

§ 7 数据分析方法基础

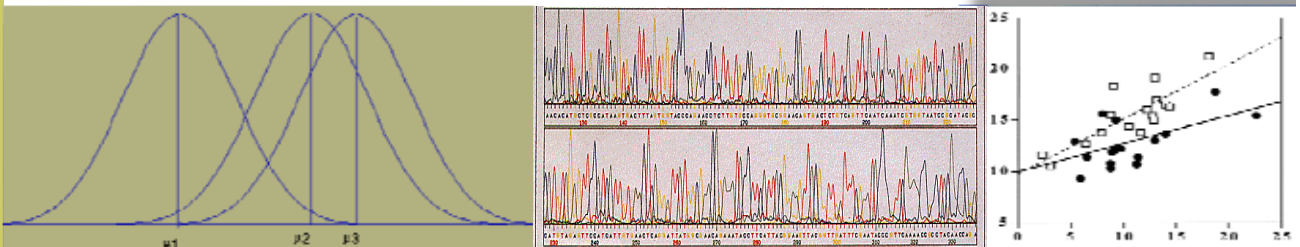
(Part I)

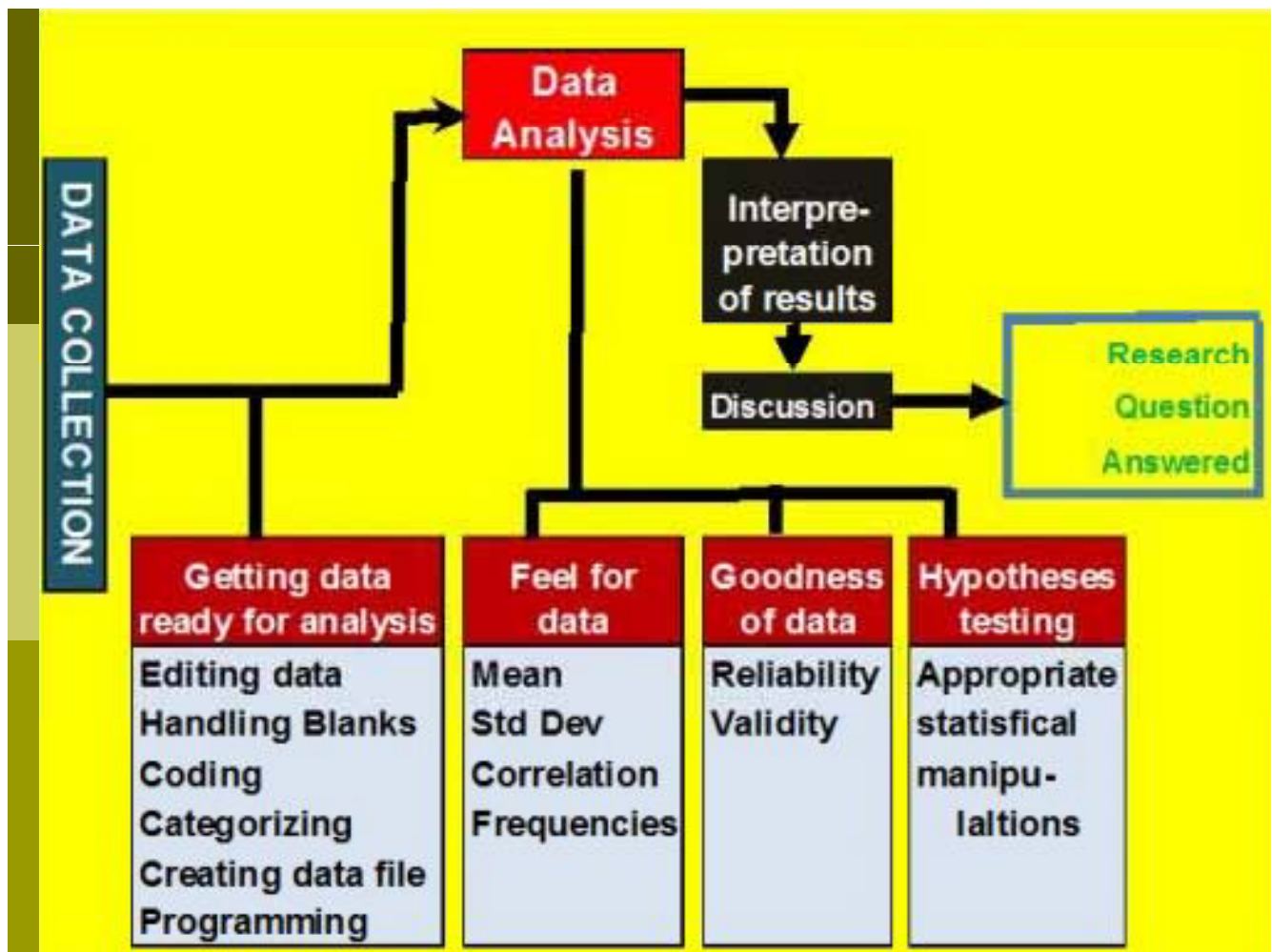
人们一直宣称，世界会由数字来统治。但我所知道的是，数字会告诉我们，它被利用得是好还是坏。

《歌德对话录》

数据，尤其是大量的数据，通常不能提供信息。而统计学家的目的是揭示这些数据所包含的信息。

——John Tabak (《Probability & Statistics—The Science of Uncertainty》in 2004)





复杂数据的多元统计分析

随机现象的多变量、多因素

医学病症诊断
复杂化学反应
工程质量控制

.....

运用数理统计方法研究多变量、多因素问题
→ 多元统计分析理论和方法

多元统计分析 (Multivariate Statistics)

研究多元变量的统计规律性，是一元统计学的推广，同时又有多元随机变量特有的问题。



多元统计分析的主要研究内容和方法

1928年, Wishart 《多元正态总体样本协方差阵的精确分布》

1、降维问题（简化数据结构）

(1) 将某些较复杂的数据结构通过变量变换等方法使相互依赖的变量变成更少的互不相关的变量

(2) 把高维空间的数据投影到低维空间, 使问题得到简化同时损失的信息不太多。

主成分分析

因子分析

对应分析

2、归类问题

对所考察的观测样本（或变量）按照相似程度进行分类、归类

聚类分析

判别分析

3、变量间的相互联系

(1) 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。建立变量间的定量关系，并用于预测或控制

回归分析

(2) 变量间的相互关系：分析两组变量间的相互关系

典型相关性分析、独立性分析

4、多元数据的统计学基础

统计描述方法

参数估计

假设检验

5、多元统计分析的概率理论基础

多维随机向量

多维正态随机向量

多元统计量

§ 7.1 经验模型与回归分析

§ 7.1.1 机械模型和经验模型

机械模型 (Mechanical model) :

实验建立在对现象已有的物理、化学或工程科学的基础之上，由此建立的数学模型。

欧姆定律： $I = U / R$

理想气体定律： $P V = C$

实际观察的数学模型 :

$I = U / R + \varepsilon$

ε : 影响系统的所有不包括在机械模型中的变异性来源的效应，表示观察值不完全遵循机械模型。

经验模型 (Empirical model) :

不分析实际过程的机理，而是根据从实际得到的与过程有关的数据进行数理统计分析、按误差最小原则，归纳出该过程各参数和变量之间的数学关系式，由此建立的数学模型。

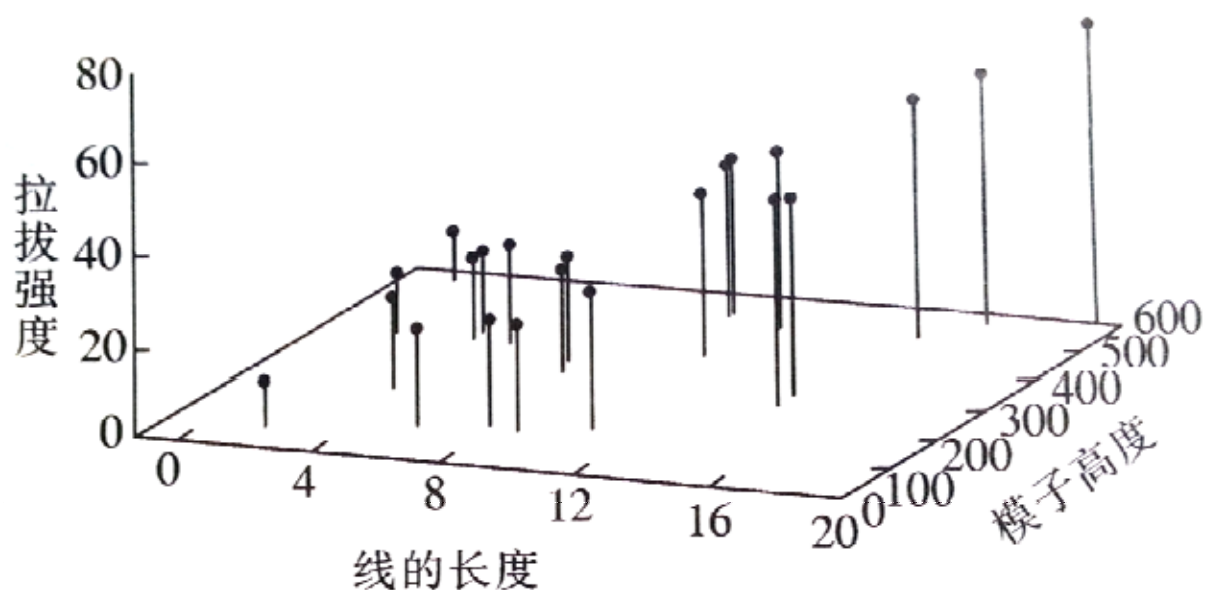
回归分析是建立经验模型的最常用手段。



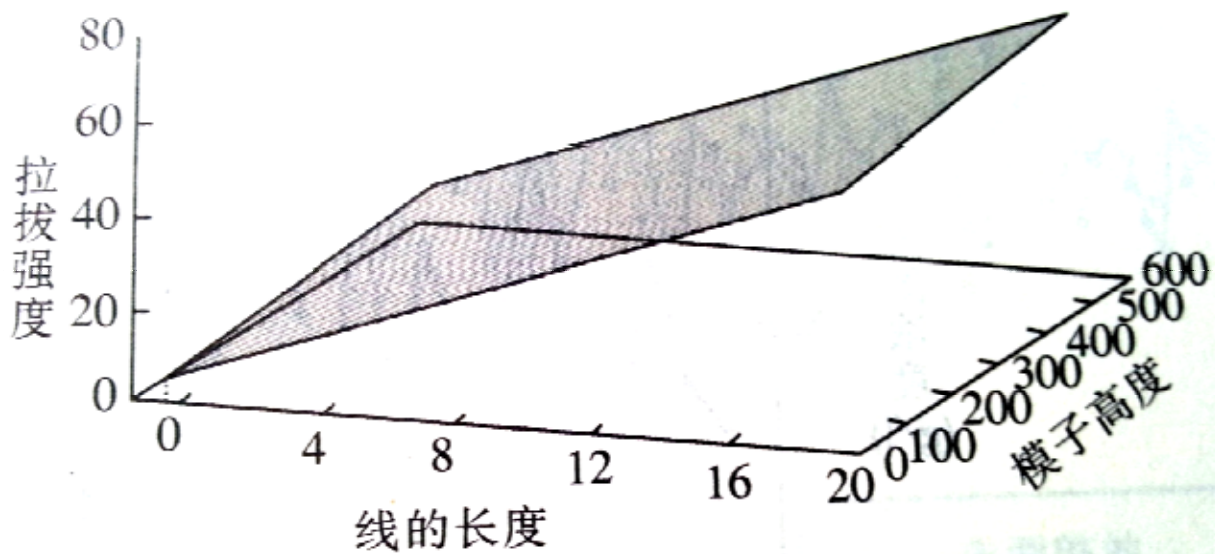
【Example 7.1】 半导体生产车间对生产的结合体部件进行拉拔强度（打破结合体需要的力量的测度）测试（见下表）。由于没有可以简单运用的物理机理，需据此建立拉拔强度与金属丝长度、模子高度的经验模型。

观察号	拉拔强度 y	金属丝长度 x_1	模子高度 x_2	观察号	拉拔强度 y	金属丝长度 x_1	模子高度 x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

拉拔强度数据的三位散点图
(样本观察值)



运用回归分析方法得出拉拔强度估计值图 (经验模型)



$$Y = a + bx_1 + cx_2 + \varepsilon$$

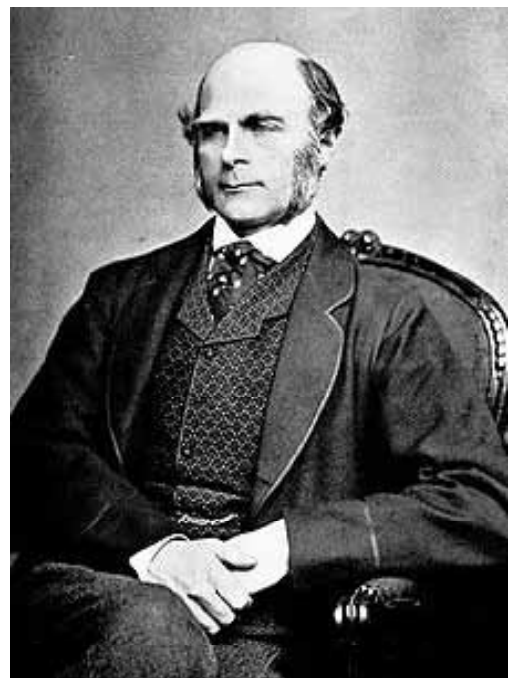
§ 7.1.2 回归分析概述

✓ **Regression:** the relation between selected values of x and observed values of y (from which the most probable value of y can be predicted for any value of x)

补充：Regression来源于F. Galton对亲子间的身高做的研究，发现父母的身高虽然会遗传给子女，但子女的身高却有逐渐“回归到中等即人的平均值”的现象

✓ 寻求表达量 Y 与 $x_1, x_2, \dots, x_n$ 的相关关系的**经验回归方程**，简称**回归方程**

✓ **因果关系**



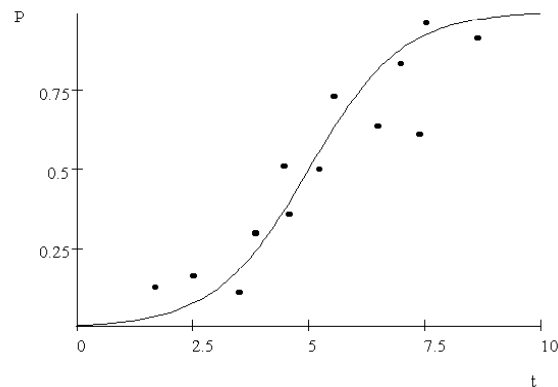
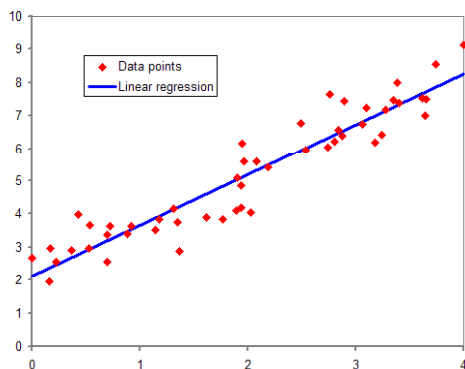
✓ 利用回归方程，在一定可靠度的要求下，预估当自变量 $x_1, x_2, \dots, x_n$ 取确定值时，随机变量 Y 的取值，称为**预测问题**；

✓ 为使 Y 在给定的范围内取值，利用回归方程，控制自变量 $x_1, x_2, \dots, x_n$ 的取值范围，称为**控制问题**。

——一元回归问题、多元回归问题

——多因变量回归问题

——线性回归问题、非线性回归问题



§ 7.1.3 一元线性回归问题

X ：可控制或可精确观测得到的数据的变量；

Y ：与 X 具有相关关系的随机变量。

环境湿度—— x_i ($i=1, 2, \dots, n$)

细菌生长数量—— y_i ($i=1, 2, \dots, n$)

数据对（样本值）： (x_i, y_i) $i=1, 2, \dots, n$

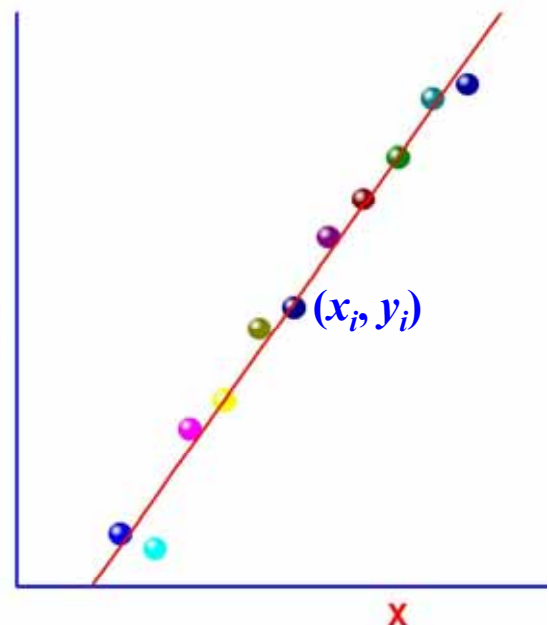
→散点图(Scatter plot)

不妨假定 Y 与 X 具有线性相关关系：

$$Y = a + bX + \varepsilon$$

其中， ε 是期望值为0的随机变量，假设 ε 服从正态分布，于是：

$$E(Y) = a + bX$$



根据数据对（样本观察值） (X_i, Y_i) , $i=1, 2, \dots, n$ 对系数 a 、 b 作出估计，并求得 $E(Y)$ 的估计值：

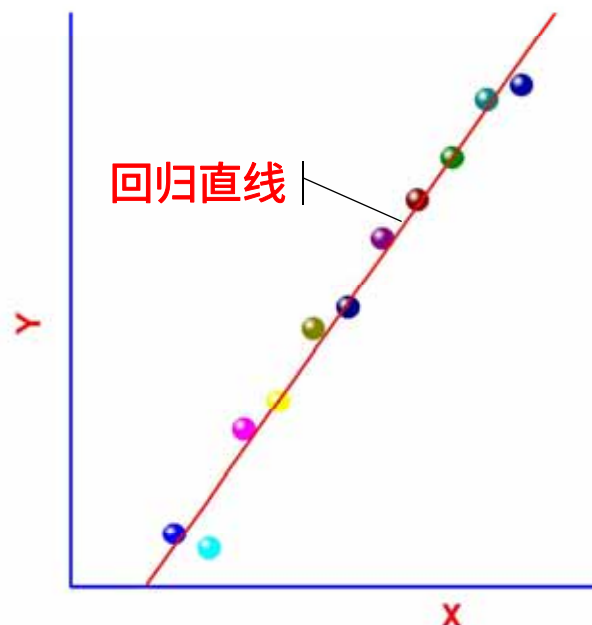
回归值 | $\hat{y} = \hat{a} + \hat{b}x$ | 回归系数

称为一元线性回归方程。

回归直线
回归值
回归系数

求回归方程的两个基本步骤：

1. 求 a 、 b 的估计值，从而求出线性回归方程；
2. 作线性相关性检验。



一. 回归系数的最小二乘法估计

平面直线 $L: y=a+bx$

样本（观察值）点 (X_i, Y_i) , $i=1, 2, \dots, n$

定义离差平方和为：

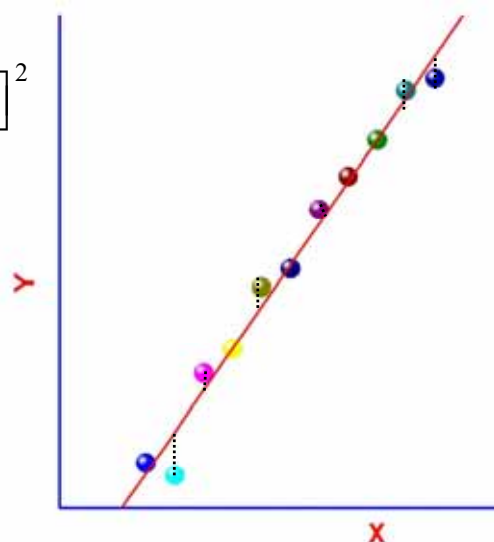
$$Q(a, b) = \sum_{i=1}^n \delta_i^2 = \sum_{i=1}^n [Y_i - (a + bX_i)]^2$$

$Q(a, b)$ 表示点 (x_i, y_i) , $i=1, 2, \dots, n$ 与直线 L 的偏离程度。

满足：

$$Q(\hat{a}, \hat{b}) = \min_{a, b \in \Theta} Q(a, b)$$

的 $\hat{a}, \hat{b}$ 称为 a, b 的**最小二乘估计值**。



根据多元函数达到极值的条件，令：

$$\begin{cases} \frac{\partial Q}{\partial a} = -2 \sum_{i=1}^n [Y_i - (a + bX_i)] = 0 \\ \frac{\partial Q}{\partial b} = -2 \sum_{i=1}^n [Y_i - (a + bX_i)] X_i = 0 \end{cases}$$

化为二元线性方程组：

$$\begin{cases} na + n\bar{X}b = n\bar{Y} \\ n\bar{X}a + \left(\sum_{i=1}^n X_i^2 \right) b = \sum_{i=1}^n X_i Y_i \end{cases}$$

其中：

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

可以证明（略），当 X_i 不全相同时，上述方程组有且存在唯一解。

解得：

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

$$\hat{b} = S_{xy} / S_{xx}$$

其中，

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

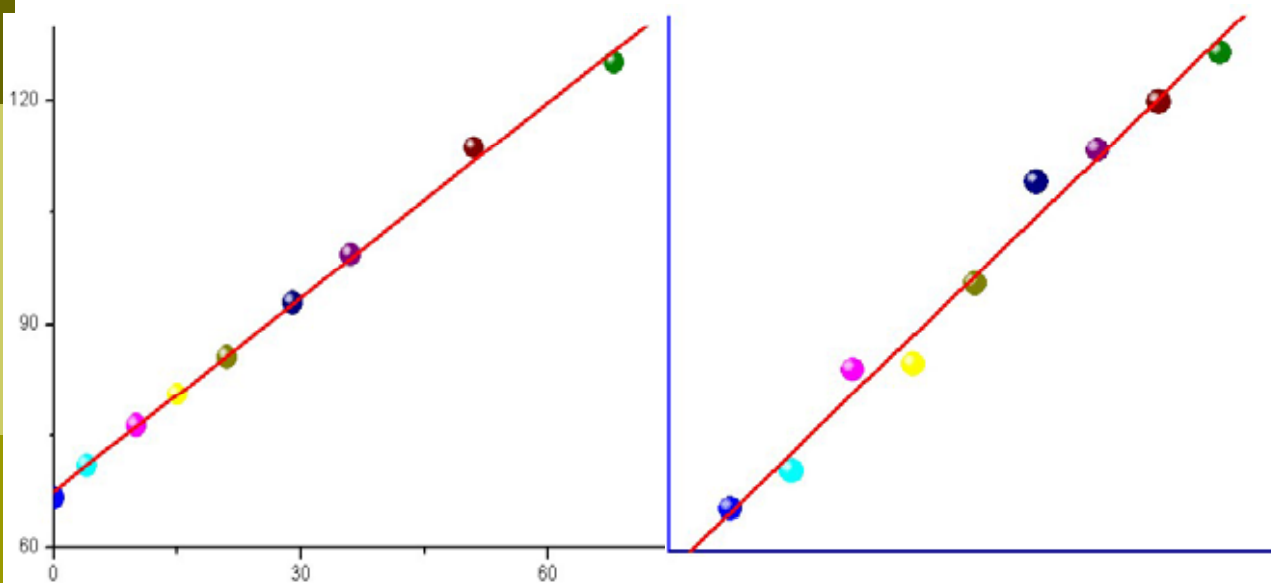
$$S_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$$

讨论：

- 回归系数 $\hat{a}, \hat{b}$ 是由样本观察值决定的随机变量，是由变量的随机涨落导致的；
- 回归系数 $\hat{a}, \hat{b}$ 是 a, b 的最小方差无偏估计（证明略）。

二、线性相关性（显著性）检验



考虑样本总离差平方和（总和）：

$$\begin{aligned} S_{yy} &= \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n [(Y_i - \hat{y}_i) + (\hat{y}_i - \bar{Y})]^2 \\ &= \sum_{i=1}^n (Y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{y}_i)(\hat{y}_i - \bar{Y}) \end{aligned}$$

$$S_{yy} = U + Q$$

$$U = \sum_{i=1}^n (\hat{y}_i - \bar{Y})^2 = (\hat{b})^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$$

$$Q = \sum_{i=1}^n (Y_i - \hat{y}_i)^2$$

$$\hat{y} = \bar{Y} + \hat{b}(x - \bar{X})$$

U：回归值的离差平方和，由n个 x_i 的离散性通过x对Y的相关关系导致。
称为回归和

Q：x对Y的非线性影响以及试验的随机误差（数据的随机涨落）导致。
称为余和

r 检验法

考虑回归和U相对于总和 S_{yy} 的比：

$$\frac{U}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \leq 1$$

定义：

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

称为相关系数。

相关系数r：|r|≤1

|r|越大，线性相关关系越显著；

r=0，Y与x不存在线性相关关系；

|r|=1，Y与x完全线性相关（完全正/负相关）

采用相关系数r为检验统计量，当：

$$|r| > r_{\alpha}(n-2)$$

数据点数目

时，认为在显著性水平 α 下，线性回归显著（接受原假设 H_0 ）。

相关系数临界值 $r_{\alpha}(n-2)$ 表

n-2 \ α	0.10	0.05	0.02	0.01	0.001
1	0.98769	0.99692	0.999507	0.999877	0.9999988
...					
7	0.5822	0.6664	0.7498	0.7977	0.8982
8	0.5494	0.6319	0.7155	0.7646	0.8721
...					
100	0.1638	0.1946	0.2301	0.2540	0.3211

F 检验法

构造检验统计量F：

$$F = \frac{U}{Q / (n-2)}$$

数据点数目

显然，F值越大，U在总和中所占比例越大，回归性也越显著。

当：

$$F > F_{1-\alpha}(1, n-2)$$

时，认为在显著性水平 α 下，线性回归显著（接受原假设 H_0 ）。

查表：F分布表（略）

【Example 7.2】

x_i	0	4	10	15	21	29	36	51	68
y_i	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

根据散点图，确定回归方程形式

$$\hat{y} = \hat{a} + \hat{b}x$$

计算得到：

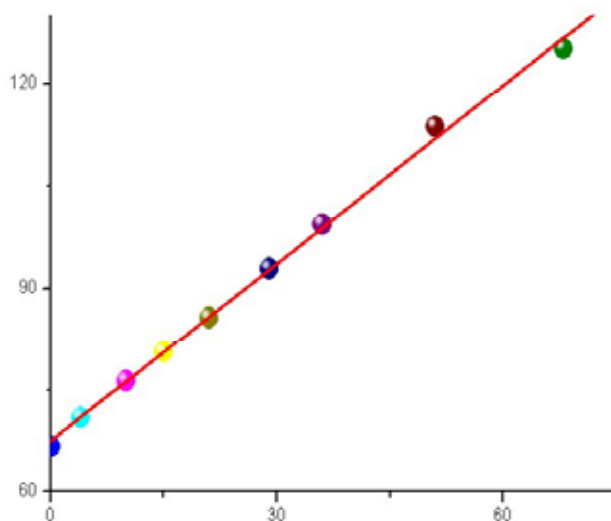
$$\bar{x} = 26.0 \quad \bar{y} = 90.1$$

$$S_{xy} = 3534.8 \quad S_{xx} = 4060$$

$$S_{yy} = 3084$$

$$\hat{b} = 0.87 \quad \hat{a} = 67.51$$

$$\hat{y} = 67.51 + 0.87x$$



线性相关性检验：

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.99896$$

查表得：

$$r_{0.01}(7) = 0.7977$$

$$r_{0.001}(7) = 0.8982$$

可线性化的曲线回归

方法：变量替换

1. 双曲线型

$$y = a + \frac{b}{x}$$

令 $u = \frac{1}{x}$, 得

$$y = a + bu$$

$$\frac{1}{y} = a + \frac{b}{x}$$

令 $u = \frac{1}{x}$, $v = \frac{1}{y}$ 得

$$v = a + bu$$

2. 指数曲线型

$$y = ae^{bx}$$

若 $a > 0$, 则令 $v = \ln y$, 得到:

$$v = \ln a + bx$$

若 $a < 0$, 则令 $v = \ln(-y)$, 得到:

$$v = \ln(-a) + bx$$

3. 幂函数型

$$y = ax^b \quad x > 0$$

若 $a > 0$, 则令 $v = \ln y$, $u = \ln x$, 得到 ($a < 0$ 情况类推):

$$v = \ln a + bu$$

4. 对数曲线型

$$y = a + b \log x$$

令 $u = \log x$ ，得到：

$$y = a + bu$$

$$\log y = a + bx$$

令 $v = \log y$ ，得到：

$$v = a + bx$$

$$\log y = a + b \log x$$

令 $u = \log x$ ， $v = \log y$ ，得到：

$$v = a + bu$$

5. S曲线型

$$y = \frac{1}{a + be^{-x}}$$

令：

$$u = e^{-x} \quad v = 1/y$$

得到：

$$v = a + bu$$

§ 7.1.4 多元线性回归问题

$x_1, x_2, \dots, x_r$: r 个可控制或可精确观测得到的数据的变量 ;
 Y : 与 $x_1, x_2, \dots, x_r$ 具有相关关系的随机变量。

不妨假定 Y 与 $x_1, x_2, \dots, x_r$ 具有线性相关关系 :

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_r x_r + \varepsilon$$

其中, ε 是期望值为0的多维随机误差, 且满足多维正态分布。
对于 n 组样本观察值 ($n > r$) :

$x_{i1}, x_{i2}, \dots, x_{ir}$ ($i=1, 2, \dots, n$)

y_i ($i=1, 2, \dots, n$)

多元线性回归的数学模型

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_r x_{ir} + \varepsilon_i$$

$$E(\varepsilon_i) = 0 \quad i = 1, 2, \dots, n$$

其中, ε_i 互不相关。

记

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_n \end{bmatrix}$$
$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1r} \\ 1 & x_{21} & \dots & x_{2r} \\ \dots & \dots & & \dots \\ 1 & x_{n1} & \dots & x_{nr} \end{bmatrix}$$

多元线性回归模型可写成：

$$\begin{cases} \mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \\ E(\boldsymbol{\varepsilon}) = \mathbf{0} \end{cases}$$

1. 回归系数b的最小二乘估计

定义离差平方和：

$$Q(\mathbf{b}) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_r x_{ir})]^2$$

定义：在多元线性回归模型中，若存在b的估计值 $\hat{\mathbf{b}}$ ，则对于任意一组实数 $b_0, b_1, \dots, b_r$ 构成的向量，都成立不等式

$$Q(\hat{\mathbf{b}}) \leq Q(\mathbf{b})$$

称 $\hat{\mathbf{b}}$ 是b的最小二乘估计。

定理：在多元线性回归模型中，设矩阵X列线性无关，则唯一存在b的最小二乘估计 $\hat{\mathbf{b}}$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

2. 计算回归系数

令 $\frac{\partial Q}{\partial b_j} = 0$, $j = 0, 1, 2, \dots, r$, 得到方程组：

$$b_0 n + b_1 \sum_{i=1}^n x_{i1} + \dots + b_r \sum_{i=1}^n x_{ir} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{ij} + b_1 \sum_{i=1}^n x_{ij} x_{i1} + \dots + b_r \sum_{i=1}^n x_{ij} x_{ir} = \sum_{i=1}^n x_{ij} y_i$$

$$j = 1, 2, \dots, r$$

问题：求解上述方程组。

$$\hat{b}_0 = \bar{y} - \sum_{j=1}^r \bar{x}_j \hat{b}_j$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \dots \\ \hat{b}_r \end{bmatrix} = \begin{bmatrix} l_{11} & l_{12} & \dots & l_{1r} \\ l_{21} & l_{22} & \dots & l_{2r} \\ \dots & \dots & \dots & \dots \\ l_{r1} & l_{r2} & \dots & l_{rr} \end{bmatrix}^{-1} \begin{bmatrix} l_{1y} \\ l_{2y} \\ \dots \\ l_{ry} \end{bmatrix}$$

其中：

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad j = 1, 2, \dots, r$$

$$l_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$l_{iy} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(y_k - \bar{y})$$

3. 显著性检验

S_{yy} ：样本离差平方和

U ：回归和

Q ：余和

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = U + Q$$

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

$$Q = S_{yy} - U$$

$$r = \sqrt{\frac{U}{S_{yy}}}$$

$$F = \frac{U / r}{Q / (n - r - 1)}$$

逐步回归问题

理想的多元回归效果

- 选取对 Y 有显著关联的自变量 $x_1, x_2, \dots, x_k$ 进行回归，剔除关联较小的自变量；
- 对于相互关联很强的自变量 $x_i, x_j, \dots, x_k$ ，只要从中选取一个对 Y 有显著关联的自变量进行回归；

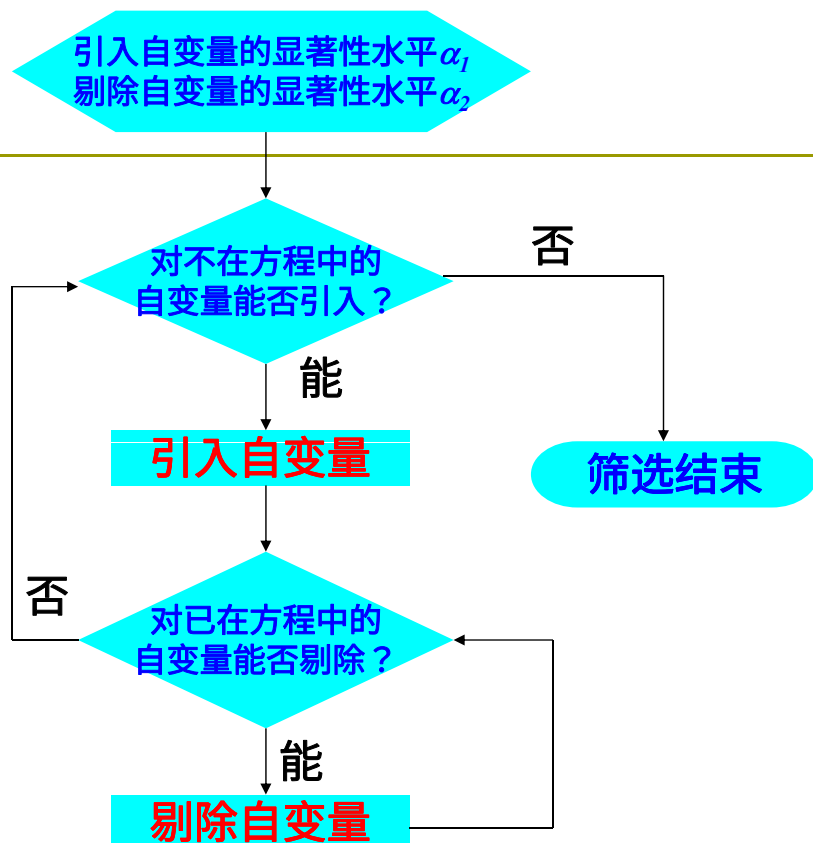


逐步回归

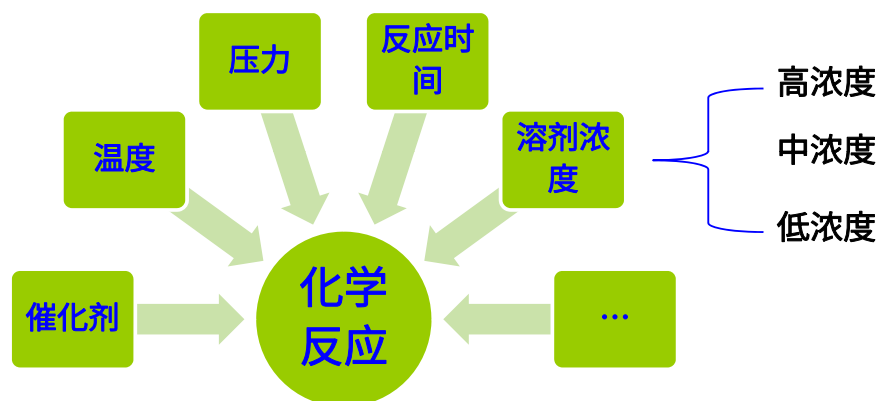
基于逐步筛选法的逐步回归

- 逐个引入自变量，每次引入对Y影响最显著的自变量，并对原有变量逐个进行检验（如F检验），把变为影响不显著的自变量逐个剔除。最终得到的回归方程既不漏掉对Y影响显著的自变量，又不包含对Y影响不显著的变量。

逐步回归的基本步骤



§ 7.2 方差分析 (ANOVA) 简介



方差分析 (ANalysis Of VAriance)

根据试验结果进行分析，鉴别各因素对试验结果的影响程度

§ 7.2.1 ANOVA的基本原理

【Example 7.3】某公司为了研究3种广告内容对某一医疗器械的销售量的影响，进行调查统计。经广告以不同的内容广泛宣传后，按寄回的广告回执的订购数量计算全年4季度的销售情况。试判断广告内容对该器械的销售量是否有影响？若有，哪种广告内容更好？

广告内容	销售量（单位：台）			
	1季度	2季度	3季度	4季度
A ₁ (强调疗效)	163	176	170	185
A ₂ (强调运行成本)	184	198	179	190
A ₃ (强调安装条件)	206	191	218	224

因素/因子 (factor)：试验中需要考察的、可以控制的条件
广告设计、售后服务、外型设计等对销售量的影响 (A, B, C, ...)

水平 (level)：因素所处的不同状态
不同内容的广告设计 (A1, A2, A3)

方差分析 (analysis of variance)：
根据试验结果（包括随机抽样）进行分析，鉴别不同的因素，
或者相同因素的不同水平对试验结果的影响程度。

单因素方差分析

双因素方差分析

多因素方差分析

单因素方差分析（一元方差分析）

考察一个因素所处的不同状态（水平）对试验结果的影响

——多个水平：多组数据

——每一组数据：来自同一总体的抽样样本（独立同分布），
多组数据对应多个总体

——任务：检验在一定假设条件下，各组样本的均值是否相等，
由此判断各组（不同水平）总体对试验结果的影响是否显著

——一般性问题：**具有相同方差的正态总体，其均值是否相等**

§ 7.2.2 单因素（一元）方差分析

一、数学模型

考虑单因素A对试验结果的影响。

取因素A的r个水平 $A_1, A_2, \dots, A_r$ ，在水平 A_i 下重复进行 n_i 次试验（抽样），得到 n_i 个数据：

$$X_{i1}, X_{i2}, \dots, X_{in_i} \quad (i = 1, 2, \dots, r)$$

X_i ：水平 A_i 下的总体， $i=1, 2, \dots, r$ ，且有 $X_i \sim N(\mu_i, \sigma^2)$ （ X_i 为r个相互独立、方差相同的总体），从中抽取样本

$$X_{i1}, X_{i2}, \dots, X_{in_i} \quad (i = 1, 2, \dots, r)$$

数据结构表

水平	试验结果（样本值）				样本均值	样本方差
1	X_{11}	X_{12}	...	X_{1n_1}	$\overline{X}_1$	s_1^2
2	X_{21}	X_{22}	...	X_{2n_2}	$\overline{X}_2$	s_2^2
...	...	...	...	...	...	...
r	X_{r1}	X_{r2}	...	X_{rn_r}	$\overline{X}_r$	s_r^2

广告内容	销售量				样本均值	样本方差
	1季度	2季度	3季度	4季度		
A_1 (强调疗效)	163	176	170	185	173.50	87.08
A_2 (强调运行成本)	184	198	179	190	187.75	66.92
A_3 (强调安装条件)	206	191	218	224	209.75	212.25

基本假定：

(1) r 个总体 $X_1, X_2, \dots, X_r$ 相互独立且有 $X_i \sim N(\mu_i, \sigma^2)$ ，其中 μ_i, σ^2 未知

(2) 各总体 X_i 下的容量为 n_i 的样本 $X_{ij}, j=1, 2, \dots, n_i$ 独立同分布，且有

$$X_{ij} \sim N(\mu_i, \sigma^2), \quad i=1, 2, \dots, r, \quad j=1, 2, \dots, n_i$$

$$e_{ij} = X_{ij} - \mu_i, \quad i=1, 2, \dots, r; \quad j=1, 2, \dots, n_i$$

e_{ij} ：随机误差，相互独立且服从正态分布 $N(0, \sigma^2)$

$$n = \sum_{i=1}^r n_i \quad (\text{样本数据总数})$$

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^r n_i \mu_i$$

$$\alpha_i = \mu_i - \bar{\mu} \quad \sum_{i=1}^r n_i \alpha_i = 0$$

α_i ：第 i 个水平 A_i 对试验结果的作用大小，称为**效应值**。

单因素方差分析的数学模型（线性模型）：

$$\begin{cases} X_{ij} = \bar{\mu} + \alpha_i + e_{ij} \\ e_{ij} \sim N(0, \sigma^2), \quad i = 1, 2, \dots, r; j = 1, 2, \dots, n_i \\ \sum_{i=1}^r n_i \alpha_i = 0 \end{cases}$$

且 e_{ij} 相互独立

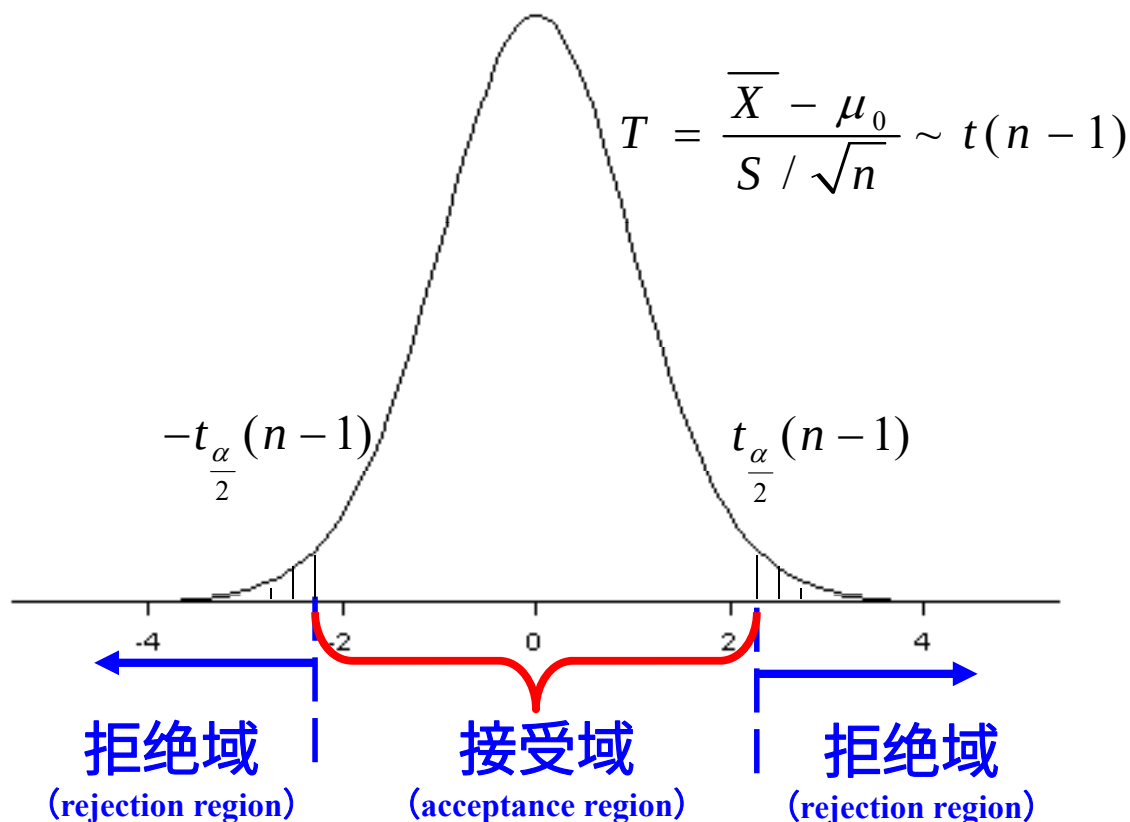
二、方差分析

不同水平对试验结果影响的差异问题归结为如下假设检验问题：

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_r \quad \text{or} \quad \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$$

$$H_1 : \exists \mu_i \neq \mu_j, \quad i \neq j, \quad i, j = 1, 2, \dots, r$$

当 $r=2$ 时，构造 t 检验统计量，进行 t 检验（参见第6讲）



当 $r>2$ 时，Fisher方法 $\Rightarrow$ 方差分析法

由英国统计学家R. A. Fisher提出，为纪念Fisher，以 F 命名，故方差分析又称 F 检验 (F test)。

《The Correlation Between Relatives on the Supposition of Mendelian Inheritance》 (1918)

《Statistical Methods for Research Workers》 (1925)

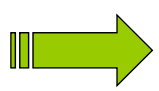


第*i*组样本的样本均值：

$$\overline{X}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad i = 1, 2, \dots, r$$

全体样本的样本均值：

$$\overline{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \overline{X}_{i\bullet}, \quad i = 1, 2, \dots, r, \quad n = \sum_{i=1}^r n_i$$


$$\begin{cases} E(\overline{X}_{i\bullet}) = \mu_i \\ E(\overline{X}) = \frac{1}{n} \sum_{i=1}^r n_i \mu_i = \bar{\mu} \end{cases}$$

$$S_T^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \overline{X})^2 \quad \text{——总离差平方和}$$

$$S_A^2 = \sum_{i=1}^r n_i (\overline{X}_{i\bullet} - \overline{X})^2 \quad \text{——组间离差平方和（即系统误差）}$$

$$S_E^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\bullet})^2 \quad \text{——组内离差平方和（即随机误差）}$$

定理（离差平方和分解公式） 上述离差平方和存在下列关系表达式：

$$S_T^2 = S_A^2 + S_E^2$$

$$S_T^2 = S_A^2 + S_E^2$$

基本思想：

通过组间离差平方和 S_A^2 的大小来反映原假设 H_0 是否成立。若

$$S_A^2 \gg S_E^2$$

说明各总体（组）间的差异显著，则 H_0 可以被拒绝。

由此构造检验统计量的形式为：

$$F = F\left(\frac{S_A^2}{S_E^2}\right)$$

定理 当原假设 H_0 成立时，有：

$$\frac{S_T^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{S_{E_i}^2}{\sigma^2} \sim \chi^2(n_i-1), \text{ and } \frac{S_E^2}{\sigma^2} \sim \chi^2(n-r), \quad \text{where } S_E^2 = \sum_{i=1}^r S_{E_i}^2$$

$$\frac{S_A^2}{\sigma^2} \sim \chi^2(r-1),$$

方差分析的检验统计量F及其分布：

$$F = \frac{S_A^2/(r-1)}{S_E^2/(n-r)} = \frac{\overline{S}_A^2}{\overline{S}_E^2} \sim F(r-1, n-r)$$

在给定显著性水平 α 的情况下， H_0 的拒绝域形式为：

$$F = \frac{S_A^2/(r-1)}{S_E^2/(n-r)} > F_{1-\alpha}(r-1, n-r)$$

单因素（一元）方差分析表

方差来源	DF (自由度)	S ² (离差平方和)	F值
因素A	r-1	S _A ²	$\frac{S_A^2/(r-1)}{S_E^2/(n-r)}$
随机误差	n-r	S _E ²	
总和	n-1	S _T ²	

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/835040012343012032>