方差分析和回归分析

讲师：Jeary

目录

目标

通过本章课程的学习，您将能够：

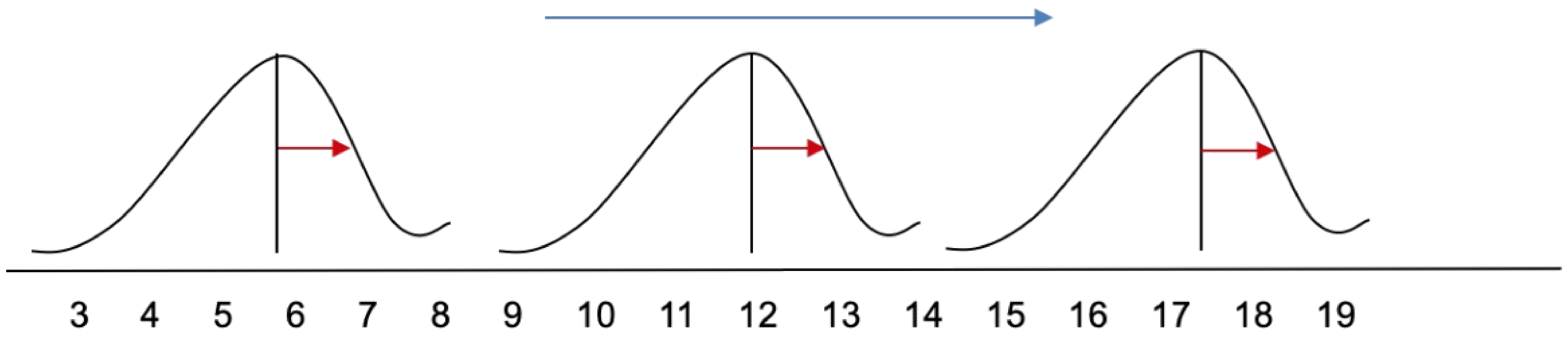
- 掌握方差分析的一般流程
- 掌握相关分析的计算
- 掌握回归分析的主要思想

01

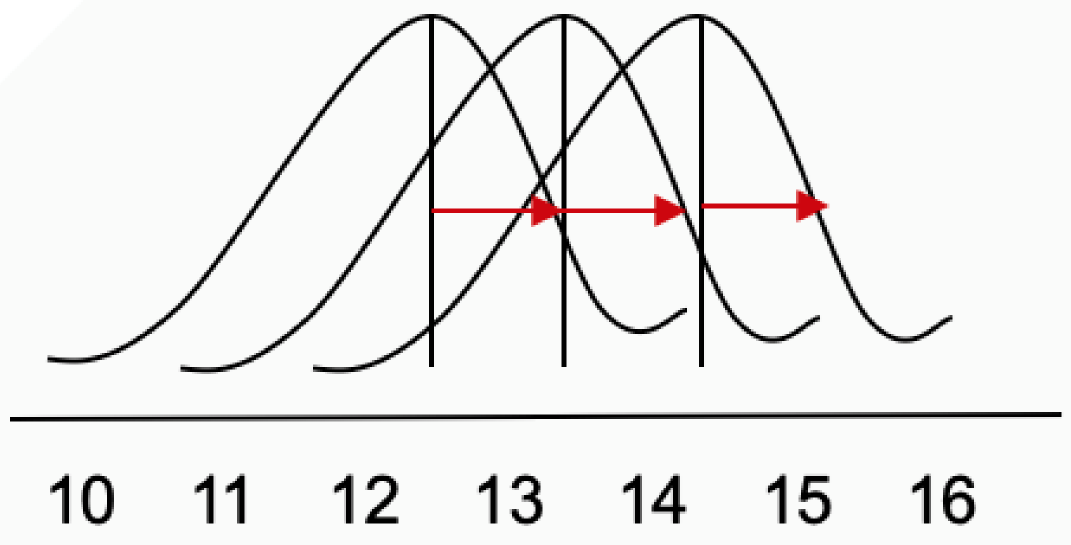
方差分析

定义

- 方差分析 Analysis of Variance，简写为ANOVA；用于两个及两个以上样本均值差别的显著性检验，研究分类型自变量对数值型因变量是否有显著性影响



不同均值



相同均值

基本假设

每个样本服从正态分布

每个样本方差相同

每个样本中个体
相互独立

分类

- 方差分析
  - 单因素方差分析
  - 多因素无交互方差分析
  - 多因素有交互方差分析

01

方差分析-单因素

案例背景

小刘、小张、小李等8个人一起参加100米游泳比赛，最后小张夺得了第一名。我们想知道是什么造成了他们8人的成绩差异。于是我们找来了他们的**身高**、**体重**、**训练时长**、**睡眠时长**等特征变量作为分析的素材。那怎么才能知道这些指标对于衡量游泳成绩是否有效果呢?

总体流程

- 设有 k 组样本，每组有 n 个独立样本， i 表示每组中的第几个样本， j 表示第几个样本组
- 定义零假设： $H_0: \mu_1 = \mu_2 = \ldots = \mu_i = \ldots = \mu_k$

⟶ **假设k个组的样本每组总体均值μ都相同，表示它们来自同一总体分布，这个特征不存在明显的差异**

- 对应的备择假设 H_1 ： $\mu_1, \mu_2 \ldots, \mu_k$ 不全相等

⟶ **表示k组的样本来自不同的总体分布**

$$\bar{X}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

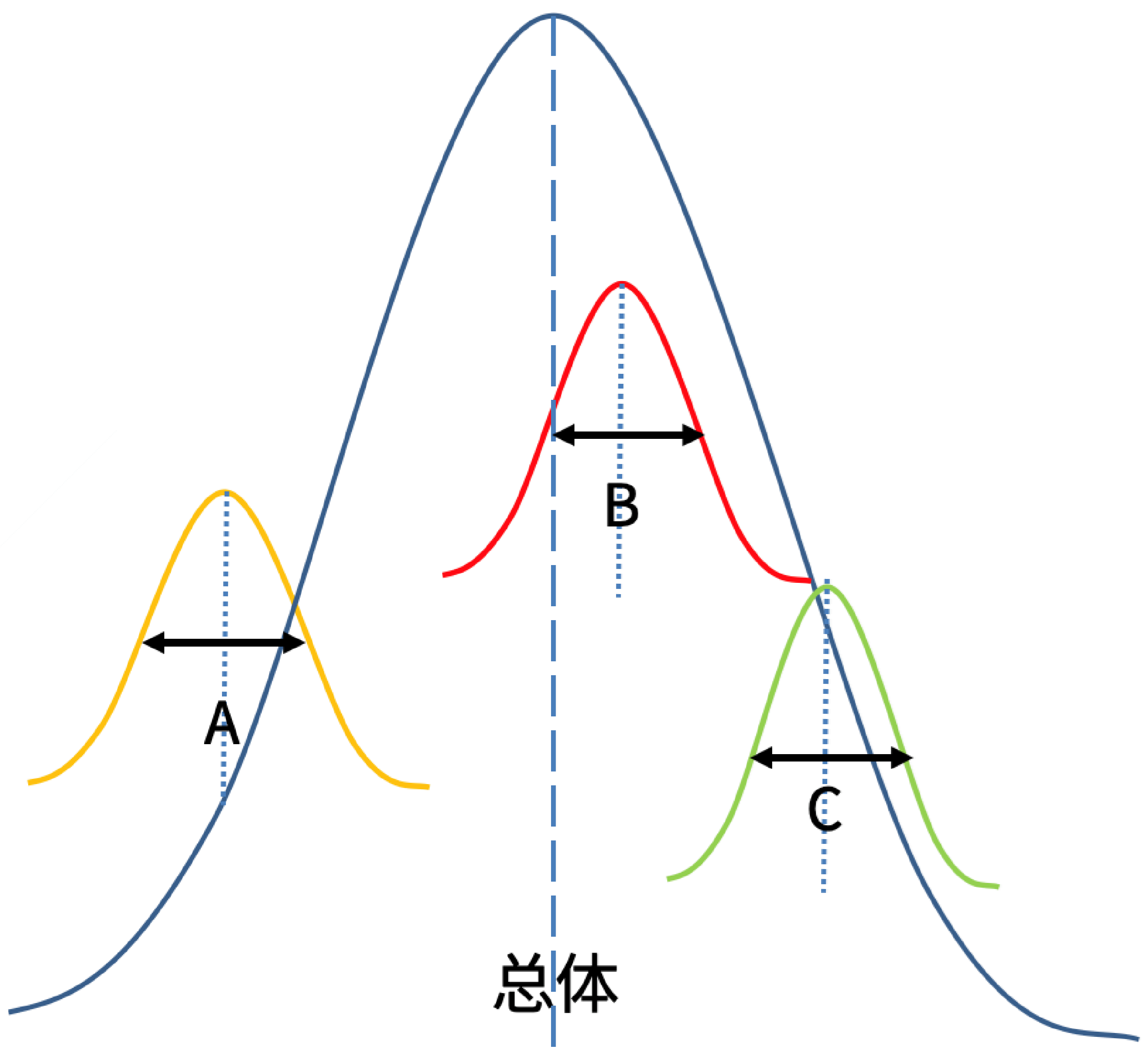
其中$\bar{X}_j$ 表示第j组的样本均值

x_{ij}为第j组的第i个样本

$$s_j^2 = \frac{\sum_{i=1}^{n_j} \left(x_{ij} - \bar{X}_j\right)^2}{n_j - 1}$$

其中s_j^2表示第j组的样本方差

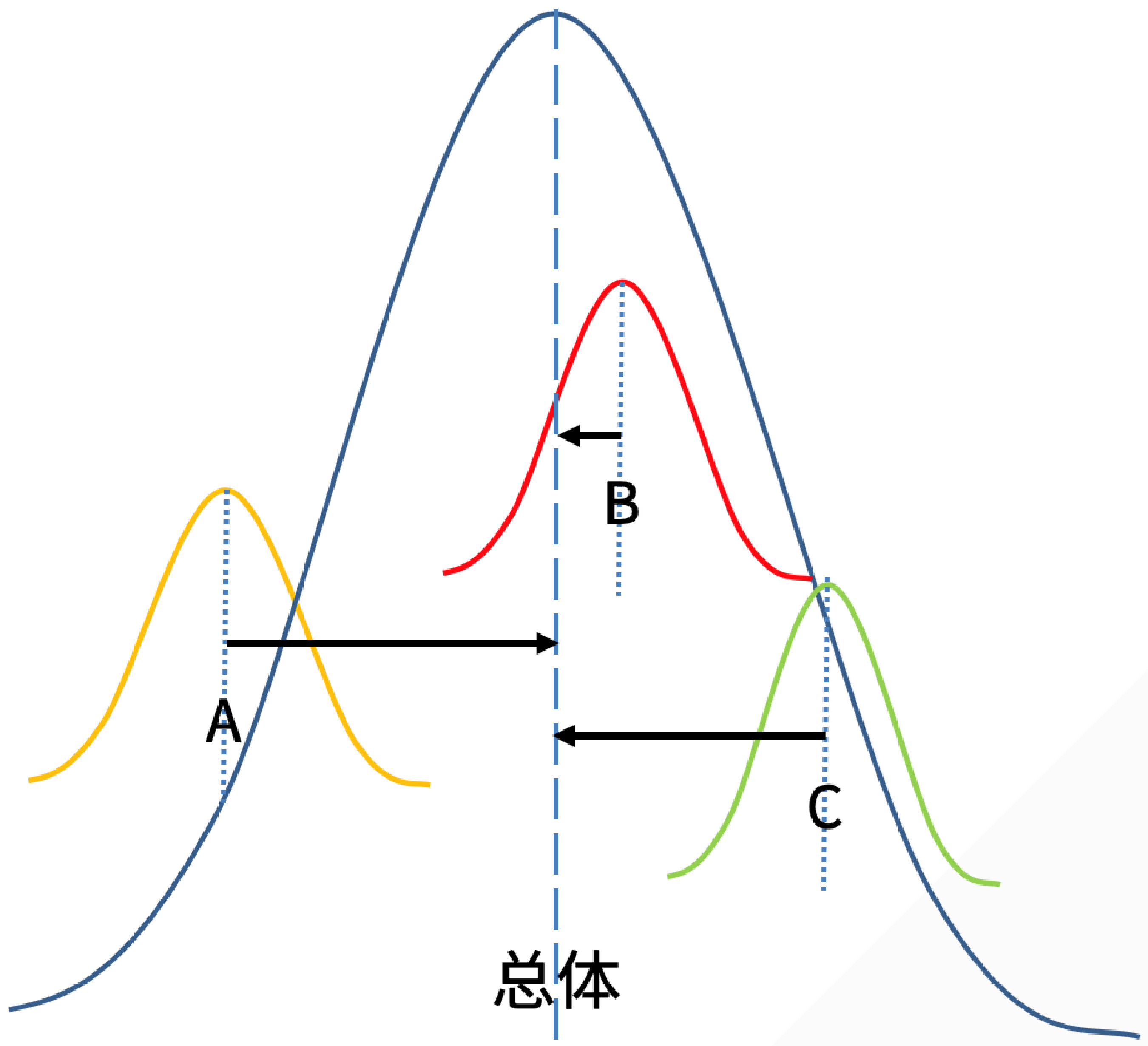
方差



组内均方差MSE

$$MSE = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{n_T - k}$$

n_T表示每个样本容量之和



组间均方差MSR

$$MSR = \frac{\sum_{j=1}^{k} (\bar{x}_j - \bar{\bar{x}})^2 \cdot n_j}{k - 1}$$

其中$\bar{\bar{x}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}}{n_T}$表示总的样本均值

F分布

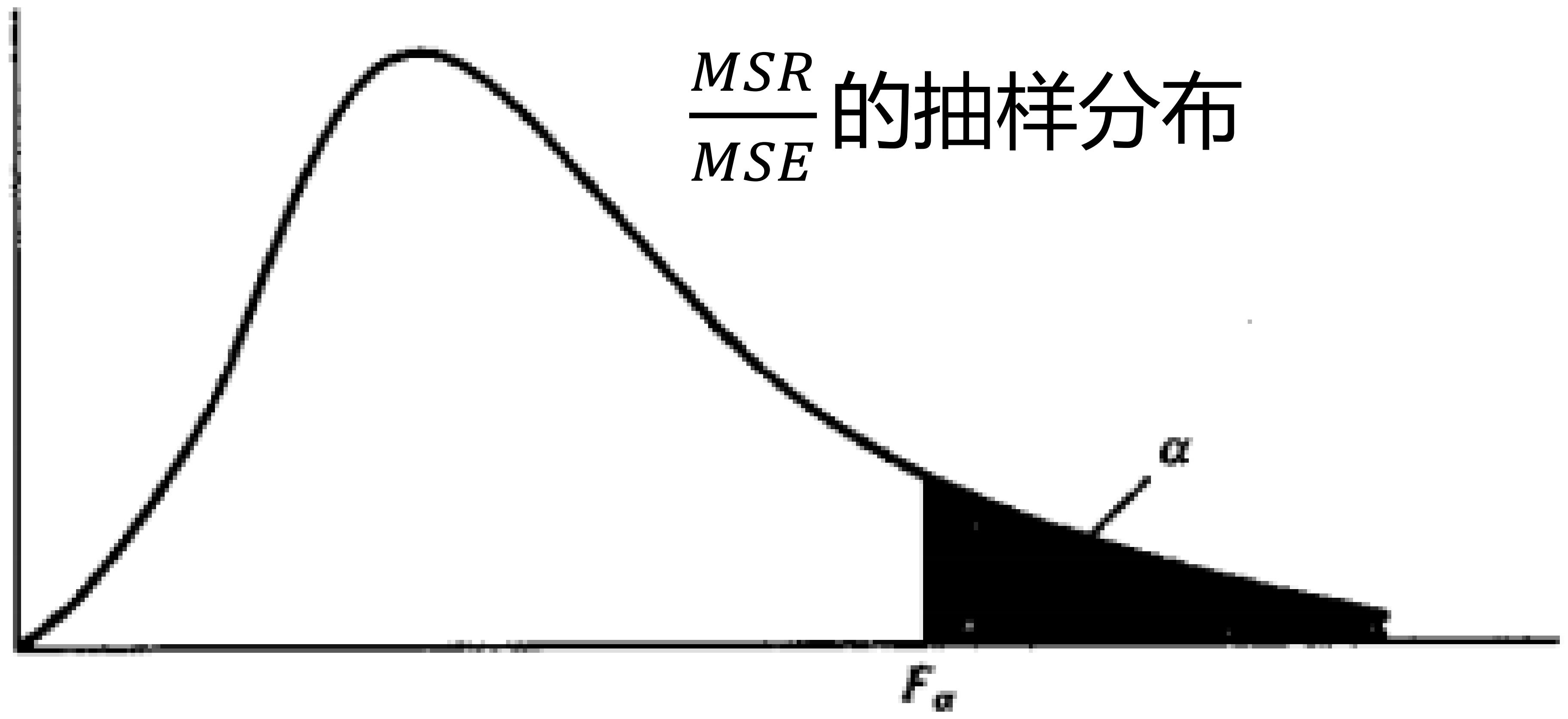
- 如果零假设为真，总体方差的组间估计和组内估计的比值，服从分子自由度为 k − 1 ，分母自由度为 $n_T - k$的 F 分布:

$$F = \frac{MSR}{MSE}$$

自由度k-1

自由度$n_T - k$

$\frac{MSR}{MSE}$的抽样分布



临界值

F分布和当中的拒绝域

F分布-案例

- 给定显著性水平α，F分布对应的临界值为F_α，当 $F=\frac{MSTR}{MSE}>F_\alpha$时，拒绝$H_0$假设、接受$H_1$假设

- 上述案列中，8个同学中有4个高于180cm，4个低于180cm，那么以身高水平将他们分为两组，测试这两组样本是否来自同一总体。

在若显著性水平α下，F的统计量大于对应的临界值F_α，则拒绝原假设H_0，接受备择假设H_1，身高对成绩的影响是显著的

01

方差分析-多因素无交互

案例背景

■ 多因素方差分析：

研究一个因变量是否受到多个自变量的影响，检验多个因素取值水平的不同组合之间，因变量的均值之间是否存在显著的差异，其中，双因素方差分析是最基础的多因素分析。

■ 无交互方差分析：

例如：在**游泳成绩**的影响因素研究中，因素A睡眠时长、因素B身高是两个相互独立因素，因素A与因素B不存在互相关系。

总体流程

- 对于不存在交互作用的观测 $\{X_{ij}\}$，采用以下的模型：

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

$$1 \le i \le \mathrm{a}，1 \le j \le \mathrm{b}$$

其中μ表示平均的效应, α_i和 β_j分别表示因素A的第 i 个水平和因素 B 的第 j 个水平的附加效应，ε_{ijk} 为误差，假定其是独立且是等方差的正态分布.

假设

- 以无交互作用双因子方差分析为例

因素A

原假设 H_{01}： $\alpha_1 = \alpha_2 = \cdots = \alpha_a$

备择假设 H_1： 至少一个α_i不等于0

因素B

原假设 H_{02}： $\beta_1 = \beta_2 = \cdots = \beta_b$

备择假设 H_2： 至少一个β_i不等于0

计算

- 总偏差平方和

$$SST=\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{\mathrm{ij}}-\bar{x})^2$$

- 因素A的偏差平方和

$$SSA=b\sum_{i=1}^{a}(x_{\mathrm{i}}-\bar{x})^2$$

- 因素B的偏差平方和

$$SSB=a\sum_{j=1}^{b}(x_{j}-\bar{x})^2$$

- 随机误差的偏差平方和

$$SSE=\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{\mathrm{ij}}-x_{\mathrm{i}}-x_{j}-\bar{x})^2$$

SST=SSA+SSB+SSE

计算

- A因素的均方，记为MSA:

$$MSA = \frac{SSA}{a-1}$$

$$F_A = \frac{MSA}{MSE}$$

- B因素的均方，记为MSB:

$$MSB = \frac{SSB}{b-1}$$

$$F_B = \frac{MSB}{MSE}$$

- 随机误差项的均方，记为MSE:

$$MSE = \frac{SSE}{(a-1)(b-1)}$$

案例

给定显著性水平α， F分布对应的临界值为F_α，当 $F_A=\frac{MSA}{MSE}>F_\alpha$、 $F_B=\frac{MSB}{MSE}>F_\alpha$时，拒绝$H_{01}$、 H_{02}假设，接受H_1、H_2假设

- 当**因素A睡眠时长**的F统计量大于显著性水平α 下临界值 F_α时，拒绝原假设H_{01}，接受备择假设H_1，睡眠时长影响游泳成绩的平均值

- 当**因素B身高**的F统计量大于显著性水平α 下临界值 F_α时，拒绝原假设H_{02}，接受备择假设H_2，身高影响游泳成绩的平均值

01

方差分析-多因素有交互

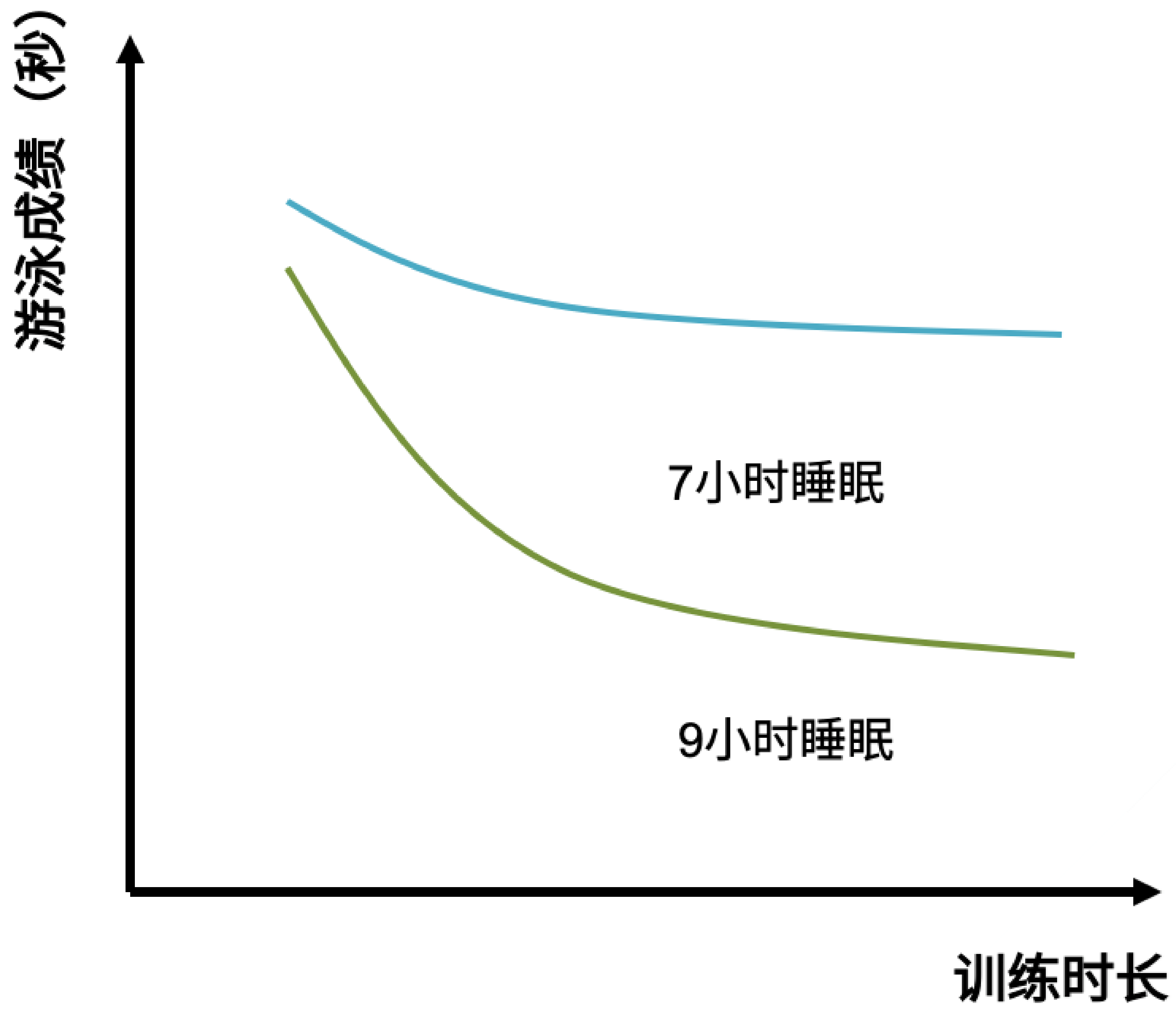
简介

- 有交互方差分析：

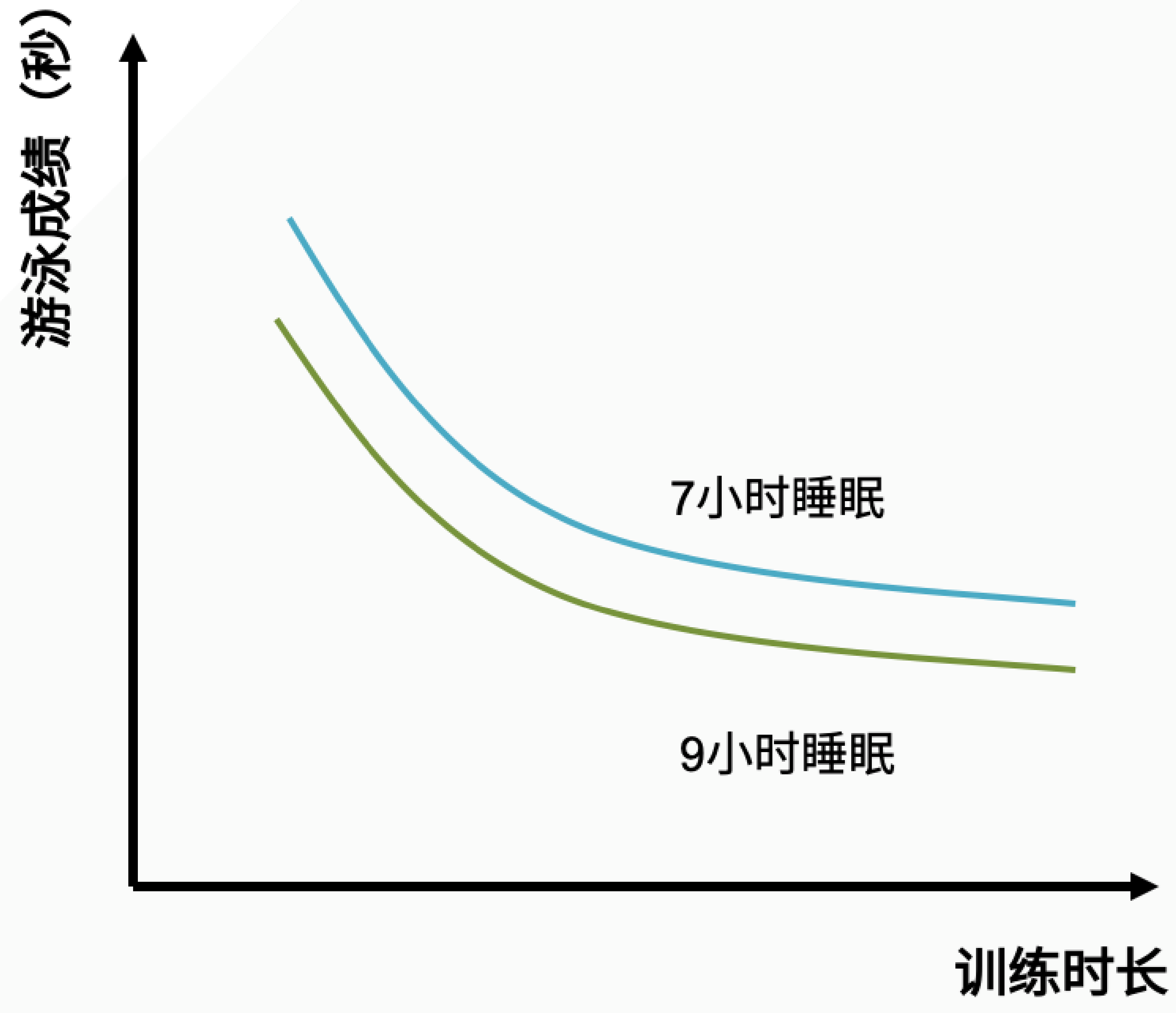
若多因素对实验结果的影响非独立，则进行**有交互方差分析**

- **例如：**在**训练时长**与**睡眠时长**对游泳成绩影响的研究中，不仅单个因素会对成绩造成影响，两个因素不同水平的搭配还会产生新的影响，则可以认为**因素A训练时长**与**因素B睡眠时长**产生交互效应

简介



睡眠时长与训练时长有交互影响



睡眠时长与训练时长无交互影响

多因素有交互分析

因素A

- 原假设 H_{01}： $\alpha_1 = \alpha_2 = \cdots = \alpha_a$
- 备择假设 H_1： 至少一个α_i不等于0

因素B

- 原假设 H_{02}： $\beta_1 = \beta_2 = \cdots = \beta_b$
- 备择假设 H_2： 至少一个β_i不等于0

交互作用

原假设 H_{03}： $\alpha\beta_{11} = \alpha\beta_{12} = \cdots = \alpha\beta_{ab}$

备择假设 H_3： 至少一个$\alpha\beta_{ij}$不等于0

多因素有交互方差分析

- 对于存在交互作用的观测 X_{ijk}，采用以下的模型：

→ **除了无交互作用双因子方差分析，可能存在两种因素同时作用，具有交互作用，(A_i, B_j)下作了r个试验**

- **公式**

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$$1 \le k \le r,\ 1 \le i \le a,\ 1 \le j \le b$$

- 其中μ表示平均的效应, α_i 和 β_j分别表示因素 A 的第 i 个水平和因素B的第 j 个水平的附加效应， γ_{ij} 表示因素A的第i个水平和因素 B 的第 j 个水平交互作用的附加效应。 ε_{ijk} 为误差，这里也假定它是独立的并且是等方差的正态分布。

计算

- 总偏差平方和

$$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} r\,(x_{ij} - \bar{x})^2$$

- 因素A的偏差平方和

$$SSA = br\sum_{i=1}^{a}(x_i - \bar{x})^2$$

- 因素B的偏差平方和

$$SSB = ar\sum_{i=1}^{b}(x_j - \bar{x})^2$$

- 因素AB交互作用的偏差平方和

$$SSAB = r\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} r\,(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$$

- 随机误差平方和

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} r\,(x_{ijk} - \bar{x}_{ij})^2$$

SST=SSA+SSB+SSAB+SSE

计算

- A因素的均方，记为MSA:

$$MSA = \frac{SSA}{a-1}$$

$$F_A = \frac{MSA}{MSE}$$

- B因素的均方，记为MSB:

$$MSB = \frac{SSB}{b-1}$$

$$F_B = \frac{MSB}{MSE}$$

- AB因素交互作用的均方，记为MSAB:

$$MSAB = \frac{SSAB}{(a-1)(b-1)}$$

$$F_{AB} = \frac{MSAB}{MSE}$$

- 随机误差项的均方，记为MSE:

$$MSE = \frac{SSE}{(a-1)(b-1)}$$

F分布-案例

- 给定显著性水平α，F分布对应的临界值为F_α，当$F_A=MSA/MSE>F_\alpha$、$F_B=MSB/MSE>F_\alpha$、$F_{AB}=MSAB/MSE>F_\alpha$时，拒绝H_{01}、H_{02}、H_{03}假设，接受H_1、H_2、H_3假设。

- 当**因素A睡眠时长**的F统计量大于显著性水平α下临界值F_α时，拒绝原假设H_{01}，接受备择假设H_1，睡眠时长影响游泳成绩的平均值

- 当**因素B训练时长**的F统计量大于显著性水平α下临界值$F\alpha$时，拒绝原假设H_{02}，接受备择假设H_2，训练时长影响游泳成绩的平均值

- 当**因素AB交互影响**的F统计量大于显著性水平α下临界值F_α时，拒绝原假设H_{03}，接受备择假设H_3，交互效应影响游泳成绩的平均值

02

相关分析

相关分析

相关分析：研究**两个**或**两个以上**处于同等地位的随机变量间的相关关系的统计分析方法

- **单相关**： 两个因素之间的相关关系叫单相关，即研究时只涉及一个自变量和一个因变量
- **复相关**：三个或三个以上因素的相关关系叫复相关，即研究时涉及两个或两个以上的自变量和因变量相关
- **偏相关**：在某一现象与多种现象相关的场合，当假定其他变量不变时，其中两个变量之间的相关关系称为偏相关

相关系数R：

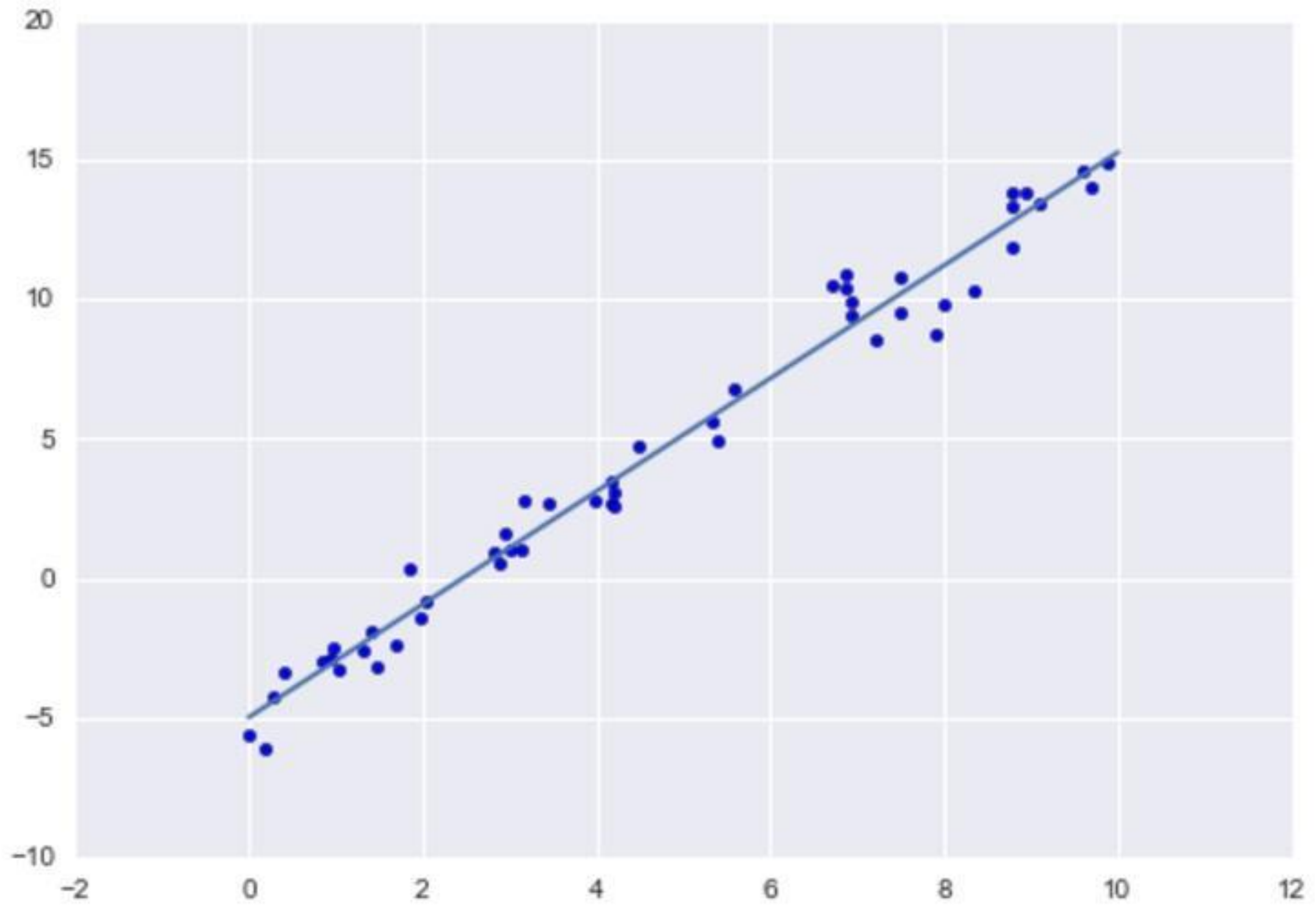
$$R = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

03

回归分析

定义

回归分析 利用数据统计原理，对大量统计数据进行数学处理，并确定因变量与某些自变量的相关关系，建立一个相关性较好的回归方程 (函数表达式) 用于预测今后的因变量的变化的分析方法



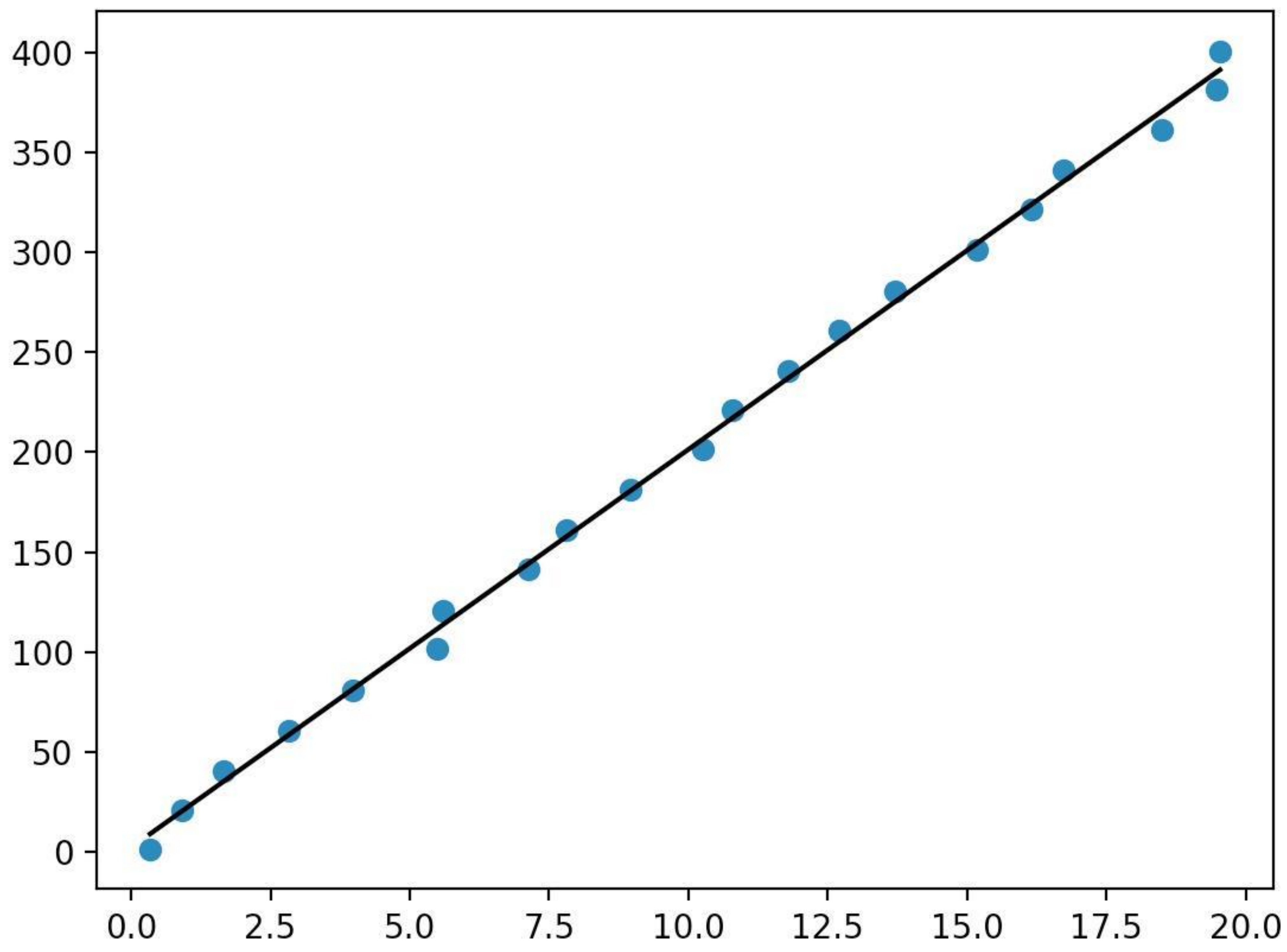
分类

- 回归分析
  - 一元线性回归
  - 多元线性回归

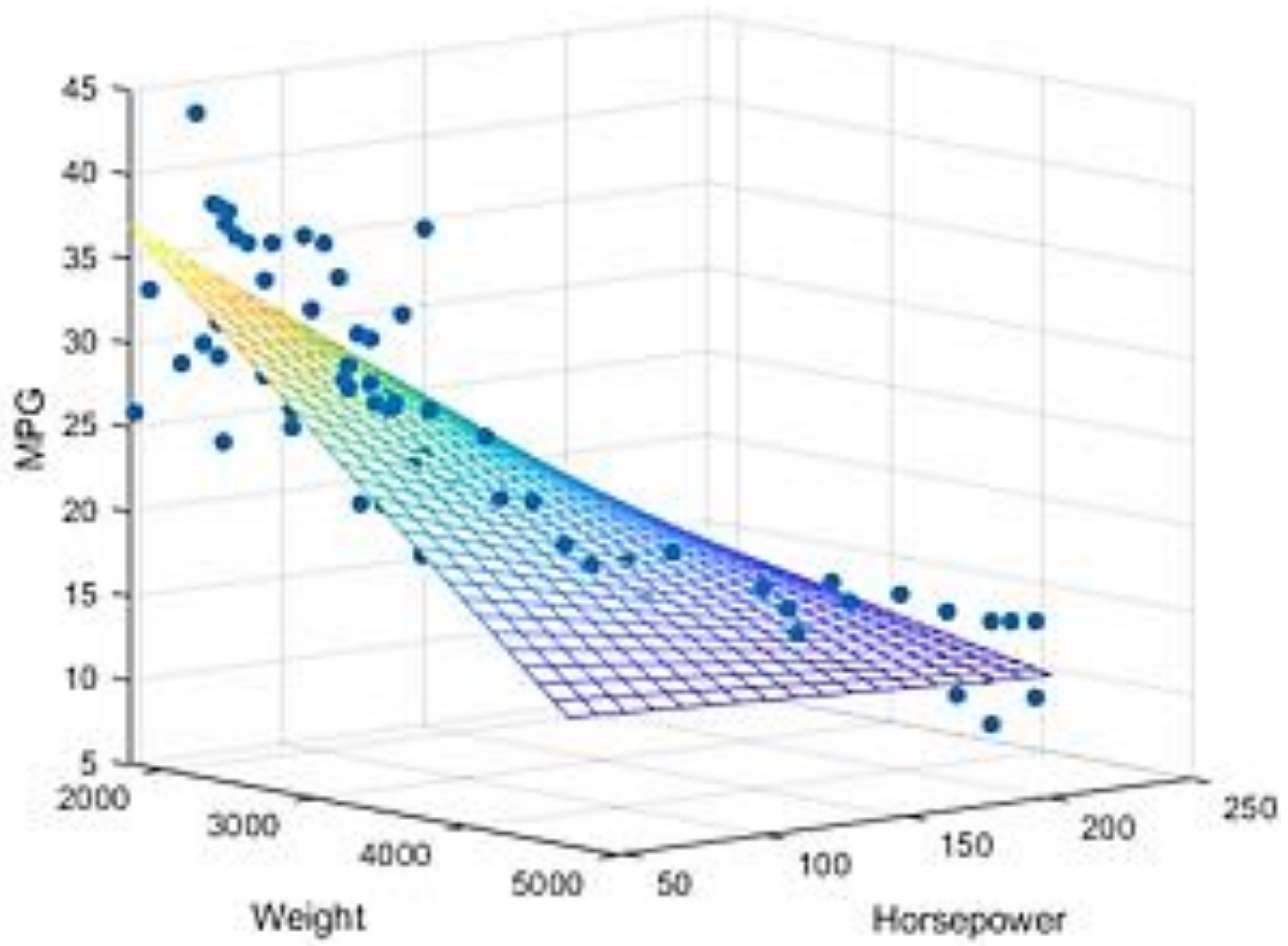
步骤

初步设定回归方程 → 求回归系数 → 确定相关系数 → 计算预测值的置信区间

一元线性回归



多元线性回归

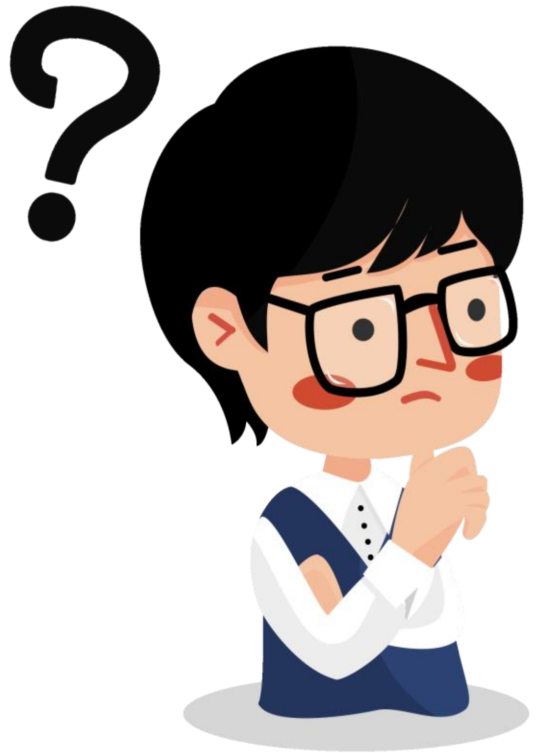


45
40
35
30
25
20
15
10
5
MPG
2000
3000
4000
5000
Weight
50
100
150
200
250
Horsepower

04

总结

思考题



01

以下说法中，不属于一元线性回归分析特点是：

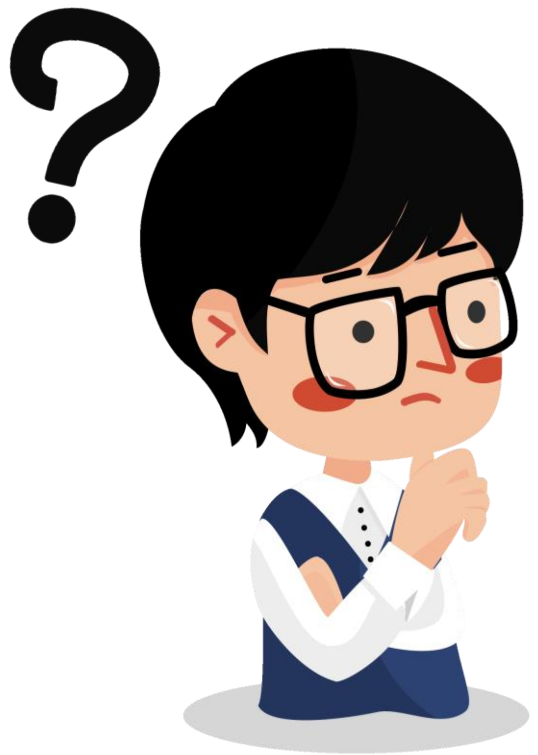
A、两个变量是非对等关系

B、可求出两个回归方程

C、利用一元回归方程，两个变量可以互相推算

D、自变量是非随机的，因变量是随机的

思考题



02

相关分析和回归分析是统计中常用的两种分析方法，那么以下关于这两种说法中错误的一项是：

A、相关分析和回归分析都可以得到变量之间关系的方向和强弱程度

B、回归分析中，因变量是随机变量，而自变量是非随机的

C、相关分析的研究目的是确定自变量和因变量、变量之间是因果关系

D、回归分析可以得到变量之间具体数量变动关系，而相关分析不可以

总结

本章包含三小节内容：

第一节

- 方差分析的相关计算方法

第二节

- 相关分析的概念以及计算

第三节

- 回归分析的概念与步骤

谢谢观看

参考书目：概率论与数理统计•第四版（浙江大学） 高等教育出版社

网易云课堂 × 微专业