量中寻找偏相关系数最高的变量进入回归方程,并进行检验;
  * 默认:回归系数检验的概率值小于PIN(0.05)才可以进入方程.
* 反复上述步骤,直到没有可进入方程的自变量为止.

自变量筛选

(三)自变量向后筛选法(backward):

* 即:自变量不断剔除出回归方程的过程.
* 首先,将所有自变量全部引入回归方程;
* 其次,在一个或多个t值不显著的自变量中将t值最小的那个变量剔除出去,并重新拟和方程和进行检验;
  * 默认:回归系数检验值大于POUT(0.10),则剔除出方程
* 如果新方程中所有变量的回归系数t值都是显著的,则变量筛选过程结束.
* 否则,重复上述过程,直到无变量可剔除为止.

自变量筛选

(四)自变量逐步筛选法(stepwise):

- 即:是“向前法”和“向后法”的结合。
- 向前法只对进入方程的变量的回归系数进行显著性检验，而对已经进入方程的其他变量的回归系数不再进行显著性检验，即：变量一旦进入方程就不回被剔除
- 随着变量的逐个引进，由于变量之间存在着一定程度的相关性，使得已经进入方程的变量其回归系数不再显著，因此会造成最后的回归方程可能包含不显著的变量。
- 逐步筛选法则在变量的每一个阶段都考虑的剔除一个变量的可能性。

线性回归分析中的共线性检测

(一)共线性带来的主要问题

- 高度的多重共线会使回归系数的标准差随自变量相关性的增大而不断增大,以至使回归系数的置信区间不断增大,造成估计值精度减低.

(二)共线性诊断

* 自变量的容忍度(tolerance)和方差膨胀因子
  * 容忍度:$Tol_i=1-R_i^2$. 其中: R_i^2是自变量x_i与方程中其他自变量间的复相关系数的平方.
  * 容忍度越大则与方程中其他自变量的共线性越低,应进入方程. (具有太小容忍度的变量不应进入方程,spss会给出警)(T<0.1一般认为具有多重共线性)
  * 方差膨胀因子(VIF):容忍度的倒数
  * SPSS在回归方程建立过程中不断计算待进入方程自变量的容忍度,并显示目前的最小容忍度

线性回归分析中的共线性检测

(二)共线性诊断

* 用特征根刻画自变量的方差
  * 若自变量间确实存在较强的相关关系，那么它们之间必然存在信息重叠，于是可从这些自变量中提取出既能反映自变量信息(方差)又相互独立的因素(成分)来.
  * 从自变量的相关系数矩阵出发，计算相关系数矩阵的特征根，得到相应的若干成分.
  * 若某个特征根既能够刻画某个自变量方差的较大部分比例(如大于0.7),同时又可以刻画另一个自变量方差的较大部分比例，则表明这两个自变量间存在较强的多重共线性。
* 条件指标
  * 0<k<10 无多重共线性； 10<=k<=100 较强； k>=100 严重

$$k_i = \sqrt{\frac{\lambda_m}{\lambda_i}}$$

线性回归分析中的异方差问题

(一)什么是差异方差

- 回归模型要求残差序列服从均值为0并具有相同方差的正态分布,即:残差分布幅度不应随自变量或因变量的变化而变化.否则认为出现了异方差现象

(二)差异方差诊断

- 可以通过绘制标准化残差序列和因变量预测值(或每个自变量)的散点图来识别是否存在异方差

(三)异方差处理

* 实施方差稳定性变换
  * 残差与y_i(预测值)的平方根呈正比：对y_i开平方
  * 残差与y_i(预测值)呈正比:对y_i取对数.
  * 残差与y_i(预测值)的平方呈正比,则$1/y_i$

第六章

聚类分析

主要内容

* 第一节　聚类分析概述
* 第二节　分层聚类的基本思想及应用
* 第三节　快速聚类法的基本思想及应用

第一节　聚类分析概述

- 概念：
  - 聚类分析是统计学中研究“物以类聚”的一种方法,属多元统计分析方法.
  - 例如：细分市场、消费行为划分

- 聚类分析是建立一种分类，是将一批样本(或变量)按照在性质上的“亲疏”程度,在没有先验知识的情况下自动进行分类的方法.其中:类内个体具有较高的相似性,类间的差异性较大.

聚类分析概述

- 依据平均得分的差距, 差距较小的为一类.
- 分类过程中, 没有事先指定分类的标准. 完全根据样本数据客观产生分类结果.

编号	购物环境	服务质量
A	73	68
B	66	69
C	84	82
D	91	88
E	94	90

两类:(A B) (C D E) 三类:(A B) (C) (D E)

聚类分析概述

- 亲疏远程度的衡量指标
  - 相似性:数据间相似程度的度量
  - 距离：数据间差异程度的度量.距离越近,越“亲密”,聚成一类;距离越远,越“疏远”,分别属于不同的类
- 定距型个体间的距离：

  把每个个案数据看成是n维空间上的点,在点和点之间定义某种距离.一般适用于定距数据
  - 欧氏距离(EUCLID)
  - 平方欧氏距离(SEUCLID)

$$EUCLID(x,y)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2}$$

聚类分析概述

- 品质型个体间的距离

姓名	授课方式	上机时间	选某门课程
张三	1	1	1
李四	1	1	0
王五	0	0	1

聚类分析概述

- 品质型个体间的距离
  - 简单匹配(simple matching)系数:适用二值变量。

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

个体i \ 个体j	1	0
1	a	b
0	c	d

a为个体i与个体j在所有变量上同时取1的个数；d为同时取0的个数

特点：排除同时拥有或同时不拥有某特征的情况；取0和1地位等价，编码方案的变化不会引起系数的变化。

聚类分析概述

- 品质型个体间的距离
  - 简单匹配(simple matching)系数:适用二值变量。

姓名	授课方式	上机时间	选某门课程
张三	1	1	1
李四	1	1	0
王五	0	0	1

（张三，李四）：a=2 b=1 c=0 d=0 d(x,y)=1/(1+2)=1/3

（张三，王五）：a=1 b=2 c=0 d=0 d(x,y)=2/(1+2)=2/3

张三距李四近

聚类分析概述

- 品质型个体间的距离
  - 根据临床表现研究病人是否有类似的病

姓名	性别	发烧	咳嗽	检查1	检查2	检查3	检查4
张三	男	1	0	1	0	0	0
李四	女	1	0	1	0	1	0
王五	男	1	1	0	0	0	0
……							

聚类分析概述

- 品质型个体间的距离
  - 雅科比（Jaccard）系数:适用二值变量

$$J(i,j)=\frac{b+c}{a+b+c}$$

个体i \ 个体j	1	0
1	a	b
0	c	d

a为个体i与个体j在所有变量上同时取1的个数；d为同时取0的个数

特点：排除同时不拥有某特征的情况；取1的状态比取0更有意义(如:临床检验中的阳性特征)；编码方案会引起系数的变化

聚类分析概述

- 品质型个体间的距离
  - Jaccard系数举例:根据临床表现研究病人是否有类似的病

姓名	性别	发烧	咳嗽	检查1	检查2	检查3	检查4
张三	男	1	0	1	0	0	0
李四	女	1	0	1	0	1	0
王五	男	1	1	0	0	0	0

$$d(\text{张三},\text{李四}) = \frac{0+1}{2+0+1} = 0.33 \qquad d(\text{张三},\text{王五}) = \frac{1+1}{1+0+1} = 0.67$$

$$d(\text{李四},\text{王五}) = \frac{1+2}{1+1+2} = 0.75$$

结论:张三和李四最有可能得类似的病;李四和王五不太有可能

聚类分析概述

- 品质型个体间的距离
  - 卡方距离:计数变量

姓名	选修课门数（期望频数）	专业课门数（期望频数）	得优门数（期望频数）	合计
张三	9（8.5）	6（6）	4（4.5）	19
李四	8（8.5）	6（6）	5（4.5）	19
合计	17	12	9	38

$$\sqrt{(\frac{(9-8.5)^2}{8.5}+\frac{(6-6)^2}{6}+\frac{(4-4.5)^2}{4.5})+(\frac{8-8.5)^2}{8.5}+\frac{(6-6)^2}{6}+\frac{(5-4.5)^2}{4.5})}=4.12$$

聚类分析概述

* 说明：
  * 聚类过程中如果数据在数量级上存在差异时，应进行标准化处理。例如：

样本号	社科活动人员数(人)	研究与发展年投入经费(元)	研究与发展课题数(项)
1	410	4380000	19
2	336	1730000	21
3	490	220000	8

样本的欧氏距离

	元	十万元
(1，2)	265000	74.07
(1，3)	416000	80.86
(2，3)	151000	154.56

聚类分析概述

- 说明：
  - 聚类分析中的变量选择问题
    - 变量应和聚类分析的目标密切相关，聚类结果仅是所选定变量所具数据特点的反应
    - 变量之间不应具有高度相关性,否则相当于给这些变量进行了加权
  - 聚类分析包括：
    - 个案聚类(Q型)和变量聚类(R型)

第二节　分层聚类

- 聚类过程具有一定的层次性，某个类是另个类的子类
- 以合并(凝聚)的方式聚类(SPSS采用)
  - 首先,每个个体自成一类
  - 其次,将最“亲密”的个体聚成一小类
  - 然后,将最“亲密”的小类或个体再聚成一类
  - 重复上述过程,即：把所有的个体和小类聚集成越来越大的类，直到所有的个体都到一起(一大类)为止
  - 随着聚类的进行,类内的“亲密”性在逐渐减低

分层聚类

- 以分解的方式聚类
  - 首先,所有个体都属于一类
  - 其次,将大类中最“疏远”的小类或个体分离出去
  - 然后,分别将小类中最“疏远”的小类或个体再分离出去
  - 重复上述过程,即：把类分解成越来越小的小类，直到所有的个体自成一类为止
  - 随着聚类的进行,类内的亲密性在逐渐增强

分层聚类

- “亲疏”程度的衡量对象
  - 个体间距离
  - 个体和小类间、小类和小类间的距离
    - 最短距离法(nearest neighbor):
      - 两类间的距离定义为两类中距离最近的两个个案之间的距离
    - 最长距离法(furthest neighbor):
      - 两类间的距离定义为两类中距离最远的两个个案之间的距离

分层聚类

- 个体和小类、类和类间的距离
  - 平均链锁法(within-groups linkage)
    - 两类之间的距离定义为两类个案之间距离的平均值。包括：
      - 组间平均链锁法(between-groups linkage)：只考虑两类间个案的距离
      - 组内平均链锁法 (With-groups linage)：考虑所有个案间的距离

分层聚类

- 聚类数目的确定
  - 聚类数目确定尚无统一标准，一般原则：
    - 各类所包含的元素都不应过多
    - 分类数目应符合分析的目的
  - 分层聚类中可以将类间距离作为确定类数目的辅助工具
    - 聚类过程中类间距离呈增加趋势
    - 类间距离小，类的相似性大；距离大，相似性小
    - 绘制碎石图（X轴为类距离，Y轴为类数）

第三节 快速聚类

* 出发点：希望克服分层聚类在大样本时产生的困难，提高聚类效率

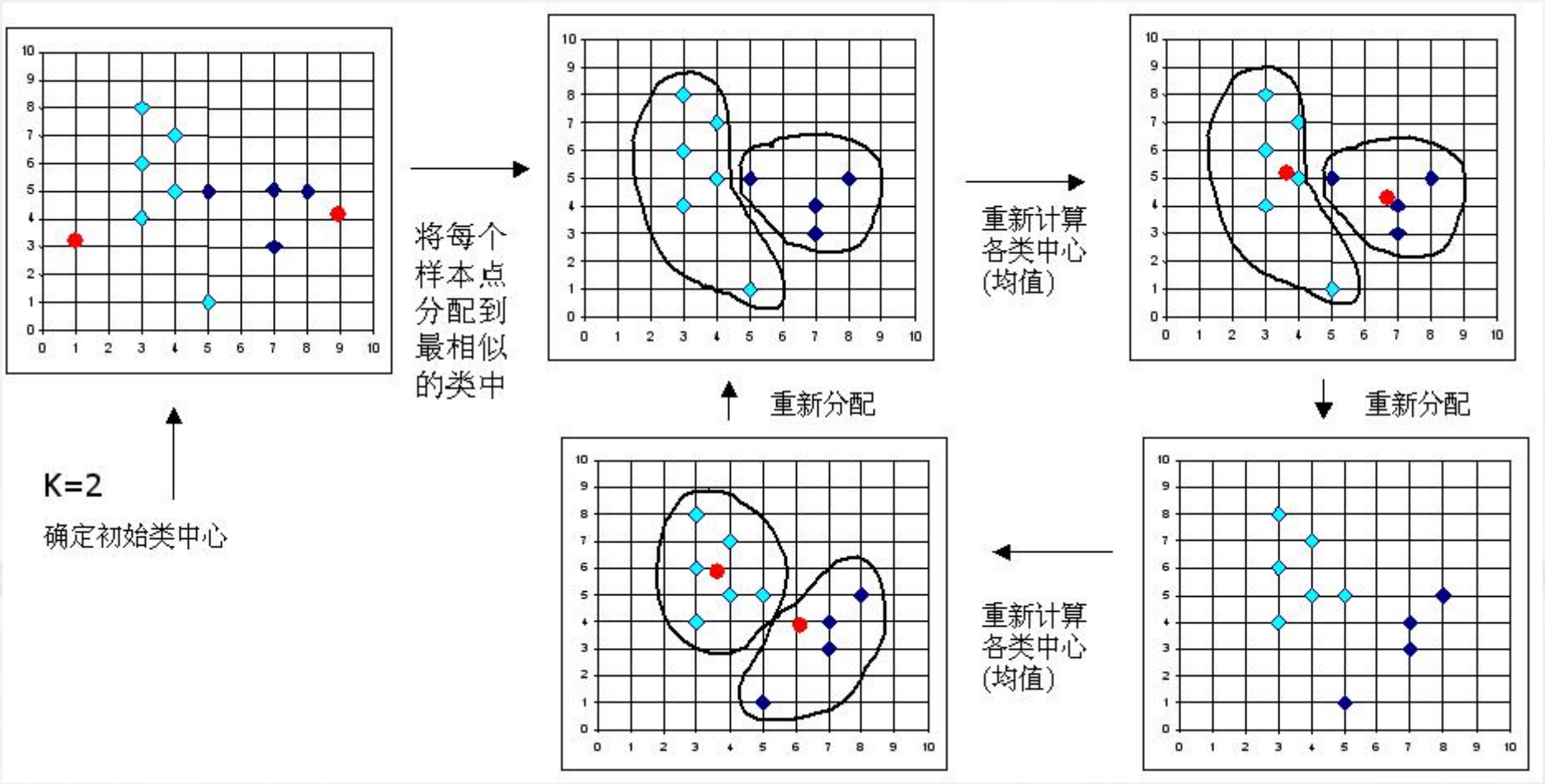
* 做法：
  * 通过用户事先指定聚类数目的方式提高效率
  * 因此，分层聚类可以对不同的聚类数而产生一系列的聚类解，而快速聚类只能产生单一的聚类解
  * K-means快速聚类

* 思路：

1. 指定最后要聚成K类
2. 指定k个样本作为初始类中心
3. 按照距k个中心距离最近的原则把每个样本分派到各中心所在的类中去,形成一个新的k类，完成一次迭代
4. 重新计算k个类的类中心(计算每类各变量的均值,以均值点作为类中心)
5. 重复3步和4步，直到达到指定的迭代次数或达到终止迭代的条件

- 达到指定迭代次数(maximum iteration),默认10次。
- 收敛标准(convergence),默认0.02，即：本次迭代产生的任意新类,各中心位置变化较小.其中最大的变化率小于2%.

K-means快速聚类


K=2
确定初始类中心
将每个样本点分配到最相似的类中
重新计算各类中心(均值)
重新分配
重新计算各类中心(均值)
重新分配

第七章

SPSS的判别分析

主要内容

第一节　判别分析概述

第二节　距离判别法

第三节　**Fisher**判别法

第四节　**Bayes**判别法

第一节 判别分析概述

- 判别分析是多元统计分析中实现数据分类的方法
  - 例如：不同类型客户的预测应用
- 特点：
  - 数据中包含用于预测的变量(自变量)，称为判别变量（定距）
  - 数据中包含已知所属类别的变量(因变量)，称为类别变量（定类，不同类别依次用整数表示）
  - 判别分析可以根据已有数据，确定类别变量与判别变量之间的数量关系，建立判别函数，并可通过判别函数实现对未知数据类别的预测

- 判别分析与聚类分析的不同点：
  - 聚类分析中的类别是未知的，完全通过数据来确定；判别分析，通过对已知类别的“训练样本”的学习，建立判别准则，具有“预测”意义
- 判别分析的一般要求：
  - 判别变量不应有较强的相关性
  - 判别变量服从正态分布
- 判别分析方法的划分：
  - 根据类数：两组判别分析、多组判别分析
  - 根据模型：线性判别、非线性判别
  - 根据判别准则：距离判别法、Fisher判别法、Bayes判别法

判别分析中的数据

- 设有分别来自k≥2个总体的k个样本，每个样本都有关于X_1，X_2，…，X_p的判别变量(p＞k)
- 总样本量为n，各样本的样本量为n_i(i=1,2,..k)
  - 例：设有两个总体G_1和G_2,从G_1中抽取n个观测，从G_2中抽取m个观测；有p个判别变量

第一组样本	$x_{11}^{(1)}$ $x_{12}^{(1)}$ … $x_{1p}^{(1)}$ $x_{21}^{(1)}$ $x_{21}^{(1)}$ … $x_{2p}^{(1)}$ … $x_{n1}^{(1)}$ $x_{n2}^{(1)}$ … $x_{np}^{(1)}$	第二组样本	$x_{11}^{(2)}$ $x_{12}^{(2)}$ … $x_{1p}^{(2)}$ $x_{21}^{(2)}$ $x_{21}^{(2)}$ … $x_{2p}^{(2)}$ … $x_{m1}^{(2)}$ $x_{m2}^{(2)}$ … $x_{mp}^{(2)}$
均值	$\bar{x}_1^{(1)}$ $\bar{x}_2^{(1)}$ … $x_p^{(1)}$	均值	$\bar{x}_1^{(2)}$ $\bar{x}_2^{(2)}$ … $x_p^{(2)}$

第二节 距离判别法

- 思路：
  - 将n个观测数据看成p维空间中的点，计算每个类别的中心(类别均值)
  - 计算任一观测点到各个类别中心的距离(通常采用平方马氏距离)
  - 根据距离最近的原则，距离哪个中心近，则属于哪个类
- 例：设$\mu^{(1)}$，$\mu^{(2)}$，$\Sigma^{(1)}$，$\Sigma^{(2)}$分别为G_1和G_2的均值向量和协差阵,则点X到G_i的距离定义为平方马氏距离为：

$$D^2(X,G_i)=(\mathbf{X}-\boldsymbol{\mu}^{(i)})'(\sum{}^{(i)})^{-1}(\mathbf{X}-\boldsymbol{\mu}^{(i)}) \quad i=1,2$$

u未知时用样本均值替代

为什么采用马氏距离

- 当各维度存在数量级的差异时，欧氏距离不恰当
  - 马氏距离：

$$D(X,G_i)=\sqrt{\frac{(x_1-\mu_1)^2}{S_1^2}+\frac{(x_2-\mu_2)^2}{S_2^2}+\ldots+\frac{(x_p-\mu_p)^2}{S_p^2}}$$

  - 除以方差

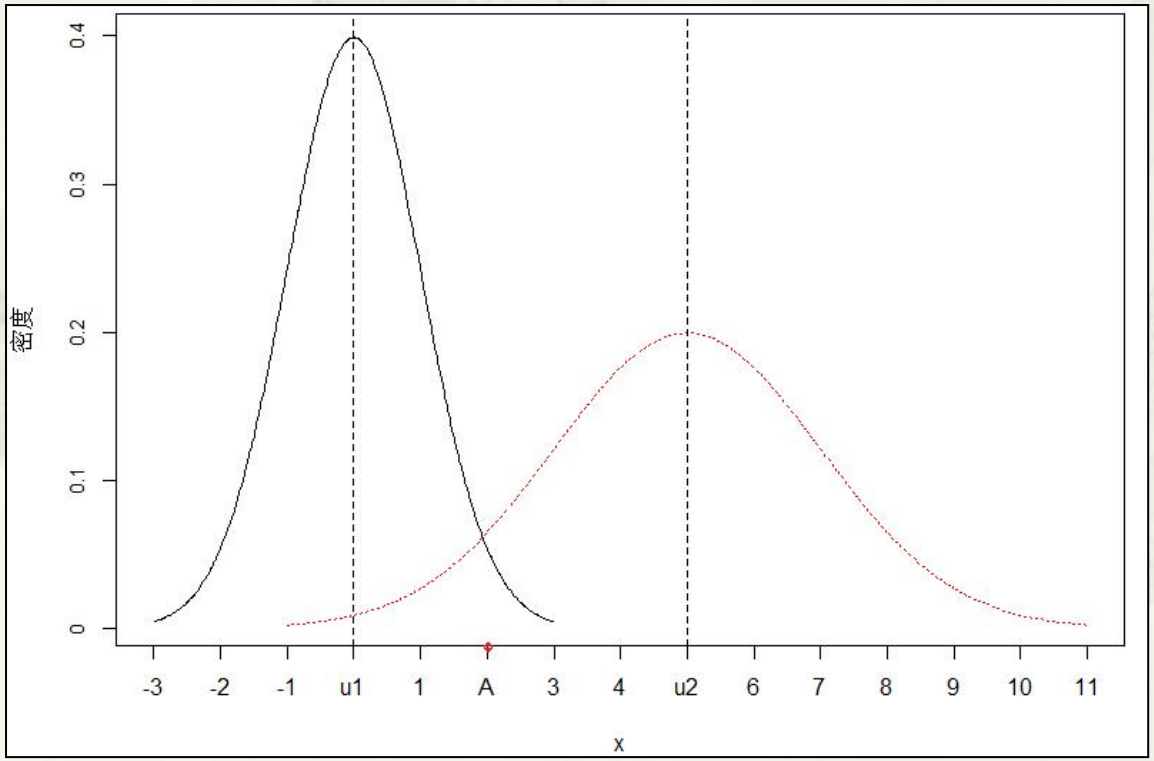
$$D^2(X,G_i)=(\mathbf{X}-\boldsymbol{\mu}^{(i)})'(\sum{}^{(i)})^{-1}(\mathbf{X}-\boldsymbol{\mu}^{(i)})\quad i=1,2$$

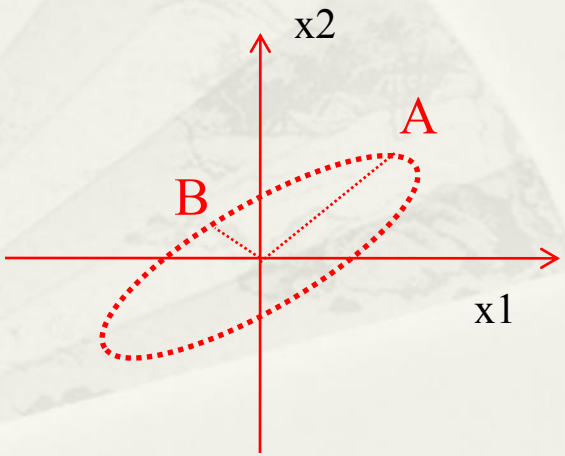
为什么采用马氏距离

- 体现了从概率角度出发的距离(A距u_2更近,A,B距离相等)

$$D^2(X,G_i)=(\mathbf{X}-\boldsymbol{\mu}^{(i)})'(\sum{}^{(i)})^{-1}(\mathbf{X}-\boldsymbol{\mu}^{(i)}) \quad i=1,2$$

  - 例如：均值为0标准差为1以及均值为5标准差为2





距离判别法

- 根据$D^2(X,G_1)$、$D^2(X,G_2)$判断：
  - 如果$D^2(X,G_1)<D^2(X,G_2)$，则：$X\in G_1$
  - 如果$D^2(X,G_2)<D^2(X,G_1)$，则：$X\in G_2$
  - 如果$D^2(X,G_1)=D^2(X,G_2)$，则待判
- 判别函数：$W(X)=D^2(X,G_2)-D^2(X,G_1)$，判断：
  - 如果$W(X)>0$，则：$X\in G_1$
  - 如果$W(X)<0$，则：$X\in G_2$
  - 如果$W(X)=0$，则待判

距离判别法

- 若各组协差阵相等：
  - 采用Σ(pooled within-groups covariance)

$$\sum = \frac{1}{n_1 + n_2 - 2}(S_1 + S_2)\ (\text{S为SSCP}); S_i = \sum_{j=1}^{n_i}(X_j^{(i)} - \bar{X}^{(i)})(X_j^{(i)} - \bar{X}^{(i)})'\ \ i = 1,2$$

- 当采用Σ时，代入距离判别函数，整理：

$$W(X) = 2(X - \bar{X})'\sum{}^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) \Rightarrow W(X) = (X - \bar{X})'\sum{}^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$$

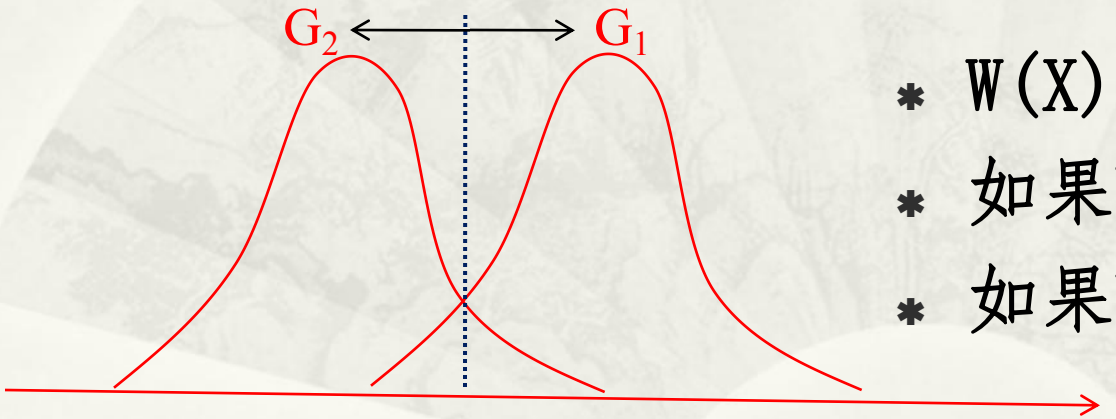
$$\bar{X} = \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})$$

$$W(X) = (X - \bar{X})'\sum{}^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$$

$$= (X - \bar{X})'a = a'(X - \bar{X})$$

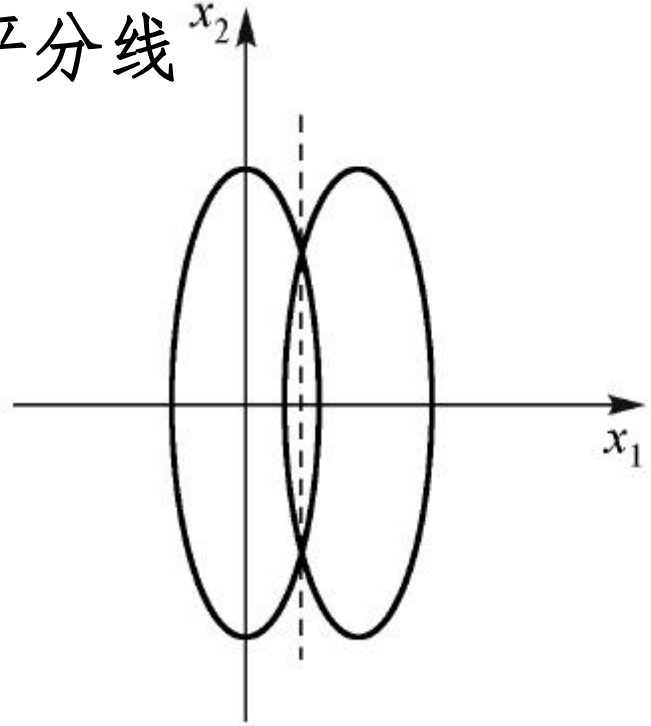
距离判别法

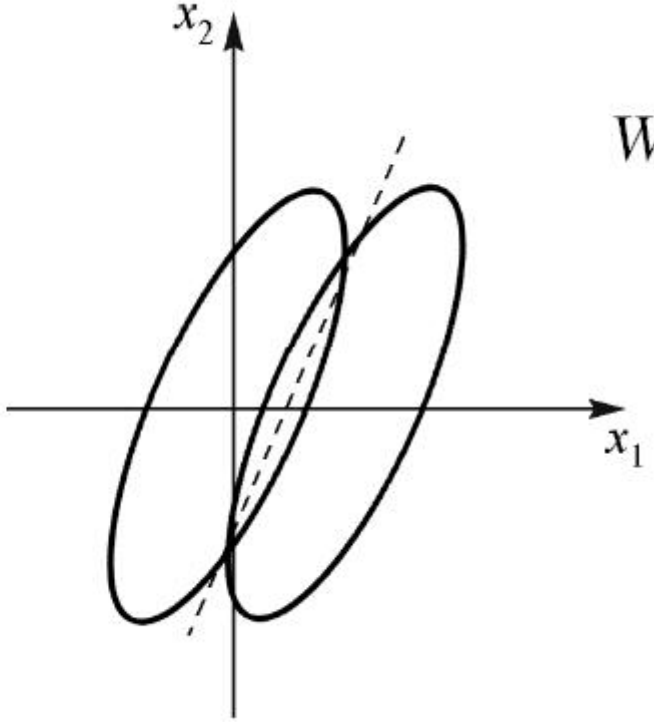
* 当采用Σ时,判别函数为线性函数
  * 判别函数表示的直线为两总体中心连线的垂直平分线（$\sigma_1^2 = \sigma_2^2$）



* W(X)>0,则:X∈G_1
* 如果W(X)<0,则:X∈G_2
* 如果W(X)=0,则待判

判别函数等于0表示的直线为两总体中心连线的垂直平分线





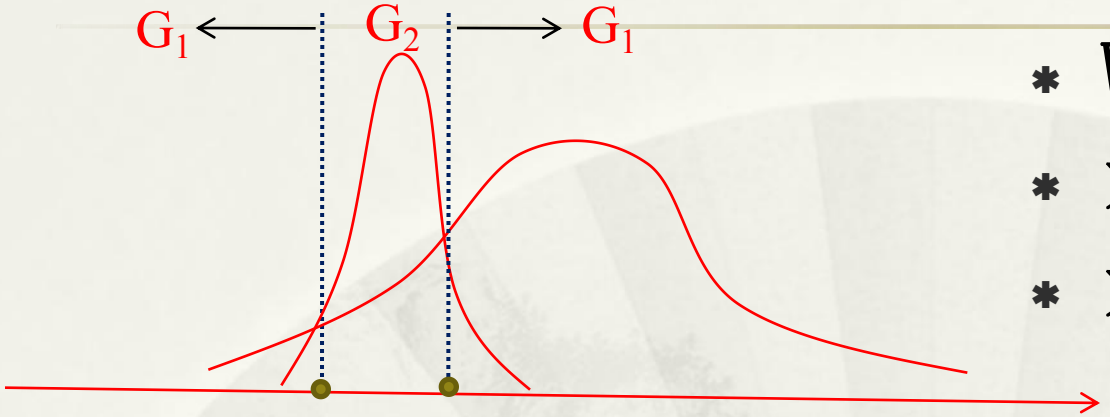
$$W(X)=(X-\overline{X})'\sum{}^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$$
$$=(X-\overline{X})'a=a'(X-\overline{X})$$

- W(X)>0,则:X∈G_1
- 如果W(X)<0,则:X∈G_2
- 如果W(X)=0,则待判
- 若两类别的均值无显著差异，错判概率高

* 若各组协差阵不相等：

  * 采用(separated-groups covariance)

$$W(X)=(X-\bar{X})'\sum{}^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$
$$=(X-\bar{X})'a=a'(X-\bar{X})$$



* W(X)>0，则：X∈G_1
* 如果W(X)<0，则：X∈G_2
* 如果W(X)=0，则待判

  * 判别函数(非线性)为：

$$W(X)=(X-\mu^{(i)})'(\sum{}^{(i)})^{-1}(X-\mu^{(i)})-(X-\mu^{(j)})'(\sum{}^{(j)})^{-1}(X-\mu^{(j)})$$

* 距离判别法的特点：直观
* 问题：
  * 多个总体的均值是否存在显著差异
  * 多个总体的协差阵是否存在显著差异

* 多个总体的均值检验:H_0:μ(1)=…= μ(k)
  * Wilks λ统计量：Wilks λ =|SSE|/|SSG+SSE|，服从Wilks分布。

$$SSG=\sum_{m=1}^{k} n_m(\bar{x}^{(m)}-\bar{x})^2 \qquad SSE=\sum_{m=1}^{k}\sum_{i=1}^{n_m}(x_i^{(m)}-\bar{x}^{(m)})^2$$

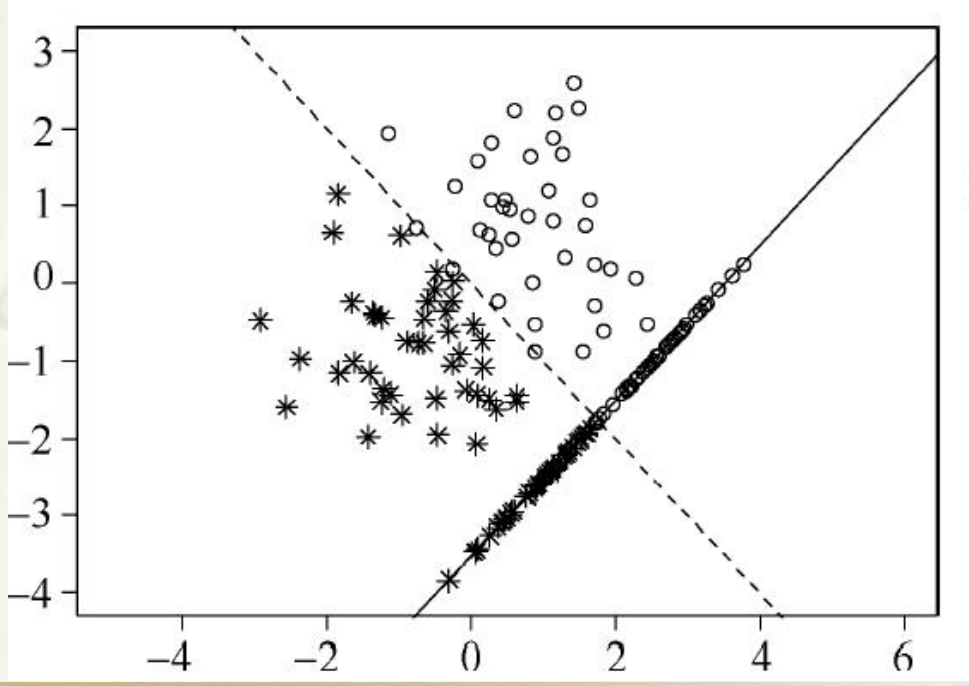
* SPSS输出一元单因素方差分析表(F统计量)
* 多个总体的协差阵检验:BOX’s检验

$$\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$$

  * H0:
  * 若协差阵等，合并的协差阵的广义方差与各类别广义方差差异不显著
  * 统计量：M（近似服从F分布）

第三节　Fisher判别法

- Fisher判别也称典型判别
- 基本思想是先投影的距离判别
  - 将原来p维X空间的判别变量通过线性变换投影到m(m<p)维Y空间中



$$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$$

- 系数如何确定
  - 可否参考主成分分析法
- m如何确定

Fisher判别法

* Fisher判别模型，是判别变量的线性函数形式：

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$$

  - 系数a_i称为判别系数，表示各判别变量对于判别函数的影响
  - Y反映的是样本在低维空间中某个维度上的坐标
  - 判别函数通常为多个，得到在低维空间中多个维度上的坐标，决定了预测点在低维空间中的位置

Fisher判别法

- 寻求能够将总体尽可能分开的方向
  - 首先在判别变量的p维空间中，找到某个线性组合，使各类别的平均值差异最大，作为判别的第一维度，代表判别变量组间方差中的最大部分，得到第一判别函数
  - 然后，按照同样规则依次找到第二判别函数、第三判别函数等，这些判别函数之间相互独立
- 得到的每个函数都可以反映判别变量组间方差的一部分，各判别函数反应的组间方差比例之和为100%
- 前面的判别函数相对重要，后面的判别函数只代表很少一部分方差，可以被忽略

* 示例：设有两个总体G_1和G_2，从G_1中抽取n个观测，从G_2中抽取m个观测；有p个判别变量

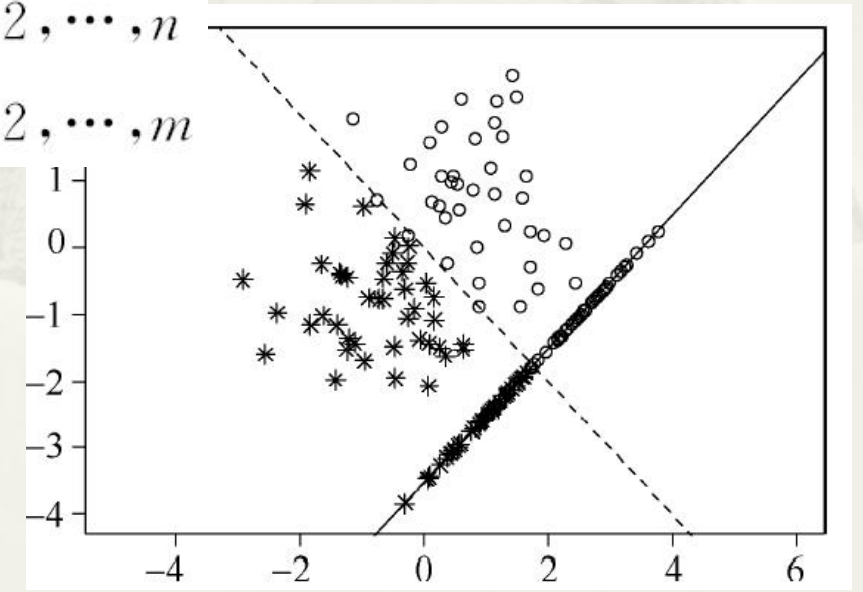
第一组样本	$x_{11}^{(1)}$	$x_{12}^{(1)}$	…	$x_{1p}^{(1)}$	第二组样本	$x_{11}^{(2)}$	$x_{12}^{(2)}$	…	$x_{1p}^{(2)}$
	$x_{21}^{(1)}$	$x_{21}^{(1)}$	…	$x_{2p}^{(1)}$		$x_{21}^{(2)}$	$x_{21}^{(2)}$	…	$x_{2p}^{(2)}$
			…					…	
	$x_{n1}^{(1)}$	$x_{n2}^{(1)}$	…	$x_{np}^{(1)}$		$x_{m1}^{(2)}$	$x_{m2}^{(2)}$	…	$x_{mp}^{(2)}$
均值	$\bar{x}_1^{(1)}$	$\bar{x}_2^{(1)}$	…	$x_p^{(1)}$	均值	$\bar{x}_1^{(2)}$	$\bar{x}_2^{(2)}$	…	$x_p^{(2)}$

* 若判别函数为：$y = a_1 x_1 + a_2 x_2 + \cdots + a_p x_p$
* 将属于不同两类的观测代入判别函数：

$$y_i^{(1)} = a_1 x_{i1}^{(1)} + a_2 x_{i2}^{(1)} + \cdots + a_p x_{ip}^{(1)}, i = 1, 2, \cdots, n$$

$$y_i^{(2)} = a_1 x_{i1}^{(2)} + a_2 x_{i2}^{(2)} + \cdots + a_p x_{ip}^{(2)}, i = 1, 2, \cdots, m$$

$$\bar{y}^{(1)} = \sum_{i=1}^{p} a_i \bar{x}_i^{(1)}; \bar{y}^{(2)} = \sum_{i=1}^{p} a_i \bar{x}_i^{(2)}$$



- 为使判别函数很好区分来自两个总体的样本,希望:
  - $\bar{y}^{(1)}$和$\bar{y}^{(2)}$相差越大越好
  - 组内的离差平方和越小越好
  - 下式越大越好

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

- 利用求极值原理，可以求出使I达到最大时的系数a

- 一般陈述：
  - 点x在以a为法方向的投影为a’x,则各组数据的投影为：

$$G_i : a'x_1^{(i)} \cdots a'x_{n_i}^{(i)}, i = 1,...,k$$

  - 将G_m组中数据投影的均值记为 $a'\overline{x}^{(m)}$ 有：

$$a'\overline{x}^{(m)} = \frac{1}{n_m}\sum_{i=1}^{n_m} a'x_i^{(m)}, m = 1,...,k$$

  - 记k组数据投影的总均值为 $a'\overline{x}$ 有：

$$a'\overline{x} = \frac{1}{n}\sum_{m=1}^{k}\sum_{i=1}^{n_m} a'x_i^{(m)}$$

- 组间离差平方和：

$$SSG = \sum_{m=1}^{k} n_m (a'\overline{x}^{(m)} - a'\overline{x})^2$$

$$= a'[\sum_{m=1}^{k} n_m (\overline{x}^{(m)} - \overline{x})(\overline{x}^{(m)} - \overline{x})']a = a'Ba;$$

- 组内离差平方和：

$$SSE = \sum_{m=1}^{k} \sum_{i=1}^{n_m} (a'x_i^{(m)} - a'\overline{x}^{(m)})^2$$

$$= a'[\sum_{m=1}^{k} \sum_{i=1}^{n_m} (x_i^{(m)} - \overline{x}^{(m)})(x_i^{(m)} - \overline{x}^{(m)})']a = a'Ea;$$

- 希望寻找a使得SSG尽可能大而SSE尽可能小，即：

$$\Delta(a)=\frac{a'Ba}{a'Ea}\rightarrow \max$$

使 $\frac{a'Ba}{a'Ea}$ 最大的值为方程$|B-\lambda E|=0$的最大特征值根λ_1

- 记方程$|B-\lambda E|=0$的全部特征值为$\lambda_1\geq\ldots\geq\lambda_r>0$，相应的特征向量为$v_1,\ldots,v_r$. 则判别函数为：$y_i(x)=v_i'x$ (= $a'x$)
- 记p_i为第i个判别函数的判别能力(效率)：$p_i=\frac{\lambda_i}{\sum_{h=1}^{r}\lambda_h}$
- m个判别函数的判别能力为：

$$\sum_{i=1}^{m}p_i=\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{h=1}^{r}\lambda_h}$$

第四节 Bayes判别法

- 在认为所有k个类别都是空间互斥的子域的条件下，利用贝叶斯方法进行判别
- 贝叶斯方法是一种研究不确定性问题的决策方法
  - 通过贝叶斯概率描述不确定性
  - 引进效用函数（Utility Function）
  - 选择使期望效用最大的最优决策
- 贝叶斯概率
  - 一种主观概率：对事物发生概率的主观估计
  - 主观概率取决于先验知识的正确性和后验知识的丰富性

Bayes判别

- 贝叶斯概率
  - 首先,用先于数据的概率描述最初的不确定性
  - 然后,将其和试验数据相结合，产生一个后于数据的修订了的概率
  - 不确定性须用概率来描述，不确定性的表述须与概率论的运算规则相结合
- 贝叶斯公式
  - 事件A与事件B独立 $P(AB) = P(A)P(B)$
  - 事件A与事件B不独立

$$P(AB) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

Bayes判别

- 贝叶斯公式

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{P(A)P(B|A)}{P(B)}=\frac{P(A)P(B|A)}{\sum_{i=1}^{k}P(A_i)P(B|A_i)}$$

  - P(A)称为先验概率;P(B|A)为条件概率，在贝叶斯判别中为似然函数;P(A|B)为后验概率

  - 后验概率可看做一种简化的效用函数
  - 最大后验概率假设是贝叶斯决策的依据

Bayes判别

- 设：
  - 有k个总体$G_1, G_2 \cdots, G_k$，观测从属于各总体的先验概率分别为$q_1, q_2, \cdots q_k$；
  - 似然函数为$f_1(X), f_2(X), \cdots f_k(X)$(x连续为密度)

- 则：样本x来自第g总体的后验概率为：

$$P(g \mid x) = \frac{q_g f_g(X)}{\sum_{i=1}^{k} q_i f_i(X)}, g = 1,2,...k$$

- 对于新样本分别计算其落入各个子域的后验概率，其所归属的类别为后验概率最大的类别（总体）

- 先验概率：一般为等概率(熵最大原则)
- 计算似然函数：

$$P(g \mid x) = \frac{q_g f_g(X)}{\sum_{i=1}^{k} q_i f_i(X)}, g = 1,2,...k$$

$$p(X \mid G_1) = \frac{1}{|\Sigma| \sqrt{2\pi}} \exp\left[-\frac{1}{2} (X - \mu^{(1)})' (\Sigma)^{-1} (X - \mu^{(1)}) \right]$$

$$p(X \mid G_1) = \frac{1}{|\Sigma| \sqrt{2\pi}} \exp\left[-\frac{1}{2} D_1^2 \right] \qquad p(X \mid G_2) = \frac{1}{|\Sigma| \sqrt{2\pi}} \exp\left[-\frac{1}{2} D_2^2 \right]$$

- 计算后验概率

$$p(G_i \mid X) = \frac{q_i p(X \mid G_i)}{\sum_{j=1}^{k} q_j p(X \mid G_j)}, \quad i = 1, 2, \cdots, k$$

$p(X \mid G_i)$ 与 $\exp\left[-\frac{1}{2} D_i^2 \right]$

$$p(G_i \mid X) = \frac{q_i \exp(-D_i^2/2)}{\sum_{j=1}^{k} q_j \exp(-D_j^2/2)}, \quad i = 1, 2, \cdots, k$$

- 只根据对分子计算对数的结果即可判断

第八章

因子分析

因子分析的提出

- 为尽可能完整描述一个事物，往往要收集相关的许多指标
- 多指标产生的问题：
  - 计算处理麻烦
  - 信息重叠
- 从众多的指标中剔除一些指标会造成信息丢失

因子分析的基本思想

- 因子分析的基本出发点
  - 将原始指标综合成较少的指标，这些指标能够反映原始指标的绝大部分信息（方差）
  - 这些综合指标之间没有相关性
- 因子变量的特点
  - 这些综合指标称为因子变量，是原变量的重造
  - 个数远远少于原变量个数，但可反映原变量的绝大部分方差
  - 不相关性
  - 可命名解释性

因子分析的基本步骤

- 确认待分析的原始变量是否适合作因子分析
- 构造因子变量
- 利用旋转方法使因子变量具有可解释性
- 计算每个样本的因子变量得分

因子分析的数学模型

- 数学模型（x_i为标准化的原始变量；F_i为因子变量；k<p）

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + a_{13}f_3 + \ldots + a_{1k}f_k + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + a_{23}f_3 + \ldots + a_{2k}f_k + \varepsilon_2 \\ x_3 = a_{31}f_1 + a_{32}f_2 + a_{33}f_3 + \ldots + a_{3k}f_k + \varepsilon_3 \\ \ldots\ldots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + a_{p3}f_3 + \ldots + a_{pk}f_k + \varepsilon_p \end{cases}$$

F：因子变量
A：因子载荷阵
$\mathbf{a_{ij}}$：因子载荷
ε：特殊因子

也可以矩阵的形式表示为：

$X=AF+\varepsilon$

因子分析的相关概念

- 因子载荷
  - 在因子变量不相关的条件下，a_{ij}就是第i个原始变量与第j个因子变量的相关系数。a_{ij}绝对值越大，则X_i与F_i的关系越强
- 变量的共同度(Communality)
  - 也称公因子方差。X_i的变量共同度为因子载荷矩阵A中第i行元素的平方和

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2$$

X_i的共同度反应了全部因子变量对X_i总方差的解释能力

因子分析的相关概念

- 因子变量F_j的方差贡献
  - 因子变量F_j的方差贡献为因子载荷矩阵A中第j列各元素的平方和

$$S_j = \sum_{i=1}^{p} a_{ij}^2$$

可见：因子变量F_j的方差贡献体现了同一因子F_j对原始所有变量总方差的解释能力

S_j/p表示了第j个因子解释原所有变量总方差的比例

是否适合作因子分析

- 计算原有变量的相关系数矩阵
  - 一般小于0.3就不适合作因子分析

确定因子变量–主成份分析

- 主成份分析法的数学模型

$$\begin{cases} y_1 = \mu_{11}x_1 + \mu_{12}x_2 + \mu_{13}x_3 + \ldots \mu_{1p}x_p \\ y_2 = \mu_{21}x_1 + \mu_{22}x_2 + \mu_{23}x_3 + \ldots \mu_{2p}x_p \\ y_3 = \mu_{31}x_1 + \mu_{32}x_2 + \mu_{33}x_3 + \ldots \mu_{3p}x_p \\ \ldots\ldots \\ y_p = \mu_{p1}x_1 + \mu_{p2}x_2 + \mu_{p3}x_3 + \ldots \mu_{pp}x_p \end{cases}$$

将原有的P个相关变量Xi作线性变换后转成另一组不相关的变量Yi

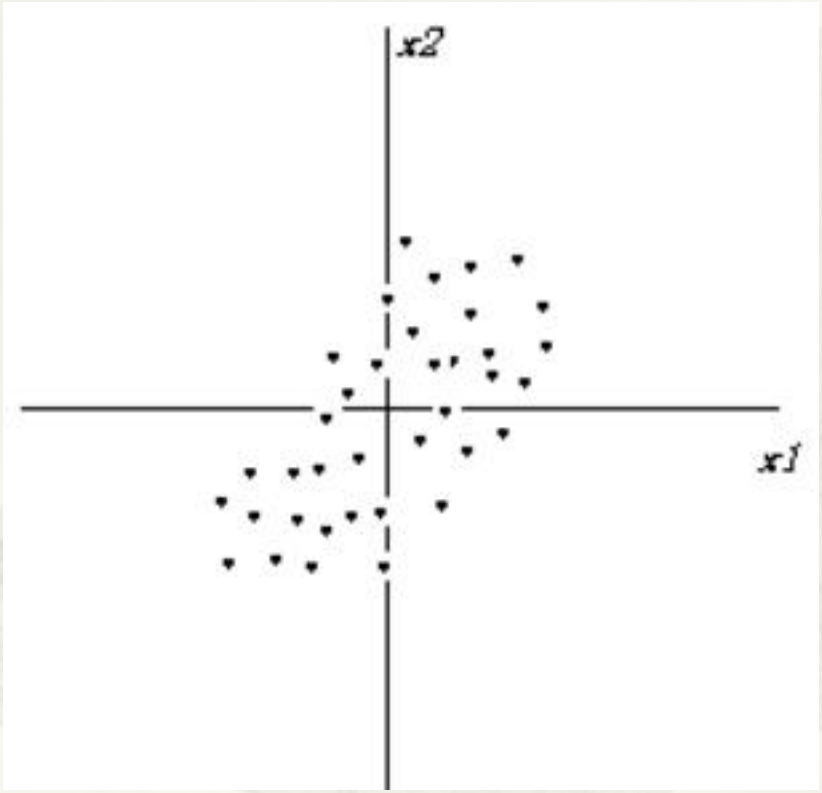
该方程组要求：

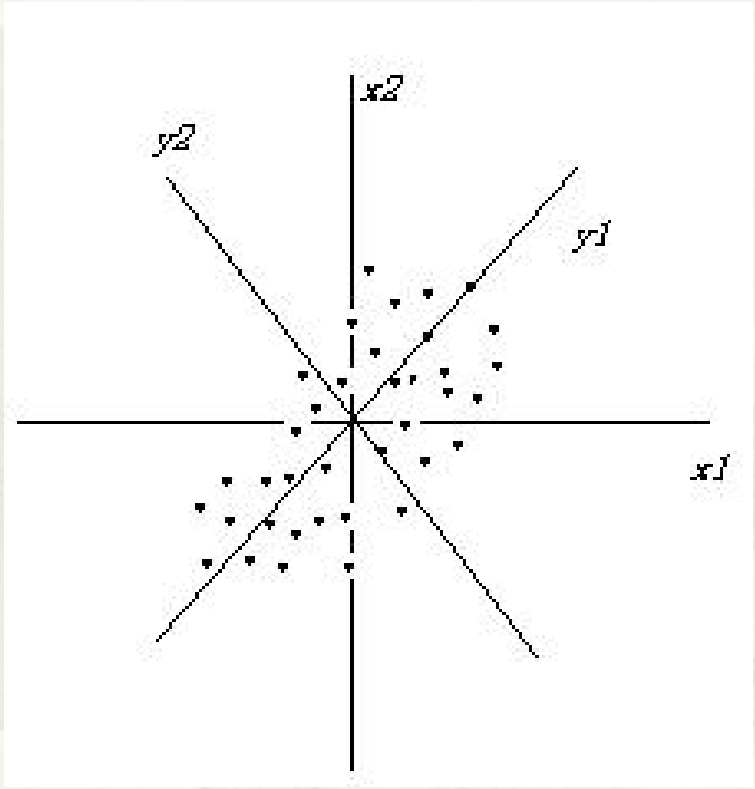
$$\mu_{i1}^2 + \mu_{i2}^2 + \mu_{i3}^2 + \ldots + \mu_{ip}^2 = 1 \ (i = 1,2,3\ldots, p)$$

确定因子变量–主成份分析

- 系数u_{ij}依照两个原则来确定
  - y_i与y_j $(i \neq j, i, j=1,2,3,\cdots p)$互不相关；
  - y_1是$x_1, x_2, x_3, \cdots, x_p$的一切线性组合（系数满足上述方程组）中方差最大的；y_2是与y_1不相关的$x_1, x_2, x_3, \cdots, x_p$的一切线性组合中方差次大的；$y_P$是与$y_1$，$y_2$，$y_3, \cdots y_p$都不相关的$x_1, x_2, x_3, \cdots, x_p$的一切线性组合中方差最小的；
  - y_1在总方差中所占比例最大，它综合原有变量的能力最强，其余变量在总方差中所占比例依次递减，即：其余变量综合原有变量的能力依次减弱。

确定因子变量–主成份分析





$$\begin{cases} y_1 = x_1 \cos\theta + x_2 \sin\theta \\ y_2 = -x_1 \sin\theta + x_2 \cos\theta \end{cases}$$

确定因子变量-主成份分析

* 主成份分析的基本步骤：
  * 将原始数据标准化
  * 计算变量间简单相关系数矩阵R
  * 求R的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots \lambda_p \geqslant 0$及对应的单位特征向量 μ_1， μ_2， $\mu_3, \cdots \mu_p$
  * 得到：$y_i=u_{i1}x_1+u_{i2}x_2+\cdots+u_{ip}x_p$

确定因子变量—计算因子载荷阵

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}=\begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{21}\sqrt{\lambda_2} & \cdots & u_{p1}\sqrt{\lambda_p} \\ u_{12}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{p2}\sqrt{\lambda_p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{1p}\sqrt{\lambda_1} & u_{2p}\sqrt{\lambda_2} & \cdots & u_{pp}\sqrt{\lambda_p} \end{pmatrix}$$

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pk} \end{pmatrix}=\begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{21}\sqrt{\lambda_2} & \cdots & u_{k1}\sqrt{\lambda_k} \\ u_{12}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{k2}\sqrt{\lambda_k} \\ \cdots & \cdots & \cdots & \cdots \\ u_{1p}\sqrt{\lambda_1} & u_{2p}\sqrt{\lambda_2} & \cdots & u_{kp}\sqrt{\lambda_k} \end{pmatrix}$$

确定因子变量个数

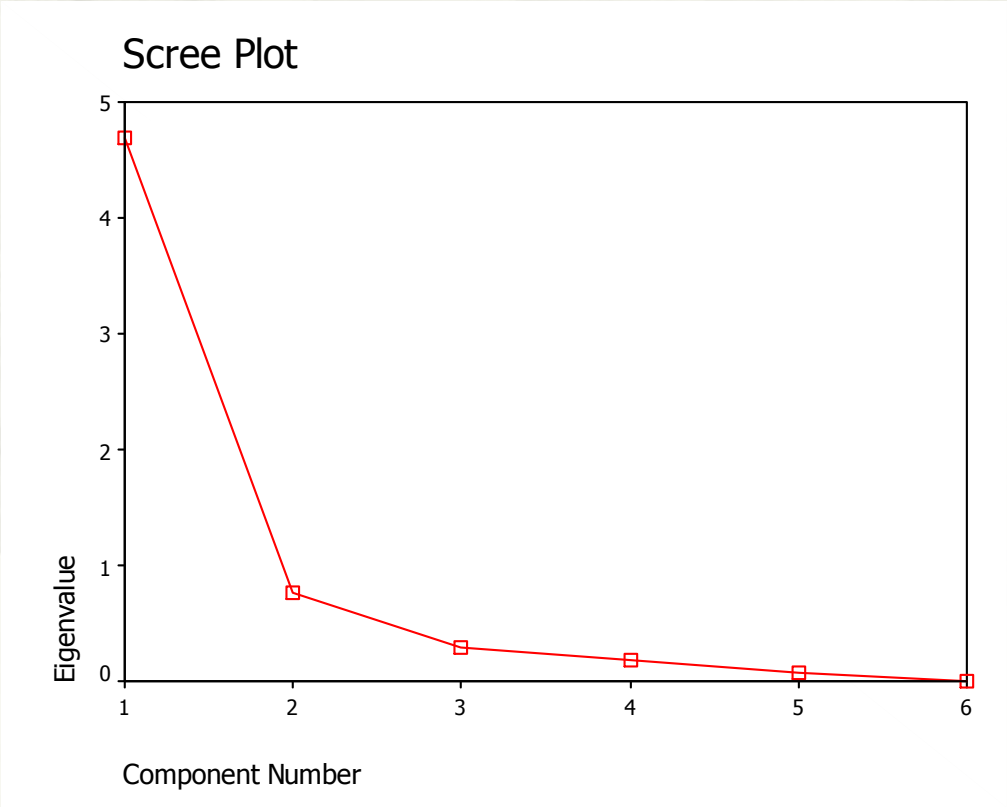
- 确定k个因子变量
  - 根据特征值 λ_i 确定：取特征值大于1的特征根
  - 根据累计贡献率：一般累计贡献率应在70%以上。

$$a_1 = S_1^2 / p = \lambda_1 / \sum_{i=1}^{p} \lambda_i \quad a_2 = (S_1^2 + S_2^2) / p = (\lambda_1 + \lambda_2) / \sum_{i=1}^{p} \lambda_i$$

$$a_k = \sum_{i=1}^{k} S_i^2 / p = \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$$

确定因子变量个数

- 确定k个因子变量
  - 通过观察碎石图的方式确定因子变量的个数。



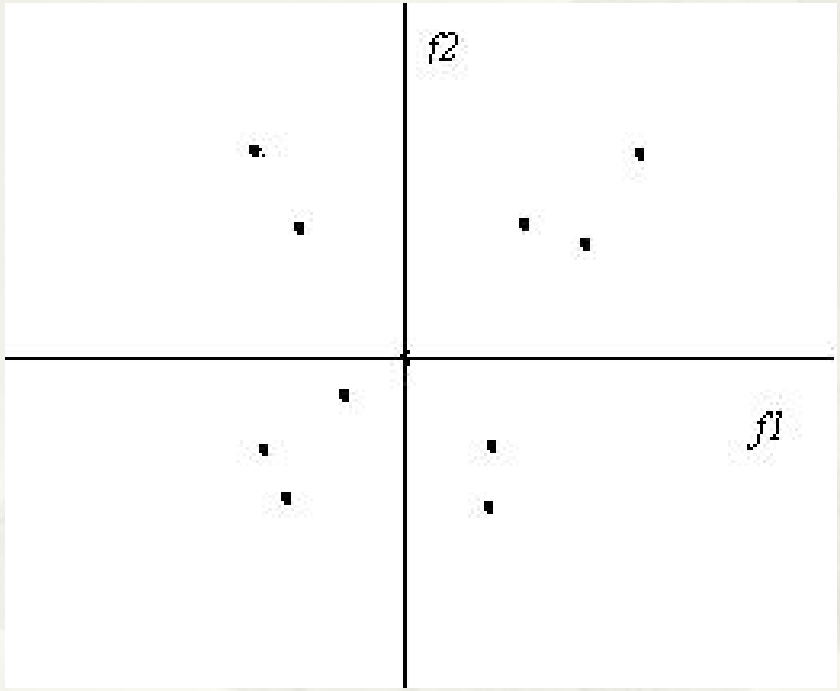
因子变量的命名解释

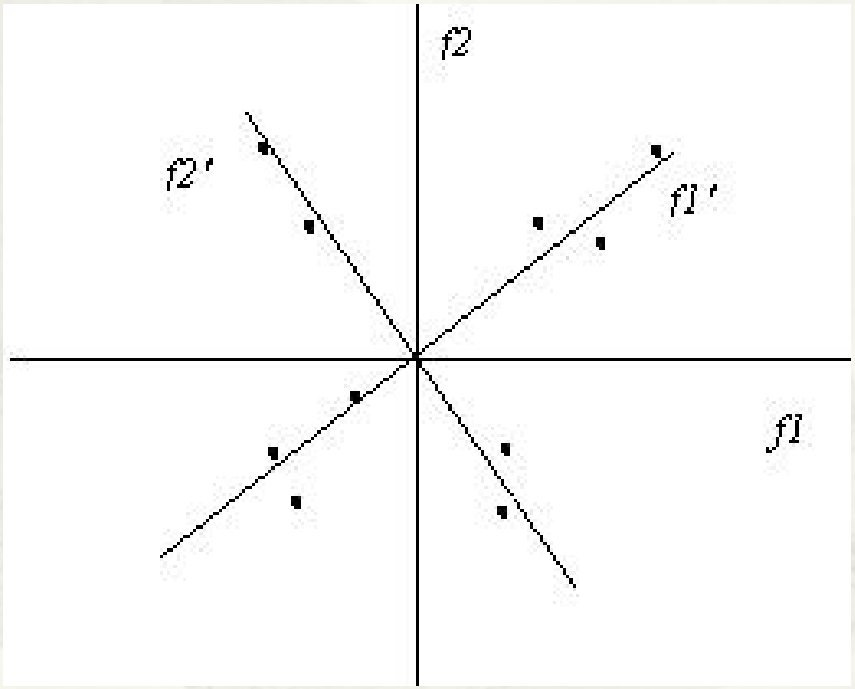
- 发现：
  - a_{ij}的绝对值可能在某一行的许多列上都有较大的取值，或a_{ij}的绝对值可能在某一列的许多行上都有较大的取值。
- 表明：
  - 某个原有变量x_i可能同时与几个因子都有比较大的相关关系，也就是说，某个原有变量x_i的信息需要由若干个因子变量来共同解释；同时，虽然一个因子变量可能能够解释许多变量的信息，但它却只能解释某个变量的一少部分信息，不是任何一个变量的典型代表。
- 结论：因子变量的实际含义不清楚

因子变量的命名解释

- 通过因子旋转使：
  - 每个变量在尽可能少的因子上又比较高的载荷，即：在理想状态下，让某个变量在某个因子上的载荷趋于1，而在其他因子上的载荷趋于0。
- 这样：一个因子变量就能够成为某个变量的典型代表，它的实际含义也就清楚了。
- 因子旋转的目的是通过改变坐标轴的位置，重新分配各个因子所解释的方差比例，使因子结构更简单。
- 因子旋转不改变模型对数据的拟和程度，不改变每个变量的方差共同度

因子变量的命名解释





因子旋转方法

- 因子正交旋转方法和斜交旋转方法
- 方差最大法(正交旋转)：
  - 从简化因子载荷矩阵的每一列出发，使和每个因子有关的载荷平方的方差最大
  - 当只有少数几个变量在某个因子上有较高的载荷时，对因子的解释是最简单的
- 即使正交旋转也不一定使因子含义清晰
- 斜交旋转：因子含义清楚，但允许因子之间相关
- 理论上：斜交旋转优于正交旋转，但如果相关性过高则不可接收，因此正交旋转应用更广泛

计算因子得分

- 因子得分是因子变量构造的最终体现，应给出因子对应每个样本上的值。
- 基本思想：是将因子表示为原有变量的线性组合，即通过因子得分函数计算因子得分
  - 第j个因子在第i个样本上的值表示为：

$$F_{ji} = \varpi_{j1}x_{1i} + \varpi_{j2}x_{2i} + \varpi_{j3}x_{3i} + \ldots + \varpi_{jp}x_{pi}$$

$$(j = 1,2,3,\ldots,k)$$

- 某样本的因子得分可看作各观测变量值的加权平均，权数的大小表示了变量对因子的重要程度

$$F_j = \varpi_{j1}x_1 + \varpi_{j2}x_2 + \varpi_{j3}x_3 + \ldots + \varpi_{jp}x_p$$

$$(j = 1,2,3,\ldots,k)$$

第九章

Logistic回归分析

主要内容

- 第一节　二项Logistic回归及应用
- 第二节　多项Logistic回归及应用

第一节 二项Logistic回归

- 研究二分类变量与其他变量之间的关系
  - 例如：研究吸烟对是否得肺癌的影响，并以年龄和性别作为控制变量，特点：
    - 被解释变量是二值变量
    - 解释变量有分类变量和定距变量
    - 吸烟与肺癌之间并非一种线性关系

- 对二项分类的被解释变量可否直接采用一般多元线性回归分析方法？
  - 结论：不可以

二项Logistic回归

- 当被解释变量为二项(0/1)分类变量时，被变量的取值范围和与自变量的关系问题：
  - 根据回归模型的意义，可知：

$$E(y_i) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \longrightarrow P_{y=1} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

  - 一般回归模型下的被解释变量的取值范围是－∞~＋∞
  - 这里，被解释变量的取值范围是0~1
  - 一般回归分析建立模型，解释变量与*P*间的关系只能是线性的。

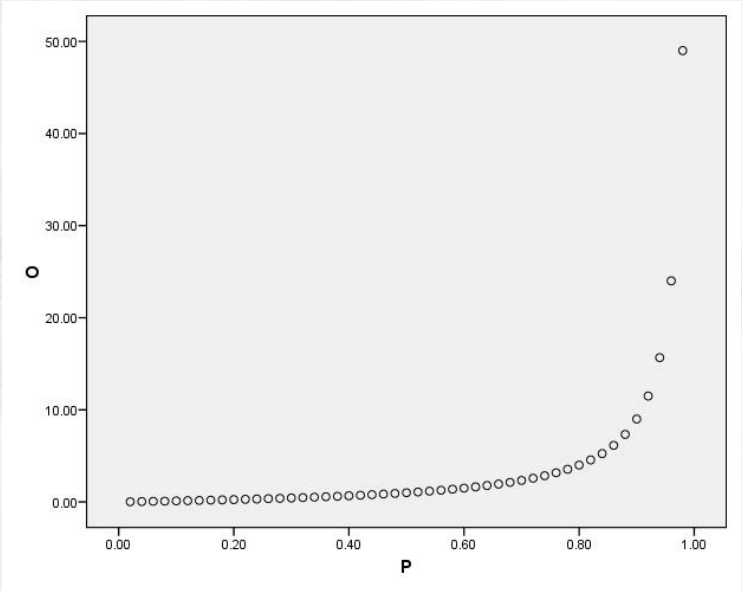
二项Logistic回归

- 解决问题的方向
  - 能否对概率P进行转换处理后，使其取值范围与一般线性回归模型吻合
  - 对概率*P*应采用非线性转化处理
  - 所有的转化都不应改变解释变量和被解释变量之间关系的方向

二项Logistic回归:理论上的处理

- 进行两步转换处理：
  - 第一步，将*P*转换成 *Ω*
    - Ω 称为优势
    - 对P的转化是非线性的
    - Ω 是P的单调增函数
    - 优势的取值范围：0~＋∞

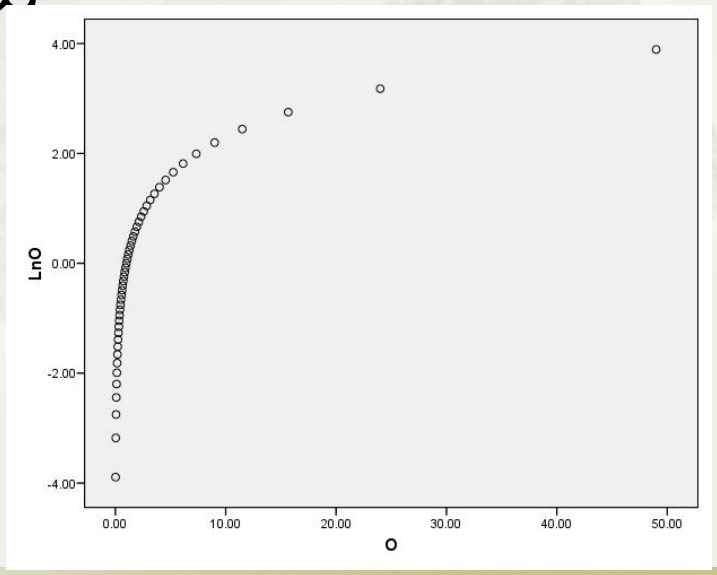
$$\Omega = \frac{P}{1 - P}$$



二项Logistic回归:理论上的处理

- 进行两步转换处理：
  - 第二步，Ω转换成ln Ω
    - ln Ω称为Logit P
    - Logit P与 Ω仍呈增长（或下降）的一致性关系
    - Logit P的取值于－∞~＋∞

$$\ln(\Omega) = \ln(\frac{P}{1-P})$$



二项Logistic回归:理论上的处理

- 二项Logistic模型：

$$\ln(\frac{P}{1-P}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

$$Logit\ P = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

二项Logistic回归

- P与自变量间为非线性关系：

$$\frac{P}{1-P} = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

$$P = (1-P)\exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

$$P = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i) - P * \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

$$P[1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)] = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

$$P = \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}$$

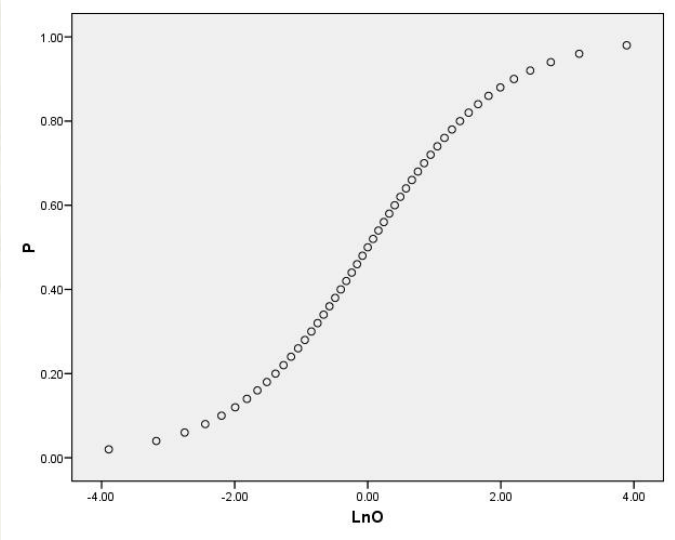
增长曲线

$$P = \frac{1}{1 + \exp[-(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)]}$$

二项Logistic回归系数的含义

- 回归系数表示当其他自变量取值保持不变时，某自变量取值增加一个单位引起Logit P平均变化β_i个单位
  - 在模型的实际应用关心的是自变量变化引起事件发生概率P变化的程度
  - 当自变量x_i变化时，对概率P的影响程度是非线性的，不易直观理解

更注重自变量对发生比Ω的影响



二项Logistic回归系数的含义

- 优势：$\Omega = P/(1-P)$，即某事件发生的概率与不发生的概率之比
  - 利用优势比可以进行组之间风险的对比分析
  - 例如，如果吸烟得肺癌的概率是0.25，不吸烟得肺癌得概率是0.10，则两组的优势比为：

$$OR_{\text{A VS..B}} = \frac{pr(D_A)}{1-pr(D_A)} / \frac{pr(D_B)}{1-pr(D_B)} = \frac{1}{3} / \frac{1}{9} = 3$$

  - 吸烟的风险近似是不吸烟的三倍，吸烟组得肺癌的风险高于不吸烟组

二项Logistic回归系数的含义

- 如果被解释变量y(肺癌1=得/0=没），自变量x只有一个(x1吸烟1=吸烟/0=不吸烟)，则logistic方程为：

$$\text{logit}\,[pr(Y=1)] = \beta_0 + \beta_1 X_1$$

  - 吸烟与不吸烟组的方程分别是：

$$\text{logit}\,[pr(Y=1)] = \ln(odd(\text{smoker})) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

$$\text{logit}\,[pr(Y=1)] = \ln(odd(\text{nonsmkers})) = \beta_0 + \beta_1 \times 0 = \beta_0$$

  - 两组优势比为：

$$\text{OR}_{\text{S VS. NS}} = \frac{odds(\text{smokers})}{odds(\text{nosmokers})} = \frac{e^{(\beta_0+\beta_1)}}{e^{\beta_0}} = e^{\beta_1}$$

  - 可见，当解释变量是1/0二组时，两组间的对比是关于回归方程相应回归系数的对比

二项Logistic回归系数的含义

- 如果被解释变量y(肺癌1=得/0=没），自变量x有三个(x1吸烟/x2年龄/x3性别)，则logistic方程为：

$$\text{logit}\,[pr(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

  - X_A=(1,45,1)与X_B=(0,45,1)的方程分别是：

$$\text{logit}\,[pr(Y=1)] = \ln(odd\,(X_A)) = \beta_0 + \beta_1 \times 1 + \beta_2 \times 45 + \beta_3 \times 1$$

$$\text{logit}\,[pr(Y=1)] = \ln(odd\,(X_B)) = \beta_0 + \beta_1 \times 0 + \beta_2 \times 45 + \beta_3 \times 1$$

  - 两组优势比为：

$$\text{OR}_{\text{XA VS. XB}} = \frac{odds\,(X_A)}{odds\,(X_B)} = e^{(1-0)\beta_1 + (45-45)\beta_2 + (1-1)\beta_3} = e^{\beta_1}$$

  - 这里的主要目的是研究吸烟对肺癌的影响，年龄和性别是作为控制变量存在的，该比率为调整比率，与不包括控制变量在内的比率不相等。（也可将定距变量作观测变量）

二项Logistic回归系数的含义

- 自变量对优势Ω的影响

$$\Omega = \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)$$

  - 当其他解释变量保持不变而研究观测变量变化一个单位对Ω的影响时，可将新的优势设为Ω*，则有优势比为：

$$\frac{\Omega^*}{\Omega} = \exp(\beta_i)$$

    - 即：当x_i增加一个单位时，将引起优势是原来的$\exp(\beta_i)$倍

二项Logistic回归系数的含义

- 如果被解释变量y(肺癌1=得/0=没），自变量x有三个(x1吸烟/x2年龄/x3性别),并考虑吸烟与年龄和对性别的交互作用)，则logistic方程为：

$$\text{logit}\,[pr(Y=1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$$

  - X_A=(1,45,1,1×45,1×1)与X_B=(0,45,1,0×45,0×1)两组的优势比是：

$$\text{OR}_{(\text{XA VS. XB}|\text{age}=45,\text{sex}=1)} = \frac{odds(X_A)}{odds(X_B)} = e^{(1-0)\beta_1+(45-45)\beta_2+(1-1)\beta_3+(45-0)\beta_4+(1-0)\beta_5}$$

$$= e^{\beta_1+45\beta_4+\beta_5} = e^{\beta_1+\beta_4 x_2+\beta_5 x_3}$$

  - 这里涉及到了多个系数，以及控制变量的不同取值

$$\text{OR}_{(\text{XA VS. XB}|\text{age}=35,\text{sex}=0)} = e^{\beta_1+35\beta_4}$$

$$\text{OR}_{(\text{XA VS. XB}|\text{age}=20,\text{sex}=1)} = e^{\beta_1+20\beta_4+\beta_5}$$

二项Logistic回归的参数估计

- 采用极大似然估计法进行参数估计：似然函数值
  - 例如：通过样本数据对购买的比例θ进行估计，其总体服从参数为θ的二项分布。假设θ只有0.2和0.6两个取值，则：

$$pr(Y;\theta)=C_m^y\theta^y(1-\theta)^{m-y} \quad pr(Y;0.2)=C_m^y 0.2^y(1-0.2)^{m-y} \quad pr(Y;0.6)=C_m^y 0.6^y(1-0.6)^{m-y}$$

似然估计函数

  - 如果m=5，则
  - 如果y=4，则θ=0.6

$$pr(Y;\hat{\theta}) > pr(Y;\theta^*)$$

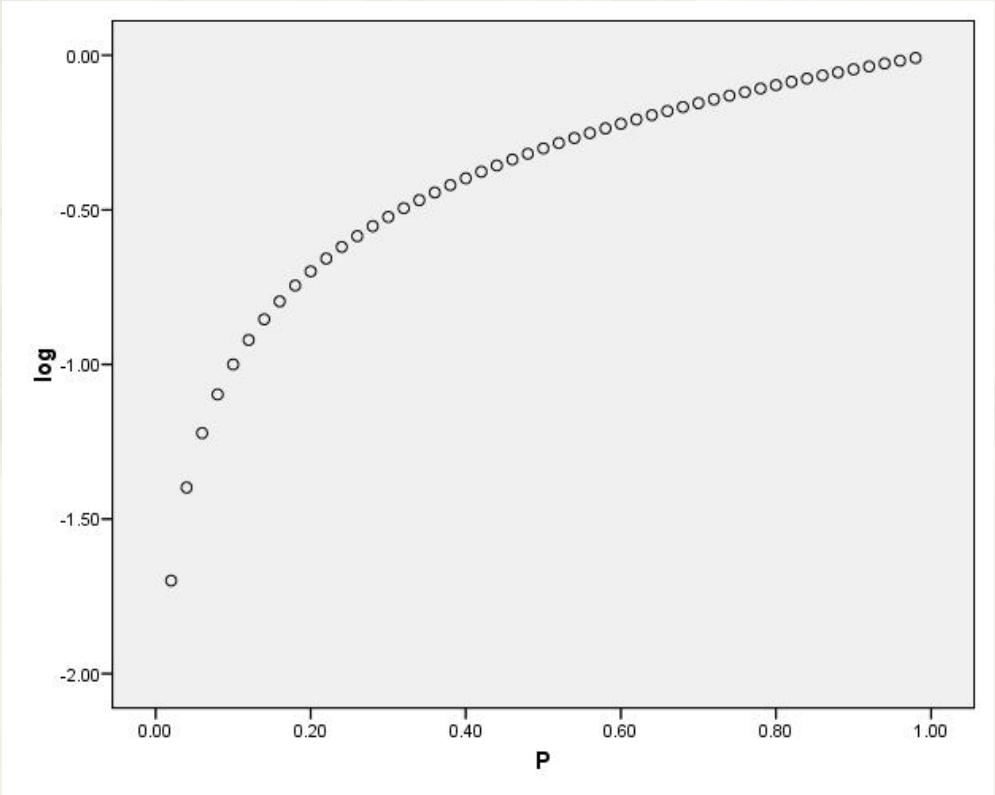
θ	y					
	0	1	2	3	4	5
0.2	0.328	0.409	0.205	0.051	0.007	0.000
0.6	0.010	0.077	0.230	0.346	0.259	0.078

  - 以似然函数值达到最大时的参数值作为总体参数的估计值，似然函数值在0至1间，反映了在所估计参数的总体中抽到特定样本的可能性行，越接近1越好

二项Logistic回归的检验

- 采用极大似然估计法进行参数估计：似然函数值
  - 求似然函数值的对数，得到对数似然函数值

对数似然函数值越大(越接近于0)，意味着模型较好地拟合样本数据的可能性越大，所得模型的拟和优度高；相反，对数似然函数值越小，意味着模型较好地拟合样本数据的可能性越小，所得模型的拟和优度低。



二项Logistic回归的检验

- 回归方程的显著性检验：自变量全体与Logit P的线性关系是否显著，原假设：回归系数同时为0
  - 采用对数似然比测度拟合程度是否提高
  - 设某自变量未引入回归方程前的对数似然函数值为：L_{x_i}
  - 某自变量引入回归方程后的对数似然函数值为：L
  - 对数似然比为：$\dfrac{L_{x_i}}{L}$
  - 如果对数似然比与1无显著差异，则说明该自变量对Logit P的线性解释无显著贡献；如果对数似然比远远大于1，与1有显著差异，则说明解释变量对Logit P的线性有显著贡献。

二项Logistic回归的检验

- 回归方程的显著性检验：自变量全体与Logit P的线性关系是否显著
  - 由于对数似然比 $\frac{L_{x_i}}{L}$ 的分布未知，但其函数(似然比卡方) $-\log(\frac{L_{x_i}}{L})^2$
  - 近似服从卡方分布

$$-\log(\frac{L_{x_i}}{L})^2 = -2\log(\frac{L_{x_i}}{L}) = -2\log(L_{x_i}) - (-2\log(L))$$

  - SPSS将自动计算似然比卡方的观测值和对应的概率*p*值

第二节 多项Logistic回归

- 分析职业、性别在选择品牌（三种）时的倾向性
  - 利用广义logit模型分析。如果因变量有K个水平，则设定一个对照水平(参照水平)，其他各水平分别与参照水平比较
  - 例如：因应变量有a、b、c三个水平，以a作为参照，则有：

$$LogitP_a = \ln[\frac{P_a}{P_a}] = \ln 1 = 0$$

$$LogitP_b = \ln[\frac{P(y=b \mid x)}{P(y=a \mid x)}] = \beta_0 + \sum_{j=1}^{k} \beta_{1j} x_j$$

$$LogitP_c = \ln[\frac{P(y=c \mid x)}{P(y=a \mid x}] = \beta_0 + \sum_{j=1}^{k} \beta_{2j} x_j \qquad P_a + P_b + P_c = 1$$

多项Logistic回归：应用案例

- 分析职业、性别在选择品牌（三种）时的倾向性

Parameter Estimates

购买品牌[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
A	Intercept	-.656	.296	4.924	1	.026			
	[x1=1.00]	-1.315	.384	11.727	1	.001	.269	.127	.570
	[x1=2.00]	-.232	.333	.486	1	.486	.793	.413	1.522
	[x1=3.00]	0[b]	.	.	0	.	.	.	.
	[x2=1.00]	.747	.282	7.027	1	.008	2.112	1.215	3.670
	[x2=2.00]	0[b]	.	.	0	.	.	.	.
B	Intercept	-.653	.293	4.986	1	.026			
	[x1=1.00]	-.656	.339	3.730	1	.053	.519	.267	1.010
	[x1=2.00]	-.475	.344	1.915	1	.166	.622	.317	1.219
	[x1=3.00]	0[b]	.	.	0	.	.	.	.
	[x2=1.00]	.743	.271	7.533	1	.006	2.101	1.237	3.571
	[x2=2.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: C.

b. This parameter is set to zero because it is redundant.

多项Logistic回归：应用案例

- 分析职业、性别在选择品牌（三种）时的倾向性

$$\log it\frac{p_a}{p_c} = -0.656 - 1.315x1(1) + 0.747x2(1)$$

•当性别相同时，第一种职业的logit(P_a/P_c)比第三种职业(参照水平)平均减少1.315，第一种职业的(P_a/P_c)是第三种职业的0.269倍。如果以P_c为基准，则第一种职业选择A品牌的倾向不如第三种职业，且统计上显著；

•当职业相同时，男性的logit(P_a/P_c)比女(参照水平)平均多0.747，男性的(P_a/P_c)是女性的2.112倍。如果以P_c为基准，则男性较女性更倾向选择A品牌，且统计上显著，即男性选择A品牌的倾向性与女性的差异显著。

多项Logistic回归：应用案例

- 分析职业、性别在选择品牌（三种）时的倾向性

$$\log it\frac{p_b}{p_c} = -0.653 - 0.656x1(1) + 0.743x2(1)$$

•当性别相同时，第一种职业的logit(P_b/P_c)比第三种职业(参照水平)平均减少0.653，第一种职业的(P_b/P_c)是第三种职业的0.519倍。如果以P_c为基准，则第一种职业选择B品牌的倾向不如第三种职业，但统计上不显著；

•当职业相同时，男性的logit(P_b/P_c)比女(参照水平)平均多0.743，男性的(P_b/P_c)是女性的2.101倍。如果以P_c为基准，则男性较女性更倾向选择B品牌，且统计上显著，即男性选择B品牌的倾向性与女性的差异显著。