



多元统计分析

Applied Multivariate Statistical Analysis

授课教师: 赵春晖

联系方式:

Email: chhzhao@zju.edu.cn

Phone: 13588312064

Room: 工控新楼308室



多元统计分析

- ◆ 什么是多元统计分析
- ◆ 多元统计分析的研究对象和方法
- ◆ 多元统计分析的应用领域
- ◆ 多元统计数据的图表示法
- ◆ 多元统计分析的简单指标
- ◆ 统计分析的预处理——标准化



三个关键词

实用：有实际使用价值的

多元：多个变量（或称因素、指标）

统计：指对某一现象有关的数据的搜集、整理、
计算和分析等。



1.1 什么是多元统计分析

- 在实际问题中，很多随机现象涉及到的变量不止一个，而经常是多个变量，而且这些变量间又存在一定的联系。我们常常需要处理多个变量的观测数据。
- 例1：表征过程运行状态的变量：压力、流量、温度、速度、浓度等
- 例2：地区经济发展的指标：总产值、利润、效益、劳动生产率、固定资产、物价、信贷、税收等
- 例3：医学诊断：血糖、血压、脉搏、白血球、体温等



1.1 什么是多元统计分析

- 如何对观测数据进行有效的分析和研究？
- 做法1：把多个随机变量分开分析（忽视了变量之间的相关性，会丢失信息，也不容易取得好的研究结果）。
一元统计分析：研究一个随机变量统计规律的学科
- 做法2：针对多个随机变量同时进行分析研究。采用**多元统计分析方法**，通过对多个随机变量观测数据的分析，来研究变量之间的相互关系以及揭示这些变量内在的变化规律。

1.1 什么是多元统计分析

下表给出从中学某年级随机抽取的12名学生中5门主要课程期末考试成绩

举例

序号	政治	语文	外语	数学	物理
1	99	94	93	100	100
2	99	88	96	99	97
3	100	98	81	96	100
4	93	88	88	99	96
5	100	91	72	96	78
6	90	78	82	75	97
7	75	73	88	97	89
8	93	84	83	68	88
9	87	73	60	76	84
10	95	82	90	62	39
11	76	72	43	67	78
12	85	75	50	34	37



1.1 什么是多元统计分析

上表提供的数据，如果用一元统计方法，势必要把多门课程分开分析，每次分析处理一门课的成绩。这样处理，由于忽视了课程之间可能存在的相关性，因此，一般说来，丢失信息太多。分析的结果不能客观全面地反映某年级学生的学习情况。

这里要讨论的多元分析方法，它同时对多门课程成绩进行分析。这样的分析对这些课程之间的相互关系、相互依赖性等都能提供有用的信息。



1.1 什么是多元统计分析

- 多元统计分析
 - 研究**多个随机变量**之间相互依赖关系以及内在统计规律性的理论和统计方法的总称。
 - 利用多元分析还可以对研究对象进行分类和简化。
- 多元统计分析**研究的对象就是多维随机向量**。
- **研究的内容**既包括一元统计学中某些方法的直接推广，也包括多个随机变量特有的一些问题。
- 多元统计分析是一类范围很广的理论和方法。



1.1 什么是多元统计分析

多元统计分析(简称多元分析)是统计学的一个重要分支。它是应用**数理统计学**来研究**多变量(多指标)**问题的理论和方法；它是一元统计学的推广和发展。



1.2 多元统计分析的对象和内容

综上所述，多元分析以 p 个变量的 n 次观测数据组成的数据矩阵

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

为依据。根据实际问题的需要，给出种种方法。

英国著名统计学家M. 肯德尔 (M. G. Kendall) 在《多元分析》一书中把多元分析所研究的内容和方法概括为以下几个方面：



1.2 多元统计分析的对象和内容

1、简化数据结构(降维问题)

- 通过变量变换等方法使相互依赖的变量变成互不相关的;
- 或把高维空间的数据投影到低维空间, 使问题得到简化而损失的信息又不太多.
- 主成分分析, 因子分析, 对应分析等多元统计方法就是这样的一类方法。

2、分类与判别 (归类问题)

- 对所考查的对象(样品点或变量) 按相似程度进行分类 (或归类) 。
- 聚类分析和判别分析等方法是解决这类问题的统计方法。



1.2 多元统计分析的对象和内容

3. 变量间的相互联系

相互依赖关系:分析一个或几个变量的变化是否依赖于另一些变量的变化?如果是,建立变量间的定量关系式,并用于预测或控制---回归分析.

- ◆ 两组变量间的相互关系---典型相关分析等.
- ◆ 一组变量依赖另一组变量的变化关系---偏最小二乘回归分析.



1.2 多元统计分析的对象和内容

4、多元数据的统计推断

参数估计和假设检验问题，特别是多元正态分布的均值向量和协差阵的估计和假设检验等问题。

5、多元统计分析的理论基础

- 包括多维随机向量及多维正态随机向量，及由此定义的各种多元统计量，推导它们的分布并研究其性质，研究它们的抽样分布理论。
- 这些不仅是统计估计和假设检验的基础，也是多元统计分析的理论基础



2 多元统计分析的实际应用

制造过程:

- 了解过程目前的运行状态，并预测可能出现的情况；
- 了解过程的哪些方面可能出现了不正常情况；
- 根据所了解的过程运行状况，进而改进过程及产品质量
- 考察某产品的质量指标与影响产品质量的因素（多个）之间的关系（多重多元回归分析法）
- 某一产品用两种不同的原料生产，产品的质量有无显著差异？



2 多元统计分析的实际应用

医学应用

■ 医生对病人的诊断是靠对病人观测若干症状来综合评定。如一个人发高烧，医生根据他的体温高低、白血球数目及其它症状来判断他是得感冒、肺炎还是其它。再如某人发现腹部有肿瘤，医生根据肿瘤的大小、生长的速度、边界是否清楚，质硬或软等症状来判断肿瘤是良性或恶性---**判别问题**



2 多元统计分析的实际应用

教育学

如何对高考的考生成绩作因素分析？学生入学后的考试成绩和入学考试的各门课程成绩有何相关关系？

体育科学

如何研究体力测试指标（反复横向跳、立定体前屈、俯卧上体后仰等）与运动能力测试指标（耐力跑、跳远、投球等）之间的相关关系？

生态学

对1000个类似的鱼类样本，如何根据测量的特征如体重、身长、鳍数、鳍长、头宽等，将这些鱼分成几个不同品种？



2 多元统计分析的实际应用

其他

军事科学

生物学

火警预报

林业科学

.....

心理学

保险科学

地震预报



3 多元统计数据的图表示法

图形有助于对所研究的数据的直观了解,一维或二维数据的图形容易得到,三维图形虽也可以画出,但并不方便.

三维以上图形如何表示? 目前尚未有公认的方法.

设变量个数为 p , 观测次数为 n , 第 k 次观测值记为 $X_{(k)} = (x_{k1}$
 $x_{k2} \dots x_{kp}) \quad (k=1,2,\dots,n)$

3 多元统计数据的图表示法

--统计过程监测图

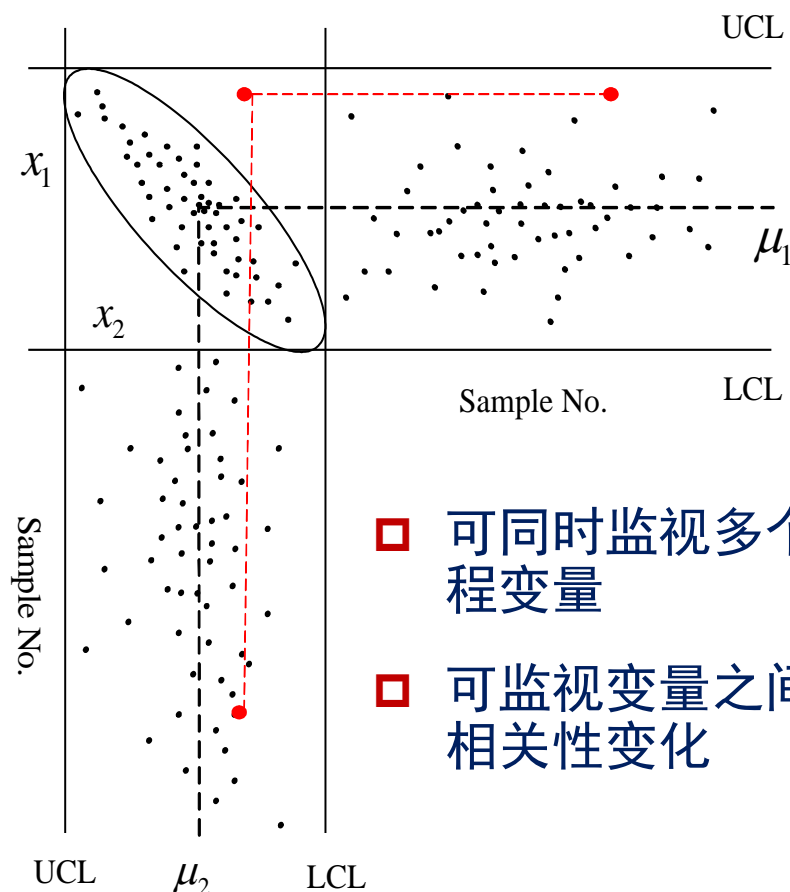
从单变量到多变量

■ 常见的单变量统计控制图

- Shewhart
- CUSUM
- EWMA

■ 从单变量到多变量统计控制图

- PCA
- PLS

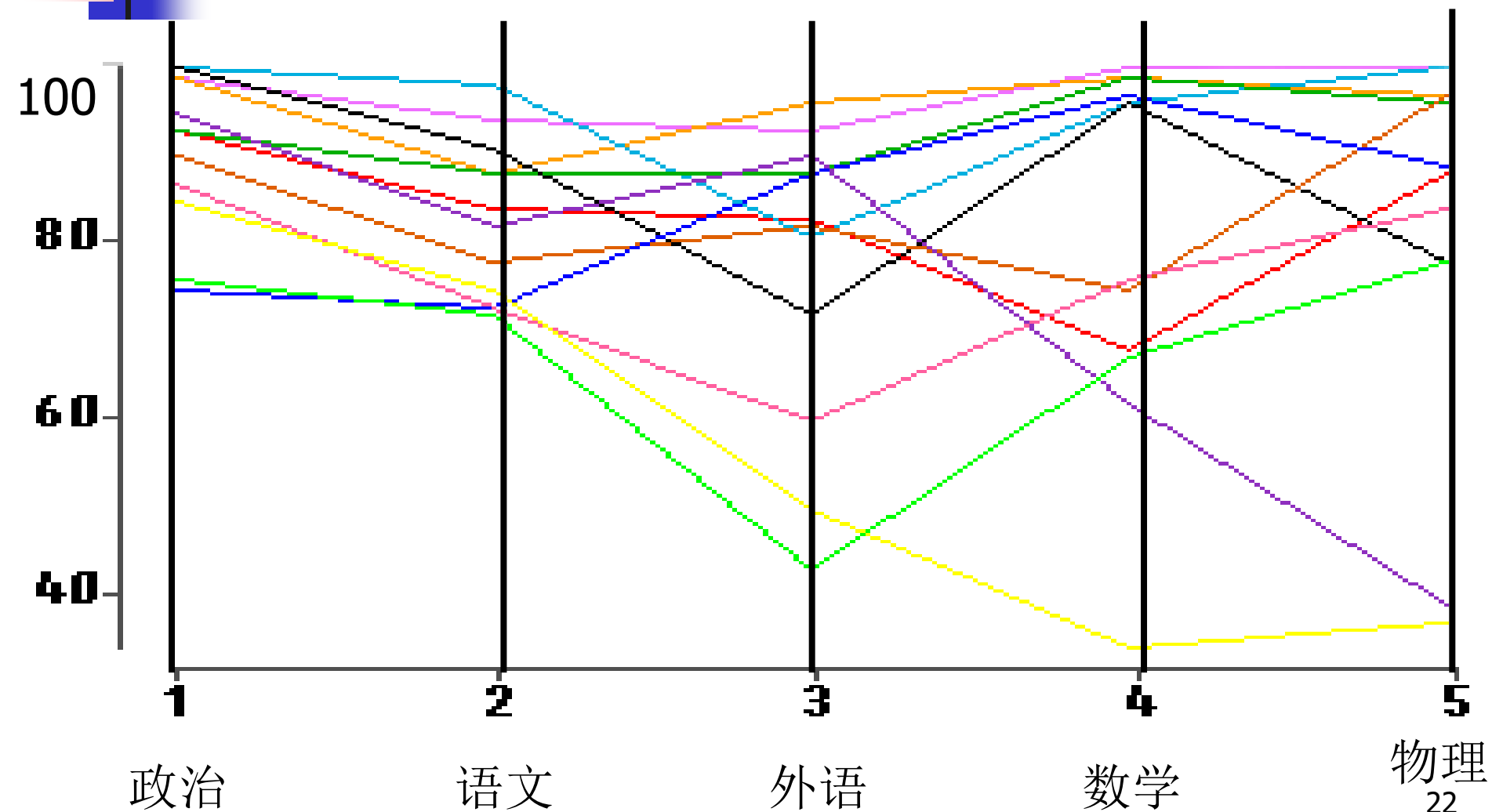


□ 可同时监视多个过程变量

□ 可监视变量之间的相关性变化

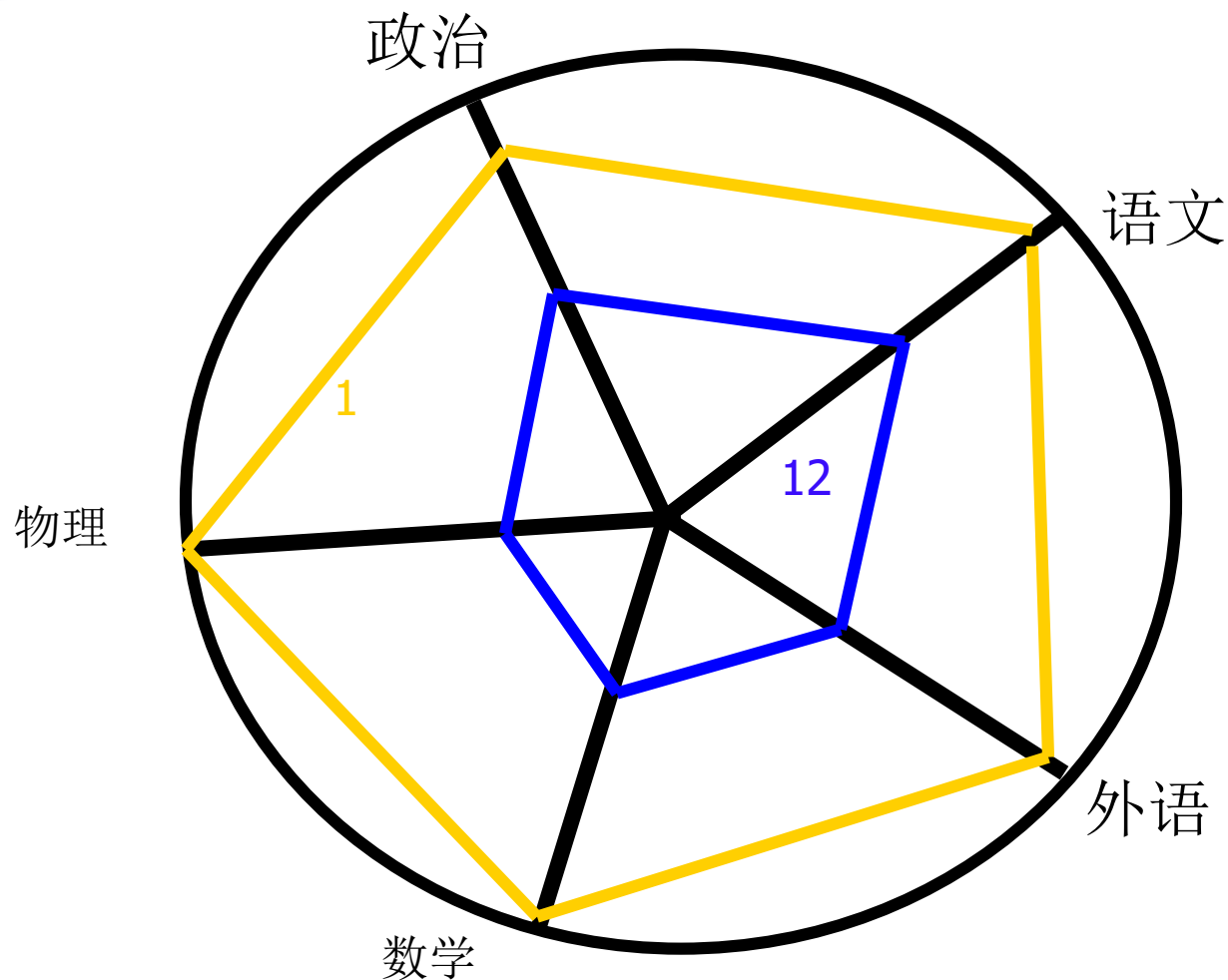
3 多元统计数据的图表示法

---轮廓图



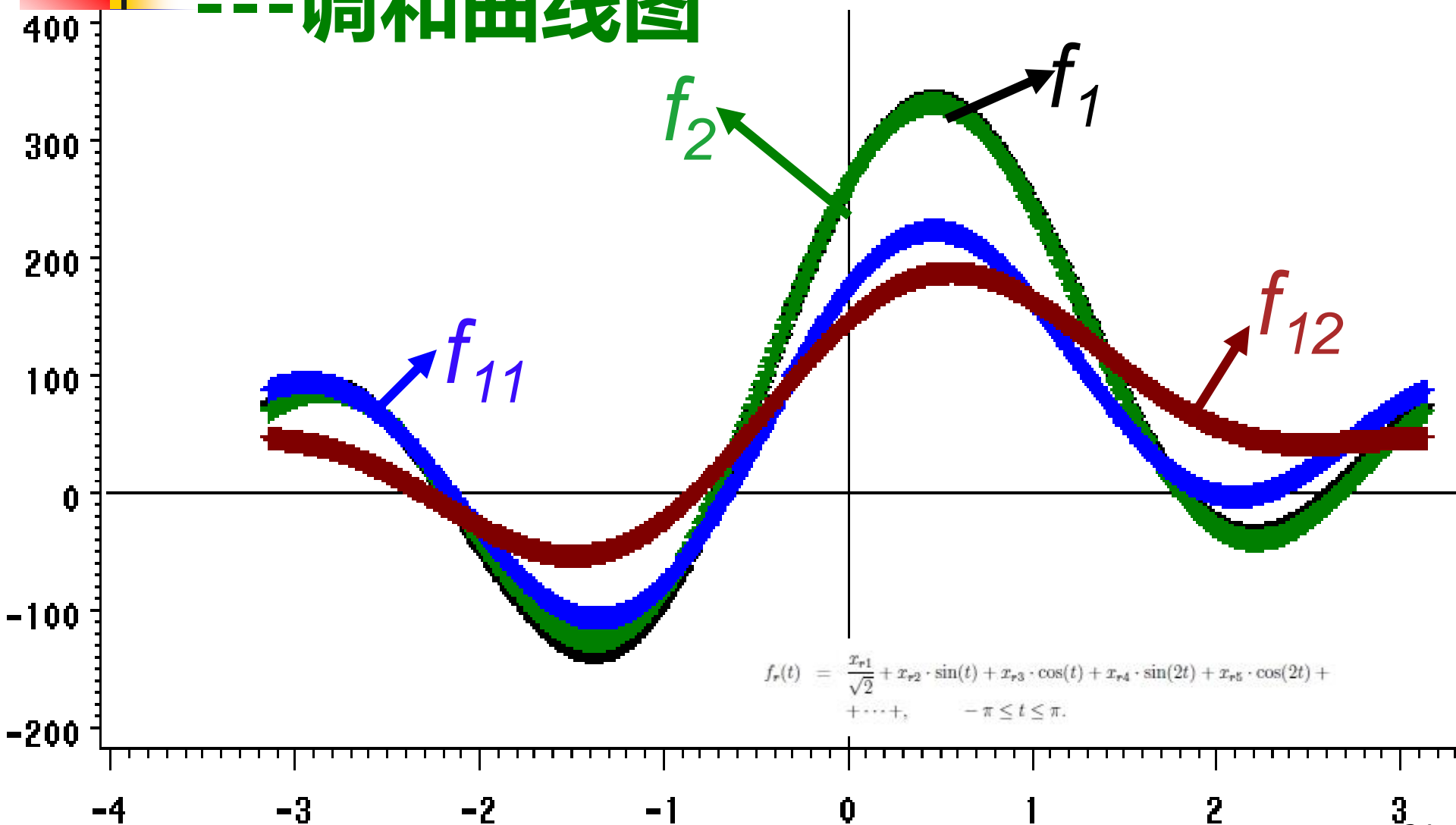
3 多元统计数据的图表示法

---雷达图



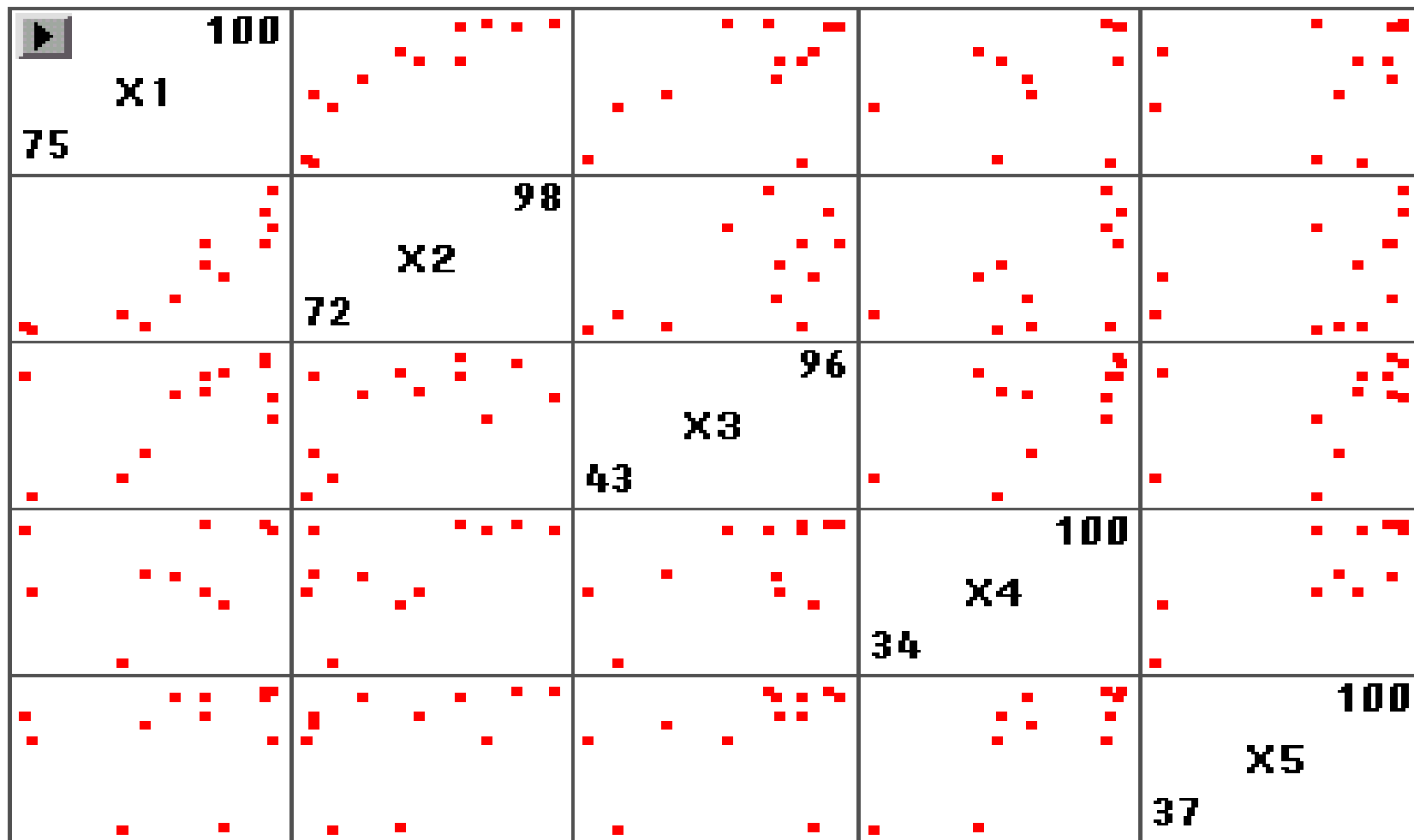
3 多元统计数据的图表示法

---调和曲线图



3 多元统计数据的图表示法

--- 散布图矩阵





3 多元统计数据的图表示法

---其它

在多元数据的图表示法中, 还有星座图、脸谱图、装饰图等表示法.

多元数据的图表示法的难点:

在于变量过多。如果有一种方法可以把高维数据投影到二维空间(平面)中去. 并且在投影过程中不会过多地损失原有数据信息的话, 就可以使用通常方法在平面上画出这些本来是高维数据的图形来. 后面将介绍的主成分分析等方法就是一些降维的方法。



3 多元统计分析的简单指标

- 统计分析是数据分析的主要工具
- 完整的数据分析过程包括
 - 数据的采集（数据可靠性、完备性、相关性，各种数据类型如极大型指标、极小型指标、居中型指标，时变的或静态的等等）
 - 数据的清洗：格式标准化、异常数据清除、错误纠正、缺损值处理、时间校准等
 - 数据的分析（模型的适用性）
- 统计学为数据分析过程提供了一套完整的科学的方法论。

3 多元统计分析的简单指标

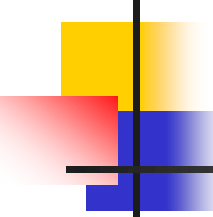
整理好的数据具有如下结构：

指标 (属性)

对象的
观察值
样本

编号	X1	X2	X3	X4	?	?	Xm
1	x_{11}						
2	x_{21}						
3	x_{31}						
?						
n	x_{n1}						

数据是信息载体，需要分析数据的主要特征。一些简单的统计指标可以对研究对象的做一些定量刻画。




基础---平均数

现在某大学有一万名教职工（不同学院），他们的睡眠时间进行一次统计，那么他们的平均睡眠时间是多少？这个值叫做**总体平均数**。

现在假定你是大学的一个基层人员，校长给你一个早上时间，让你估算一下全校教职工的平均睡眠时间，精力有限，你怎么办？

从全校的教职工中随机选取一批职工，了解他们的睡眠时间报上来，这样你就得到了 n 个教职工的睡眠时间，这 n 个教职工就是你的**样本**。

拿这 n 个教职工的睡眠时间相加并除以 n ，你就得到了**样本平均数**。你把这个数报告给校长，这个数就是你估算出来的全校教职工平均睡眠时间。



基础---平均数

□ 样本平均数会不会等于总体平均数？

很显然这和你的“手气”有关——不过大多数情况下是不会相等的。

□ 既然样本平均数不等于总体平均数，要它还有用吗？

有！因为样本平均数是总体平均数的无偏估计——也就是说只要你采用这种方法进行估算，估算的结果的期望值（你可以近似理解为很多次估算结果的平均数）既不会大于真实的平均数，也不会小于之。换句话说：你这种估算方法没有系统上的偏差，而产生误差的原因只有一个：随机因素（也就是你的手气好坏造成的）。



3 多元统计分析的简单指标

- (1) 单变量的均值 (mean)

- 均值作为一组数据的代表，反映该组数据平均水平，计算公式如下：

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

- 性质1:
$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Matlab命令: `mean(x)`



3 多元统计分析的简单指标

❖ (2)方差 (variance)

- 方差用于衡量数据的集中或分散程度，公式为：

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

Matlab命令：var(x)

- 标准差 (standard deviation) 是观测值与均值间的平均距离，公式为：

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

Matlab命令：std(x)

3 多元统计分析的简单指标

(2)方差 (variance)

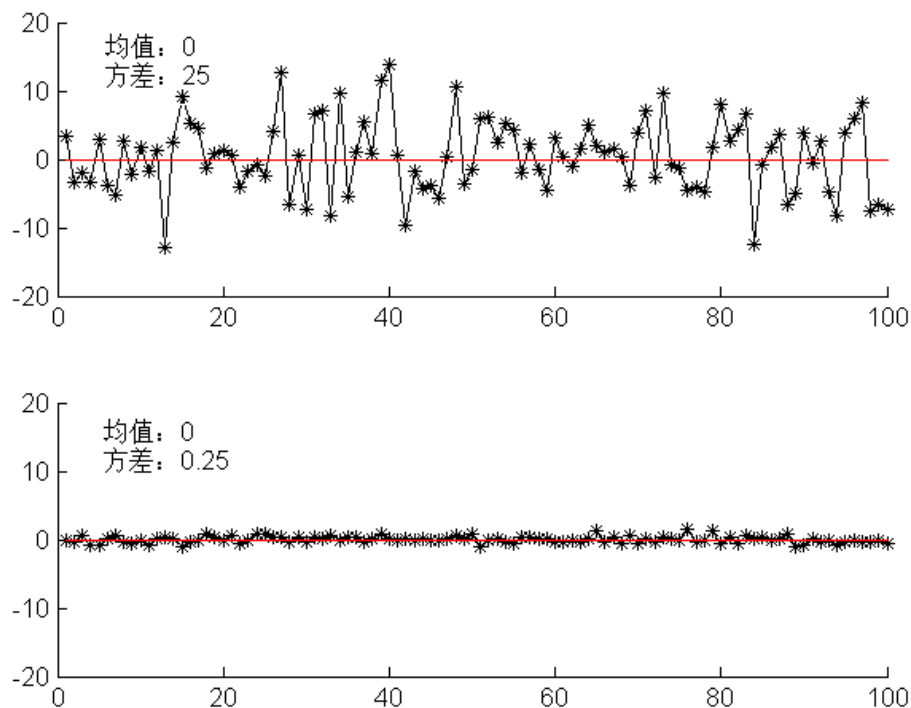


图1.不同方差数据示意图:变异性越大,说明指标对各种场景的遍历性越强,提供的信息越充分,信息量越大。



3 多元统计分析的简单指标

❖ (3)两个变量的协方差 (covariance)

➤ 协方差用于衡量数据的协变趋势，公式为：

$$\text{COV} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

记为 $c(x, y) = \sigma_{xy}$. 若 $x = y$, $c(x, x) = v(x) = \sigma^2$

若X和Y的均值为零，协方差 $c(x, y) = \frac{1}{n} x^T y$

如果X 与Y 是不相关的，二者之间的协方差就是0

matlab命令： `cov(x,y)`



3 多元统计分析的简单指标

❖ (4) 相关系数(correlation coefficient)

- 相关系数是对于变量而言，第 j 个和第 k 个变量之间的相关系数公式为：

$$\text{corr} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}}$$

- 相关系数大小在区间 $[-1, 1]$ 之间，也可写为：

$$\text{corr} = \frac{\text{cov}(x_j, x_k)}{\sqrt{\text{var}(x_j) \text{var}(x_k)}}$$

- (标准化变换不改变相关系数)

4 统计分析的预处理—标准化

- 假定有 n 组样本， m 个变量，其原始数据矩阵 X 为：

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}_{n \times m} = [x_1, x_2, \dots, x_m]$$

- 对矩阵进行标准化，其公式为：

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_j}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m$$

从而使得矩阵的每一列均值为0，方差为1

- 标准化2
$$x_{ij}^* = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)}$$



4 统计分析的预处理—标准化

标准化的优点：消除数据量纲的影响

例如：

杭州市的温度： $-10\sim 45^{\circ}\text{C}$

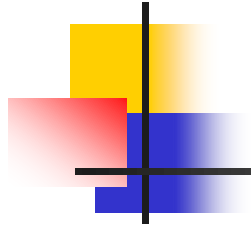
大气压力： 10^5Pa

湿度： $0\%\sim 100\%$

怎么分析温度、大气压力和湿度对心情的影响？



相关系数矩阵、协方差阵？



END