假设检验目的： 根据样本去推断是否拒绝原假设H_0

1.检验统计量确定：与枢轴变量形式一致

2. 拒绝域的确定：确定H_0的拒绝域时应遵循有利准则

对H_1成立有利的区域作为拒绝域.

3.两类错误原因：样本随机性和推导的原理(小概率事件实际不发生)

4. 两类错误:　　第一类：弃真；　　第二类：纳伪

不可能使两类错误同时都尽可能小！
减小一类错误，必然使另一类错误增大.

先控制犯第一类错误的概率α，然后再使犯第二类错误的概率尽可能地小$\beta(\mu)$.

例：$X\sim N(\mu,\sigma_0^2)$,犯第一类错误概率：$P_{\mu_0}\{|U|>u_{\frac{\alpha}{2}}\}=\alpha$(落在拒绝域)

犯第二类错误概率：$\beta(\mu)=\Phi(u_{\alpha/2}-\dfrac{\mu-\mu_0}{\sigma_0/\sqrt{n}})-\Phi(-u_{\alpha/2}-\dfrac{\mu-\mu_0}{\sigma_0/\sqrt{n}})$(落在接受域)

$\neq 1-\alpha$

解决问题

Q1： H_0和H_1的确定？

H_0 :(常规假设或维持原状);或一般没有充分理由不能轻易否定的命题.

H_1 :对立假设或备择假设. (新事物或新情况)

Q2: 拒绝域的确定？ Q3: 误判的类型及原因？

Q4： 显著性水平α作用？ Q5(补充)：不拒绝=接受？

例(补充思考):1. 司法(现在)：H_0 : 无罪(避免冤家错案)； H_1 : 有罪

司法(以前)：H'_0 : 有罪(有利于法官)； H'_1 : 无罪

2. 不拒绝 ≠ 接受； → 不依赖样本：所有样本成立

→ 一次样本得出结论

注：1) 假设检验只提供拒绝H_0的证据，没提供判断正确 H_0的证据，即 不拒绝 ≠ 接受；

2) 判断是否拒绝:一次观察, 检验统计量值落入拒绝域 → 拒绝

第九章 回 归 分 析

Regression Analysis

§9.1 相关关系与回归分析

§9.2 一元回归分析
(因果关系模型之一)

注:深度学习(也称深度神经网络模型)：本质上是一个数学模型，提高其预测效果的方法主要源于对机器学习或者统计学习的思考，而不是源于对人类神经网络的深入模拟。

其中的线性模型：

涉及线性回归模型，处理回归问题；

logistic模型处理分类问题。

"回归"一词的由来

F.Galton，**英国生物统计学派的奠基人，他的表哥达尔文的巨著《物种起源》问世以后，触动他用统计方法研究智力遗传进化问题，第一次将概率统计原理等数学方法用于 生物科学，明确提出"生物统计学"的名词.**

现在统计学上的"相关"和"回归"的概念也是高尔顿第一次使用的。

高尔顿的学生卡尔·皮尔逊(**Karl Pearson**)测量了**1078**个父亲及其成年儿子的身高。

高尔顿的学生卡尔·皮尔逊(Karl Pearson)测量了1078个父亲及其成年儿子的身高数据, 发现规律:

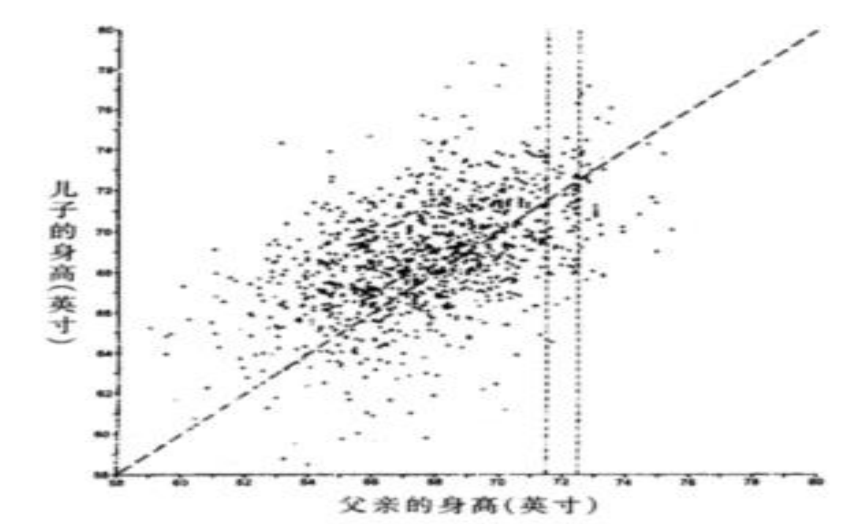


图1 1078对父子身高的散点图

1.高个子的父亲有着较高身材的儿子,而矮个子父亲的儿子身材也比较矮；
2.高个子父母的子女，其身高有低于其父母身高的趋势;
3.而矮个子父母的子女，其身高有高于其父母的趋势;

儿代有向平均身高靠拢的趋势，引出回归分析。即有“回 归”到平均数去的趋势统计学上最初出现“回归”时的涵义.

解释: 大自然有一种约束机制，使人类身高保持某种稳定形态，而不作两级分化. 这就是“一种使身高”回归于中心的作用.

“回归”一词的由来

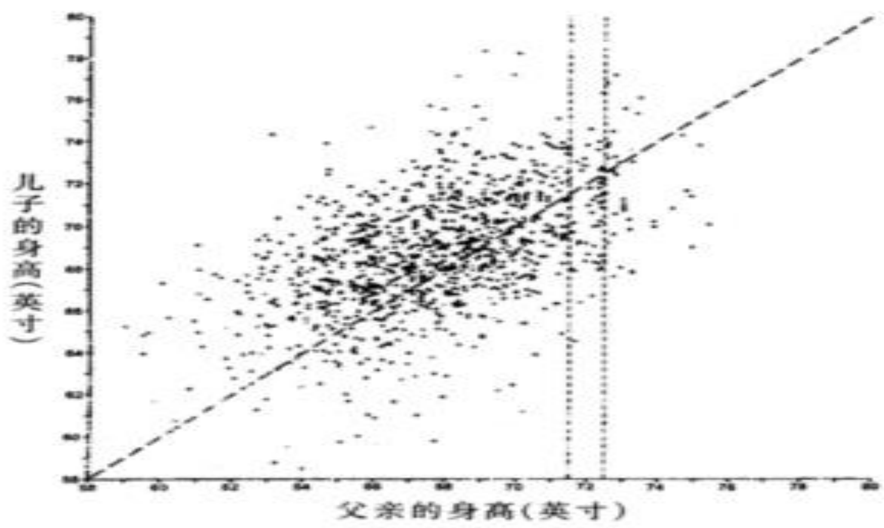


图1 1078对父子身高的散点图

高尔顿根据1078个父亲及其成年儿子的身高值，研究发现二者间的关系为(cm)

成年儿子身高=85.67+0.516×父亲身高±9.51

湖北体育科学研究所得到的公式为：

成年儿子身高=56.699+0.419×父亲身高+0.265×母亲身高 ±3

成年女儿身高=40.089+0.306×父亲身高+0.431×母亲身高 ±3

§1 相关关系与回归分析

(Regression Analysis)

一. 相关关系与回归函数

在现实世界中存在大量的变量, 它们之间的关系一般分为两类:

1.确定性关系与非确定性关系

确定性关系: 例如,正方形的边长L与它的面积S之间有确定关系 $S = L^2$

2. 非确定性关系：

例：1)股票的价格$\underline{P}$与时间T之间存在关系.

2) 人的身高$\underline{H}$与人的体重W之间的关系.

3) 农作物产量$\underline{Y}$与降雨量 X_1，施肥量X_2，播种量X_3 之间的关系.

该例中的变量关系无法用确定的函数来明确描述.

问题 如何描述各变量间的关系?

3) 农作物产量Y与降雨量 X_1 , 施肥量X_2， 播种量X_3之间的关系.

将作为考察目标的变量称为**因变量**,记为Y (随机变量), 而将影响它的各个变量称为**自变量**或**可控变量**(非随机:可测定或控制),记为 $X_1, X_2, \cdots X_n$

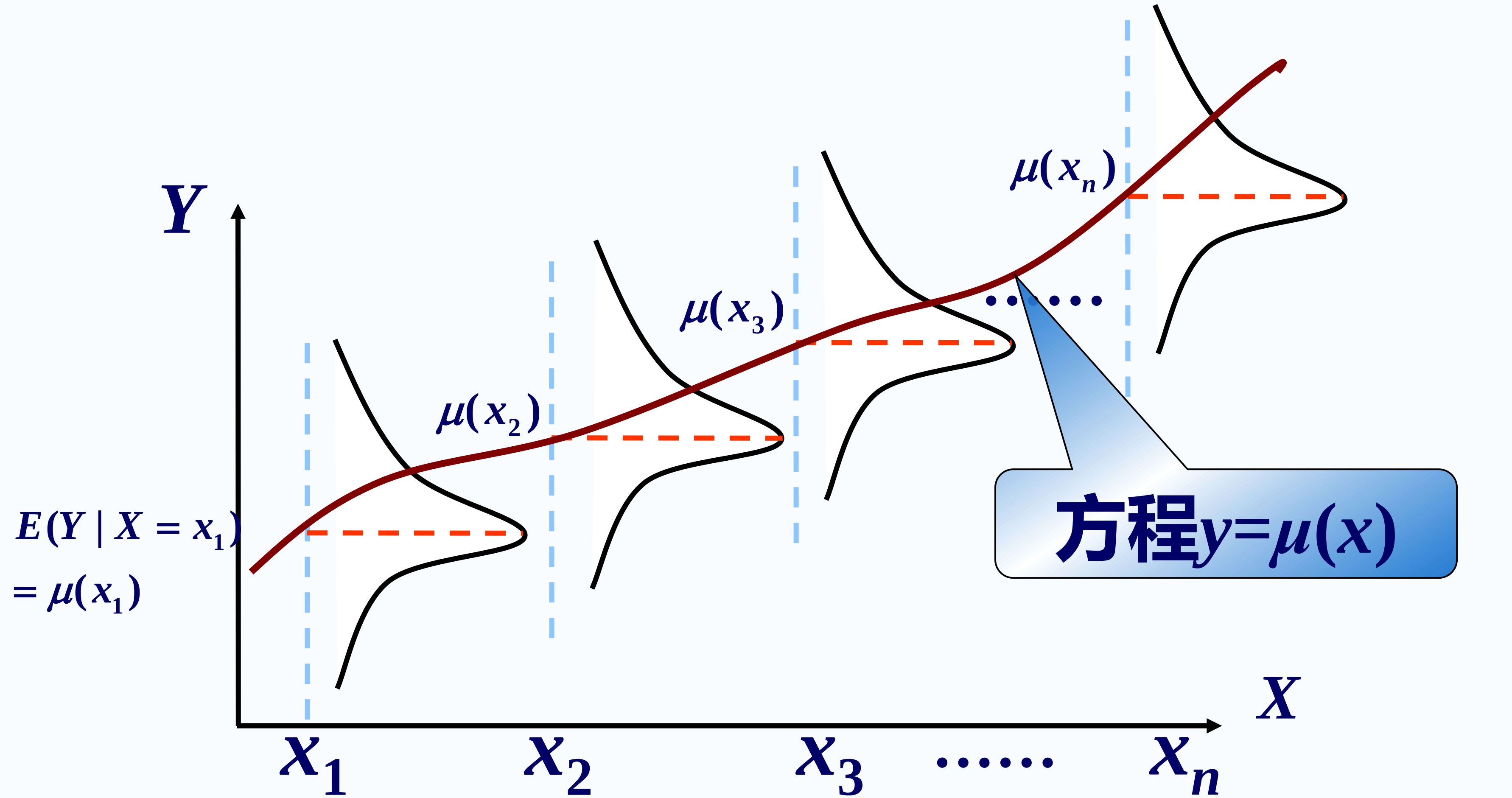
1.确定性的函数关系

用第三章方法可求随机变量函数的分布, 如若已知随机变量L的分布就可以确定函数

$$S = L^2$$ 的分布.

2. 非确定性的相关关系

例: 对X的不同取值x_1，x_2，…, x_n, Y服从条件正态分布



注：$\mu(x)$可理解为在“$X=x$”的条件下,随机变量Y取值最集中的点.

2. 非确定性的相关关系

可控变量$X_1, X_2, \cdots, X_n$的取值记为$x_1, x_2, \cdots, x_n$，若因变量Y的条件**数学期望**：

$$E(Y \mid X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

存在，称Y与$X_1, X_2, \cdots X_n$具有**相关关系**.

相关关系是一种非确定性关系

记：$\mu(x_1, x_2, \cdots, x_n) = E(Y \mid X_1 = x_1, X_2 = x_2, \cdots X_n = x_n)$

定义9.1.1 称

$$\mu(x_1, x_2, \cdots, x_n) = E(Y \mid X_1 = x_1, X_2 = x_2, \cdots X_n = x_n)$$

为Y关于X_1，X_2，$\cdots$ X_n的(理论)**回归函数**，方程

$$y = \mu(x_1, x_2, \cdots, x_n)$$

称为Y对X_1，X_2，$\cdots$ X_n的(理论)**回归方程**.

注 回归函数是确定性的函数.

高尔顿根据1078个父亲及其成年儿子的身高值，研究发现二者间的关系为(cm)

成年儿子身高=85.67+0.516×父亲身高±9.51

回归分析即以回归函数为基础处理相关关系的一种方法.

3.多元回归模型的建立

若Y关于X_1，X_2，$\cdots$ X_n的(理论)回归方程为

$$y = \mu(x_1, x_2, \cdots, x_n)$$

设想：$Y=\mu(x_1, x_2, \cdots, x_n)$ + 随机误差

其它未知的、未考虑的因素以及随机因素的影响所产生.

得数学模型:

$$Y = \mu(x_1, x_2, \cdots, x_n) + \varepsilon$$

有 $\varepsilon = Y - \mu(x_1, x_2, \cdots, x_n)$

ε可视为随机误差，通常要求:

$$Y = \mu(x_1, x_2, \cdots, x_n) + \varepsilon$$

ε可视为随机误差，通常要求:

1) $E(\varepsilon)=0$;　　即$E(Y) = \mu(x_1, x_2, \cdots, x_n)$

2) $D(\varepsilon)=\sigma^2 = E(\varepsilon^2)$**尽可能小.**

注意到　$\sigma^2 = E[Y - \mu(x_1, x_2, \cdots, x_n)]^2$

σ^2是用回归函数近似因变量Y产生的均方误差.

建立模型涉及三个问题:

1) **确定对因变量Y影响显著的自变量;**

2) **确定回归函数$\mu(x)$的类型;**

(由经验或“样本”来假设)

3) **对参数进行估计.**

本章内容

二. 回归函数类型的确定

实际问题中，通常未知回归函数形式.

回归分析的基本思想：

根据自变量X_1，X_2，$\cdots$ X_n与因变量Y的观察值去估计回归函数.

回归分析是寻求变量间近似的函数关系的一种方法.

回归分析的作用，比如(在生产实践上)预测和控制.

本节仅讨论最简单的情形：因变量关于单个自变量的回归函数 $\mu(x)=E(Y|X=x)$.

问题的提法 对两个变量X、Y间的(理论)回归函数$y=\mu(x)$，选择某个(已知类型)函数$f(x)$作为其估计函数：

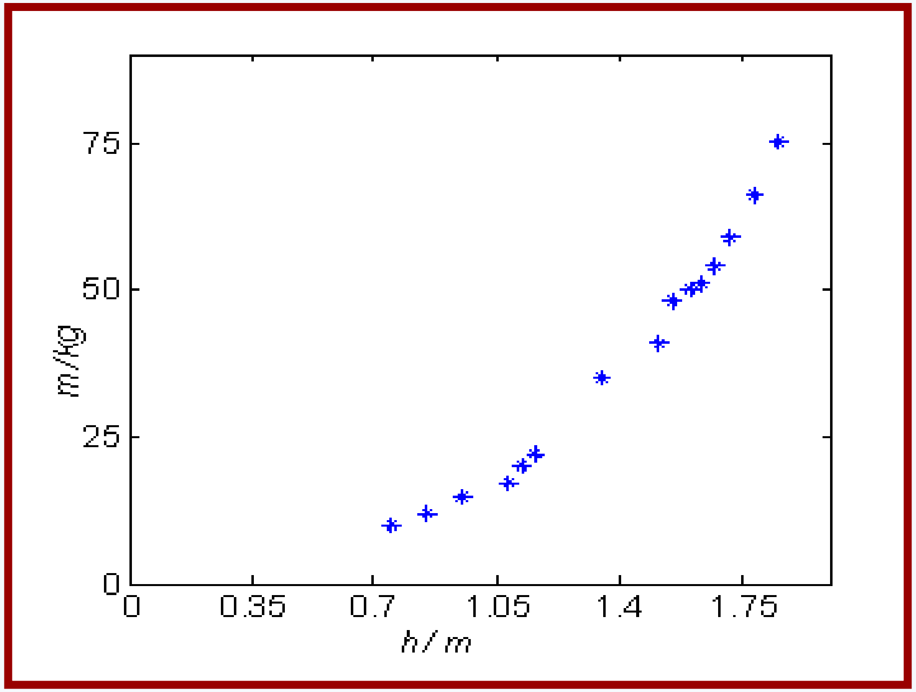
$$\hat{\mu}(x)=f(x)$$

为估计回归函数，可依据问题的背景，确定或假定回归函数的形式.

常通过分析数据散布图获得对变量间相关关系的初步认识.

例9.1.1 身高体重关系

现有15对某地区人的身高h 和体重数据m，希望用简洁的函数关系式描述该地区人的身高体重的对应关系.



呈现函数的增长趋势，可设

$$m = \hat{\mu}(h) = bh^{a}$$

其中a, b是待定参数.

思考 是否能由数据散布图完全确定回归函数?

结论 仅是初步感性的认识，还需进行检验.

变量之间关系

- 确定性
- 非确定性 $\xrightarrow{\text{其中一种}}$ 相关关系(研究重点)

$\mu(x_1,\cdots,x_k)=E(Y\mid X_1=x_1,\cdots,X_k=x_k)$存在

考察目标作为因变量Y（随机变量）

- 可控的、非随机变量X_i
- 随机误差（随机变量）ε

建立多元回归模型：

$$Y=\mu(x_1,\ldots,x_k)+\varepsilon,\qquad E(\varepsilon)=0, D(\varepsilon)=\sigma^2$$

$\mu(x_1,\ldots,x_k)$：因变量Y对自变量的(理论)回归函数

因变量Y对自变量的(理论)回归方程：$\boxed{y=\mu(x_1,\cdots,x_k)}$