第二章　生物统计基础

蒋杭进 & 王涛

本章内容

王涛

蒋杭进

第一节　生物统计简介

第二节　参数估计和假设检验

第三节　统计模型

第四节　高维统计方法

第五节　统计学习基础

第六节　统计因果推断

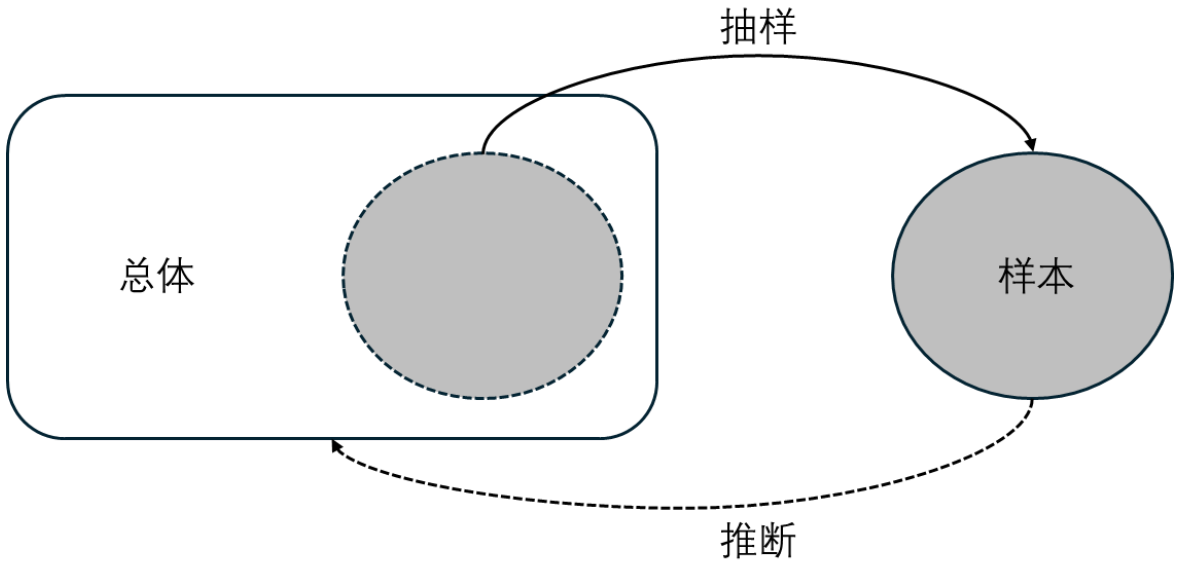
第一节 生物统计简介

1. 总体与样本的关系

2. 数据特征及描述

3. 常见的图表

总体与样本的关系



统计推断是在一定的假设下，基于数据反向推断真实的情况

数据特征及描述

数据类型		定义	示例
数值型 (numeric)	连续型 (continuous)	表示在一定范围内可以取任意数值的数据	身高、体重、温度等
	离散型 (discrete)	表示在有限数值集合中取值的数据	年龄、家庭人数等
分类型 (categorical/nomical)		表示不同类别或属性的数据，没有数值意义，只表示不同的类别	性别、血型等
顺序型 (ordinal)		表示按照一定顺序排列的数据，但不能进行精确的数值计算	肿瘤分期、烧伤程度等

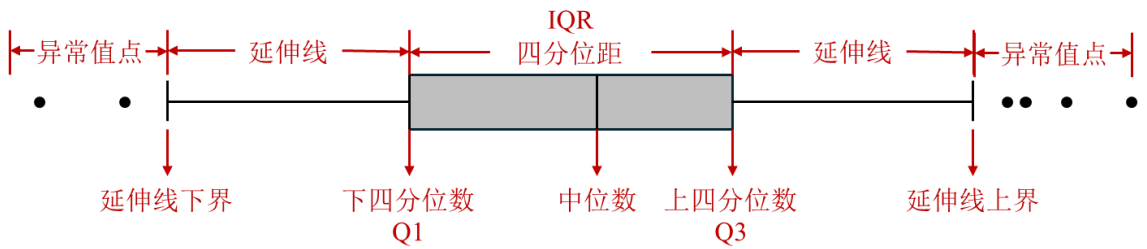
数据特征及描述

特征	样本-统计量	总体-参数
均值	$\bar{X}$	μ
中位数	$\tilde{X}$	$\tilde{\mu}$
方差(标准差)	$s^2(s)$	$\sigma^2(\sigma)$
比例	$\hat{p}$	p
众数	$\widehat{Mo}$	Mo
极差	r	R
四分位距	IQR	—
变异系数	$s/\bar{X}$	σ/μ
偏度	g_1	γ
峰度	g_2	κ

用来衡量数据相对于其均值的离散程度

衡量数据分布的不对称程度

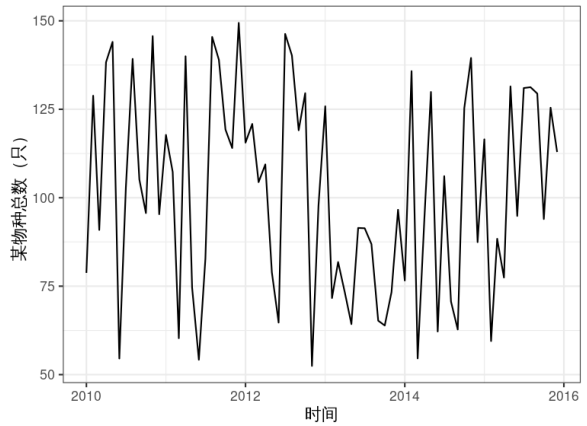
反映数据分布的尖峭或扁平程度



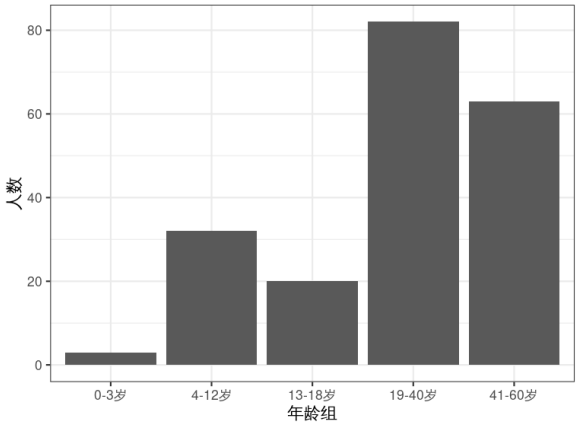
常见的图表

频数频率表

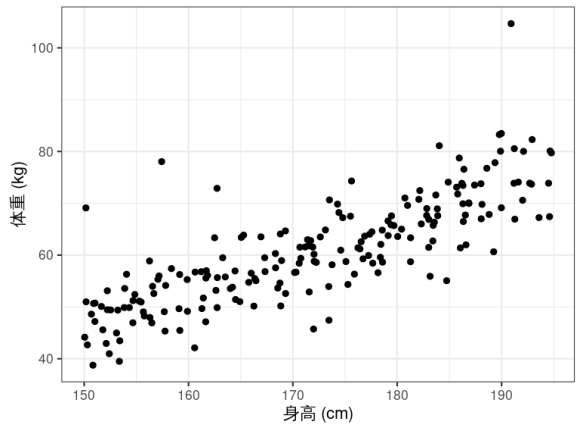
物种	A	B	C	D	总和
频数	55	38	87	20	200
频率	0.275	0.190	0.435	0.100	1.000



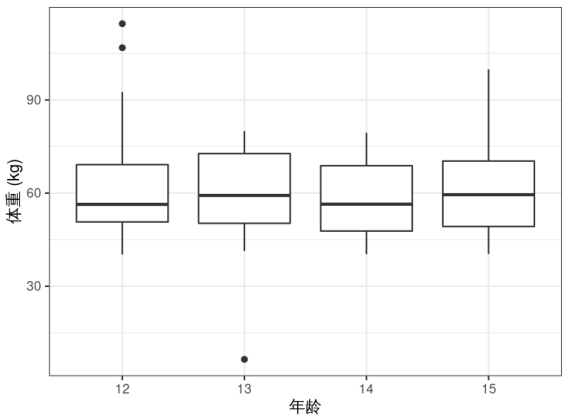
折线图



直方图



散点图



箱线图

第二节 参数估计与假设检验

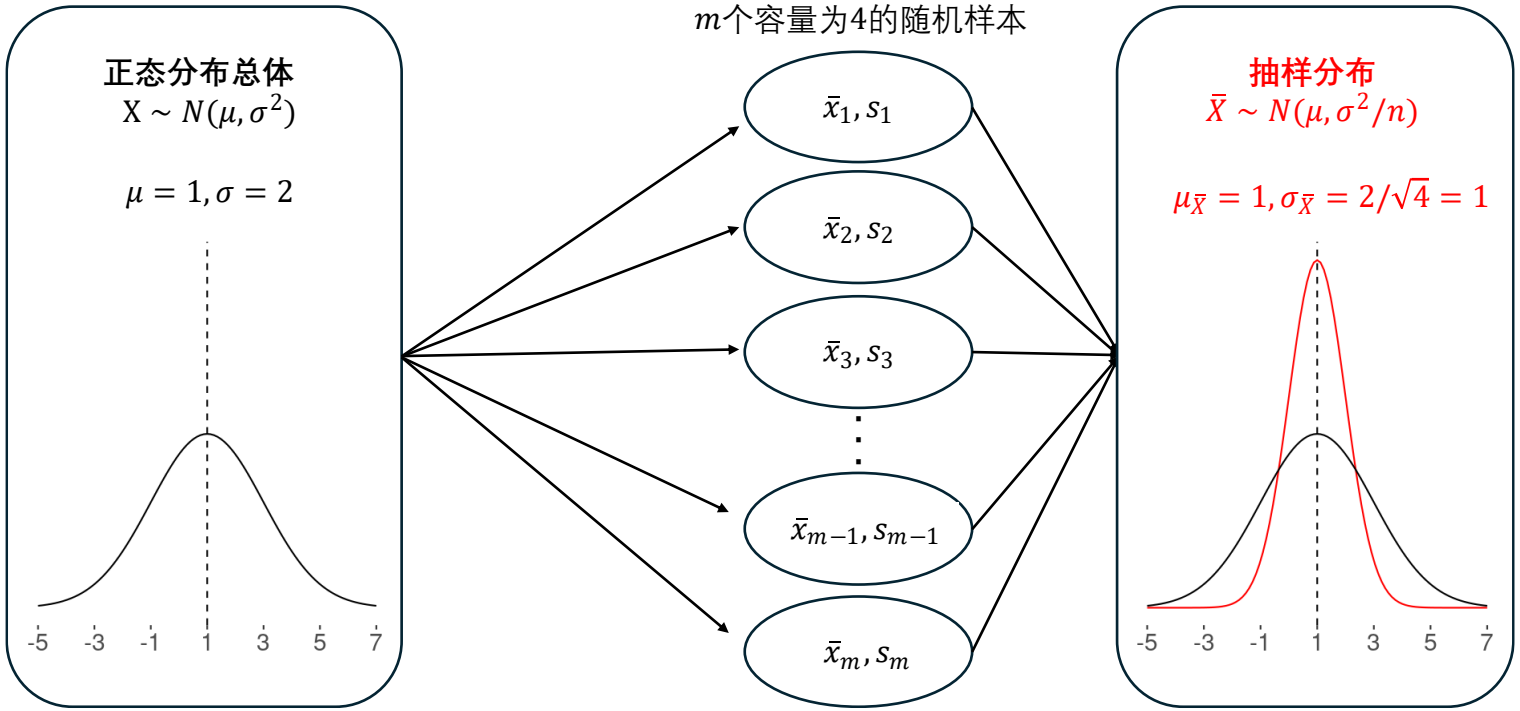
1. 参数估计

2. 假设检验

第二节 参数估计与假设检验

统计推断（statistical inference）是根据样本信息对总体分布或总体的特征进行推断

- 参数估计（**parameter estimation**）是使用样本来估计总体分布中包含的未知参数或参数的函数的方法。
- 假设检验（**hypothesis testing**）是在抽样分布和小概率原理的基础上，使用样本对总体的分布或分布中所含参数的假设进行检查的方法和过程。



第二节 参数估计与假设检验

统计推断 = 参数估计 + 假设检验

$\{x_1, x_2, \ldots, x_n\} \sim f$ 比如：100只同样品种小鼠的基因A的表达量

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

估计 ⟹

$$EX = \int x f(x) dx$$

统计量

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

假设检验 ⟹ 该品种小鼠的基因A高表达?

第二节 参数估计与假设检验

统计推断 = 参数估计 + 假设检验

$\{x_1, x_2, \ldots, x_n\} \sim f$ **比如：100只同样品种小鼠的基因A的表达量**

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

假设检验 ⟹ 该品种小鼠的基因A高表达?

单样本检验：t-test, z-test, Wilcox 符合秩检验, Wilcox 符合秩和检验

已知$\sqrt{n}(\bar{X} - \mu)/\sigma$服从标准正态分布，故

$$P(-Z_{1-\alpha/2} < \sqrt{n}(\bar{X} - \mu)/\sigma < Z_{1-\alpha/2}) = 1 - \alpha,$$

置信区间（confidence interval）是在给定的置信水平$1-\alpha$下，基于样本统计量构造的总体参数的一种区间估计。

统计推断 = 参数估计 + 假设检验

$\{x_1, x_2, \ldots, x_n\} \sim f$ **比如：100只同样品种小鼠的基因A的表达量**

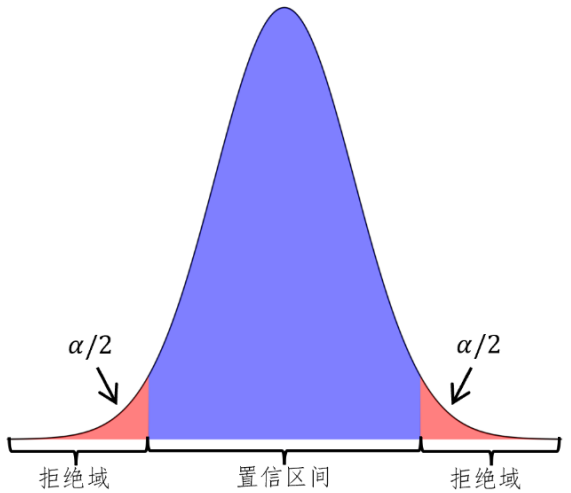
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

假设检验 ⇒ 该品种小鼠的基因A高表达?

单样本检验：t-test, z-test, Wilcox 符号秩检验

已知$\sqrt{n}(\bar{X} - \mu)/\sigma$服从标准正态分布，故

$$P(-Z_{1-\alpha/2} < \sqrt{n}(\bar{X} - \mu)/\sigma < Z_{1-\alpha/2}) = 1 - \alpha,$$



置信区间（confidence interval）是在给定的置信水平$1-\alpha$下，基于样本统计量构造的总体参数的一种区间估计。

第二节 参数估计与假设检验

统计推断 = 参数估计 + 假设检验

$\{x_1, x_2, \dots, x_n\} \sim f$ **100只品种1的基因A表达量**

$\{y_1, y_2, \dots, y_n\} \sim g$ **100只品种2的基因A表达量**

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

统计量

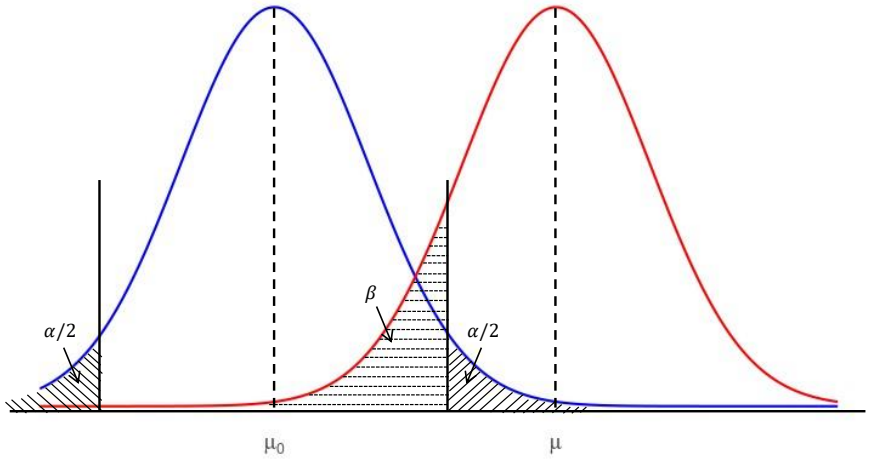
$\bar{x} - \bar{y}$ —**假设检验**→ 品种1的表达量高于品种2的表达量?

双样本均值检验： 双样本t-test, z-test, Wilcox 秩和检验

配对检验： 配对t-test,配对z-test, Wilcox 符号秩检验

统计假设检验中的错误

	原假设正确	原假设错误
拒绝原假设	拒真 (Type I error)	Power
接受原假设		纳假 (Type II error)



统计假设检验的思想是在控制第一类错误的前提下，第二类错误越低越好。

统计量T, 拒绝域C_0, 在 H_0下，$P(T \in C_0) \leq \alpha$

多重假设检验矫正

	H_0 is true	H_0 is false (H_1 is true)	Total
Reject H_0(Test is declared significant)	V	S	R
Fail to reject H_0(Test is declared non-significant)	U	T	$m-R$
Total	m_0	$m-m_0$	m

$$\mathrm{FDR} = E\frac{V}{\max\{1, V+S\}} = E\frac{V}{\max\{R, 1\}}$$

BH 矫正

$$\mathrm{FWER} = P(V \geq 1)$$

Bonferroni 矫正

FDR 和 FWER 是两种不同的统计学概念，不能够直接比较

方差分析: 单因素方差分析和多因素方差分析

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k \Leftrightarrow H_A$: $\mu_j, j = 1, .., k$不全相等。

来源	平方和	自由度	均方	F统计量	拒绝域
组间	SS_B	$k-1$	MS_B	$F = MS_B / MS_W$	$F > F_{1-\alpha}(k-1, n-k)$
组内	SS_W	$n-k$	MS_W		
总和	SS_T	$n-1$			

其他检验方法

拟合优度检验（goodness-of-fit test）是一种用于检验观测数据与理论分布之间是否存在显著差异的统计方法。

分组	1	2	…	k	总和
观测频数O_i	O_1	O_2	…	O_k	n
期望频数E_i	$E_1 = np_1$	$E_2 = np_2$	…	$E_k = np_k$	n
X^2	$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$				

其他检验方法

独立性卡方检验（**chi-square test for independence**）是一种用于检验两个分类变量之间是否独立的统计方法。

		变量B				总和
变量A		1	2	…	k	
	1	O_{11} $(E_{11} = R_1C_1/n)$	O_{12} $(E_{11} = R_1C_1/n)$	…	O_{1k} $(E_{1k} = R_1C_k/n)$	R_1
	2	O_{21} $(E_{21} = R_2C_1/n)$	O_{22} $(E_{22} = R_2C_2/n)$	…	O_{2k} $(E_{2k} = R_2C_k/n)$	R_2
	…	…	…	…	…	…
	m	O_{m1} $(E_{m1} = R_mC_1/n)$	O_{m2} $(E_{m2} = R_mC_2/n)$	…	O_{mk} $(E_{mk} = R_mC_k/n)$	R_m
总和		C_1	C_2	…	C_k	n
X^2		$\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ 或 $\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(\lvert O_{ij}-E_{ij}\rvert-0.5)^2}{E_{ij}}$ （连续性校正）				

其他检验方法

Fisher精确性检验

		变量B		总和
变量A		1	2	
	1	a	b	$a+b$
	2	c	d	$c+d$
总和		$a+c$	$b+d$	n
$P(O_{11}=a)$		$\frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{n!\,a!\,b!\,c!\,d!}$		

需要注意的是，p值的含义是如果原假设为真，观察到当前数据或更极端数据的概率。因此计算Fisher精确性检验的p值时，需要同时考虑当前数据出现的可能性以及更极端数据出现的可能性，这些可能性的总和才是最终的p值。

其他检验方法

Fisher精确性检验

例子：对于数据$a=1,b=9,c=10,d=4$，
在保持行列总和不变的情况下，更极端数据为$a=0,b=10,c=11,d=3$。

此时，两种数据p值分别为

$$P(O_{11}=1)=\frac{(1+9)!\,(10+4)!\,(1+10)!\,(9+4)!}{24!\,1!\,9!\,10!\,4!}=0.004\ ,$$

$$P(O_{11}=0)=\frac{(0+10)!\,(11+3)!\,(0+11)!\,(10+3)!}{24!\,0!\,10!\,11!\,3!}=0.00015。$$

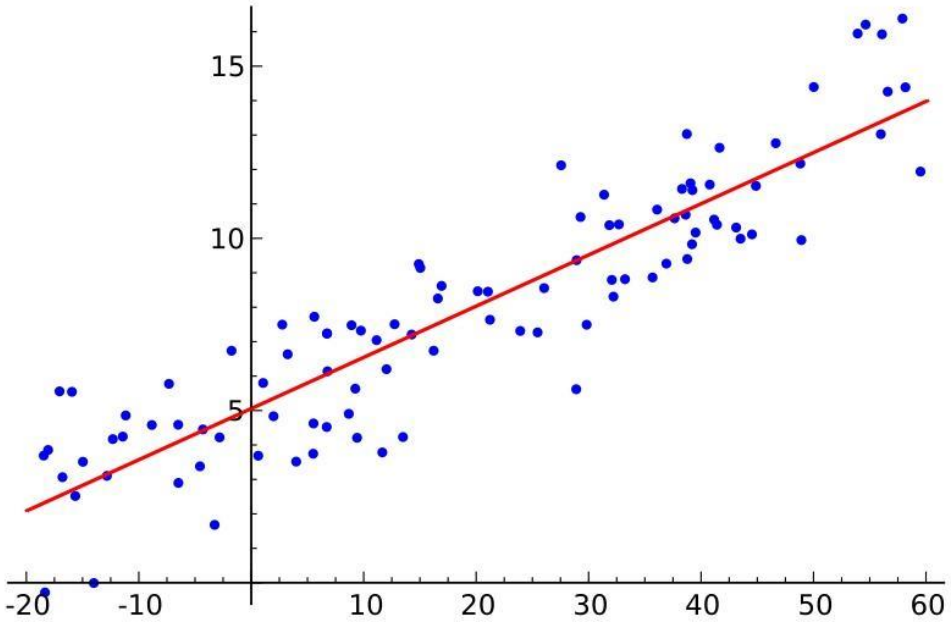
因此，p值$=P(O_{11}=1)+P(O_{11}=0)=0.00415$。

第三节 统计模型

1. 线性模型
2. 极大似然估计方法*
3. Logistic 回归模型
4. 泊松回归模型
5. Cox 比例风险模型
6. 线性混合效应模型
7. 隐马尔科夫模型
8. 贝叶斯统计基础

线性模型

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$



损失函数

$$L(y, \hat{y}) = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

$L(y, \hat{y})$ 需要满足以下两个条件：

1. 对任意的$\hat{y}$都有$L(y, \hat{y}) \geq 0$，当$y = \hat{y}$时，$L(y, \hat{y}) = 0$
2. $L(y, \hat{y})$ 是$\hat{y}$的凸函数；

$$L_q(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij}|^q, q \geq 1。$$

$$\hat{\beta} = (Z'Z)^{-1}Z'Y，\ q = 2$$

极大似然估计方法*

假设$(x_1, x_2, \ldots, x_n)$是随机变量X的n个独立同分布的观测样本，其真实密度函数为$g(x) = f_{\theta_0}(x)$

假设用$f_\theta(x)$估计密度函数$g(x)$，其中$\theta \in R$是未知参数。为估计未知参数，需要一个损失函数来评价用$f_\theta(x)$估计$g(x)$的效果，并通过最小化损失函数求得未知参数θ的估计。

$$KL(f_\theta, g) = \int g(x) \ln \frac{g(x)}{f_\theta(x)} dx$$

$$\int g(x) \ln g(x)\, dx \geq \int g(x) \ln f_\theta(x)\, dx$$

$$\mathrm{E}[\ln f_\theta(x)] \cong \frac{1}{n} \sum_{i=1}^{n} \ln f_\theta(x_i)$$

Logistic 回归模型

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

线性模型不适用于Y为0或1的情况。

$$U = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

$$Y = \begin{cases} 0, & U < 0 \\ 1, & U \geq 0 \end{cases}$$

根据此模型，$U < 0$ 的概率为

$$P(U < 0) = P_\varepsilon(\varepsilon + \beta_0 + \sum_{j=1}^{p} \beta_j X_j < 0)$$

类似地，可以求得$P(U \geq 0)$。

泊松回归模型

$$P(Y=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

$$\log(\lambda)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon$$

其中(x_i,k_i)是第i个样本的观测，$x_i=(x_{i1},x_{i2},\ldots,x_{ip})$。根据模型，可以得到其似然函数为

$$L(\Theta)=\prod_{i=1}^{n}\frac{\lambda(x_i)^{k_i}}{k_i!}\exp\{-\lambda(x_i)\}$$

其中$\Theta=(\beta_0,\beta_1,\ldots,\beta_p)'$，$\log\lambda(x_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}$. 极大化上述似然函数即可得到参数的极大似然估计。

Cox比例风险模型

Cox模型的基本形式为

$$h(t,x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$$

$$RR = \frac{h(t,x_i)}{h(t,x_j)} = \frac{h_0(t)\exp(\beta^T x_i)}{h_0(t)\exp(\beta^T x_j)} = \exp\left[\beta^T(x_i - x_j)\right]$$

6. 线性混合效应模型

$$y = X\beta + Zu + \varepsilon$$

其中y是$n \times 1$维的结果变量，$X \in R^{n\times p}$和$Z \in R^{n\times q}$是设计矩阵，β是p维的固定效应的参数向量，u是q维的随机效应的参数向量，ε为n维的误差项。

隐马尔科夫模型

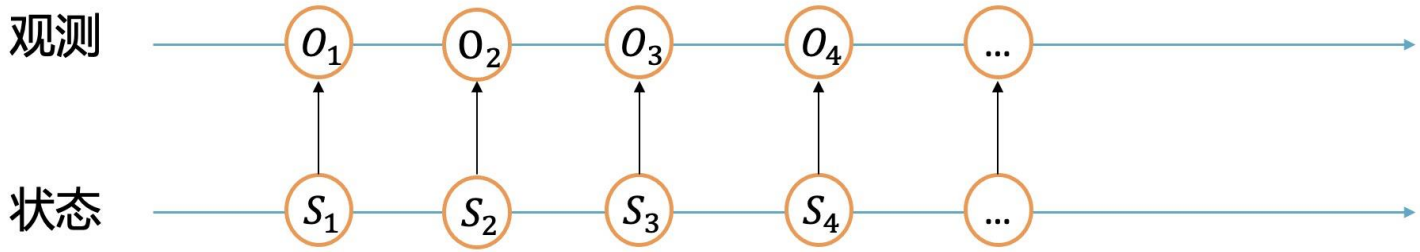
$$P(X_n = j \mid X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P_k\left(X_n = j \mid X_{n-k,\dots,} X_{n-1} = j\right)$$

随机过程$\{X_n : n = 1,2, \dots .\}$为**k阶马尔科夫过程。**

若$P_k(X_{n+1} = j \mid X_{n-k,\dots,} X_n = j)$不依赖于$k$则称该马尔科夫链为齐次马尔科夫链。

$$P = \begin{Bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{Bmatrix}$$

隐马尔科夫模型



状态转移 $$P(S_n = j \mid S_{n-1} = s_{n-1}, \dots, S_0 = s_0) = P\left(S_n = \mathrm{s_n} \mid S_{n-k,\dots,} S_{n-1} = s_{n-1}\right)$$

观测方程 $$P(Y_n = j \mid S_{n-1} = s_{n-1}, \dots, S_0 = s_0) = P(Y_n = j \mid S_n = s_n)$$

隐马尔科夫模型

三要素

- 状态转移概率矩阵$A = (a_{ij})_{N\times N}$，$a_{ij} = P(S_{t+1} = q_j | S_t = q_i)$
- 观测概率矩阵$B = (b_{jk})_{N\times M}$，$b_{jk} = P(Y_t = v_k | S_t = q_j)$.
- 初始状态概率向量$\pi = (\pi_1, \pi_2, \dots, \pi_N)$，其中$\pi_i = P(S_1 = q_i)$.

三个问题

1. 概率计算问题(评估问题): 给定模型$\lambda = (A, B, \pi)$和观测序列$O = (Y_1, Y_2, \dots, Y_T)$，计算在模型$\lambda$的观测下序列$O$出现的概率$P(O|\lambda)$. 通常使用向前算法来求解。
2. 预测问题(解码问题):已知观测序列O和模型参数λ, 求对给定观测序列下最可能的隐藏状态序列，即最大化概率$P(I|O)$. 通常使用维特比算法求解。
3. 学习问题：已知观测序列O，估计模型$\lambda = (A, B, \pi)$中的参数，使得在该模型下$P(O|\lambda)$最大化，即用极大似然估计的方法估计参数，通常使用Baum-Welch算法来解决。

第三节 统计模型

贝叶斯统计基础

贝叶斯公式

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta)$$

Thomas Bayes
(c. 1702 – April 17, 1761)

先验分布 $f(\theta)$

先验知识

似然函数 $f(x|\theta)$

汇总了数据中的信息

后验分布 $f(\theta|x)$

汇总了先验信息和数据中的信息

和频率学派的区别

1. 未知的参数是一个随机数
2. 利用数据来更新未知参数

贝叶斯统计基础

贝叶斯公式

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta)$$

$$f(\theta) \propto 1 \qquad \arg_\theta \max f(\theta|x) = \arg_\theta \max f(x|\theta)$$

$$\arg_\theta \max \frac{1}{n}\sum_{i=1}^{n} \log f(\theta|x) = \arg_\theta \max \sum_{i=1}^{n} \frac{1}{n} \log f(x|\theta) + \log f(\theta)/n$$

Maximum a posterior equals to MLE (Condition ?)

后验分布 $f(\theta|x)$

参数估计: 后验均值/中位数/极大后验

$$\hat{\theta} = E(\theta \mid x), \text{median}(\theta \mid x), \text{argmax}_\theta \, f(\theta \mid x)$$

置信区间: Credible Interval

假设检验: 贝叶斯因子 $\frac{f(x|M_1)}{f(x|M_2)}$

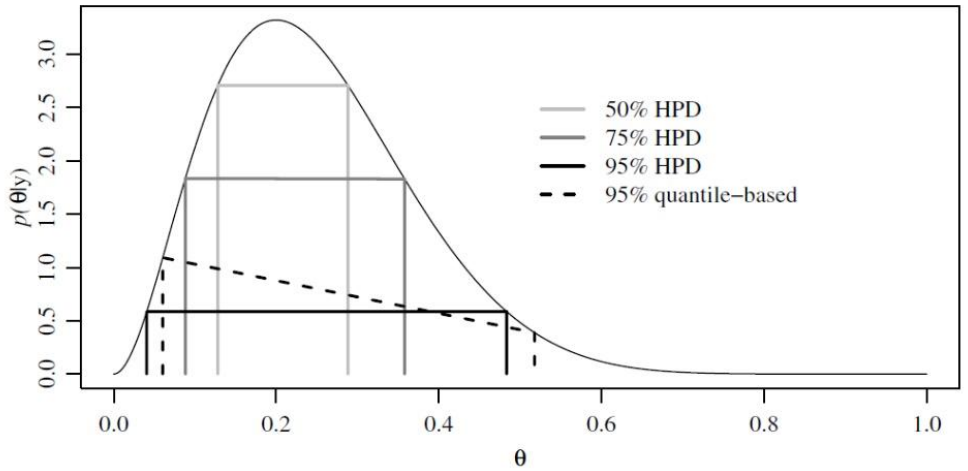


TABLE 15.1: Bayes factor scale as proposed by Jeffreys (1939).

BF_{12}	**Interpretation**
> 100	Extreme evidence for $\mathcal{M}_1$.
$30 - 100$	Very strong evidence for $\mathcal{M}_1$.
$10 - 30$	Strong evidence for $\mathcal{M}_1$.
$3 - 10$	Moderate evidence for $\mathcal{M}_1$.
$1 - 3$	Anecdotal evidence for $\mathcal{M}_1$.
1	No evidence.
$\frac{1}{1} - \frac{1}{3}$	Anecdotal evidence for $\mathcal{M}_2$.
$\frac{1}{3} - \frac{1}{10}$	Moderate evidence for $\mathcal{M}_2$.
$\frac{1}{10} - \frac{1}{30}$	Strong evidence for $\mathcal{M}_2$.
$\frac{1}{30} - \frac{1}{100}$	Very strong evidence for $\mathcal{M}_2$.
$< \frac{1}{100}$	Extreme evidence for $\mathcal{M}_2$.

第三节 统计模型

先验分布 $f(\theta)$

对系统的先验认知

先验信息的强度?

- 无信息先验 $f(\theta) \propto 1$
- 有信息先验 $f(\theta) \sim N(\mu, \sigma^2)$
- 弱信息先验

$$\int f(\theta|x)d\theta < \infty$$

分布族

- 共扼先验 $f(\theta)$ and $f(\theta|x)$ belongs to the same distribution family
- 非共扼先验

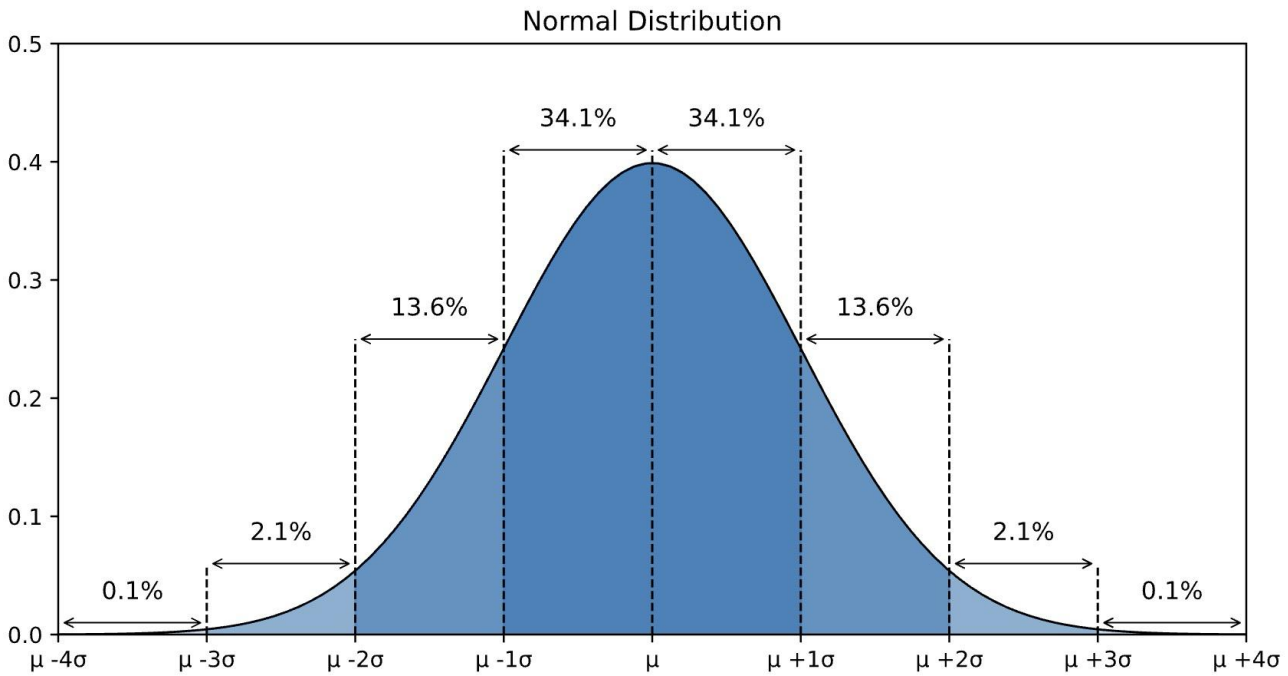
第三节 统计模型

信息量?

- 无信息先验
- 有信息先验
- 弱信息先验

$$f(\theta) \sim N(\mu, \sigma^2)$$

**方差越大
信息量越少**



$$P(\mu - 3\sigma \leq \theta \leq \mu + 3\sigma) > 99\%$$

$$\sigma \to \infty, f(\theta) \propto 1 \qquad \sigma = 0, f(\theta) = I(\theta = \mu)$$

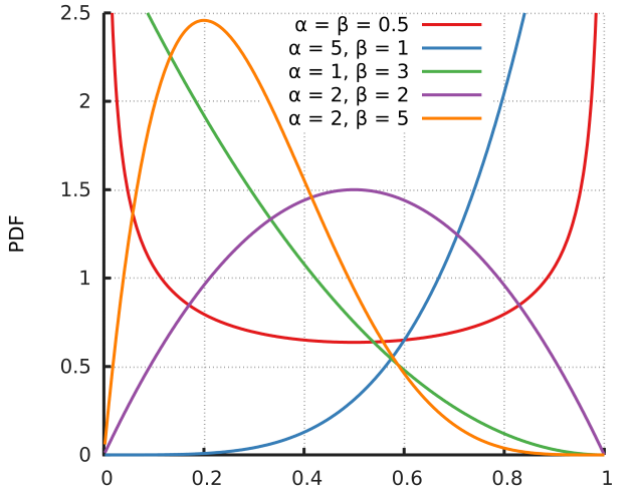
第三节 统计模型

例：小明投篮投了10次投中了0次(10次)，请估计小明投篮的命中率。

$$x_i : i = 1, 2, \cdots, n \sim B(1, p) \quad p \in [0, 1]$$

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

$$Ep = \frac{\alpha}{\alpha+\beta}, Var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^m(1-p)^{n-m}, m = \sum_{i=1}^{n} x_i$$

$$f(p|x_1, ..., x_n) \propto p^{\alpha+m-1}(1-p)^{n-m+\beta-1} \sim \text{Beta}(\alpha+m, n-m+\beta)$$

例：小明投篮投了10次投中了0次(10次)，请估计小明投篮的命中率。

$$f(p|x_1,...,x_n) \propto p^{\alpha+m-1}(1-p)^{n-m+\beta-1} \sim \text{Beta}(\alpha+m, n-m+\beta)$$

$$E(p|D) = \frac{\alpha+m}{\alpha+\beta+n} \qquad D = \{x_1, x_2, \cdots, x_n\}$$

$$\alpha = \beta = 1, n = 10$$

$$m = 0 \qquad E(p|D) = \frac{1}{12} \qquad \frac{m}{n} = 0$$

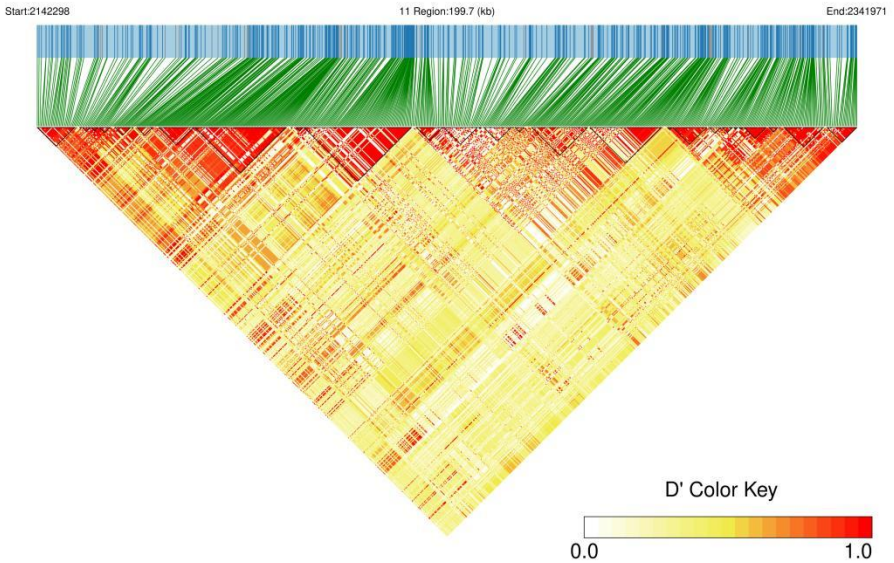
$$m = n \qquad E(p|D) = \frac{11}{12} \qquad \frac{m}{n} = 1$$

第四节 高维统计方法

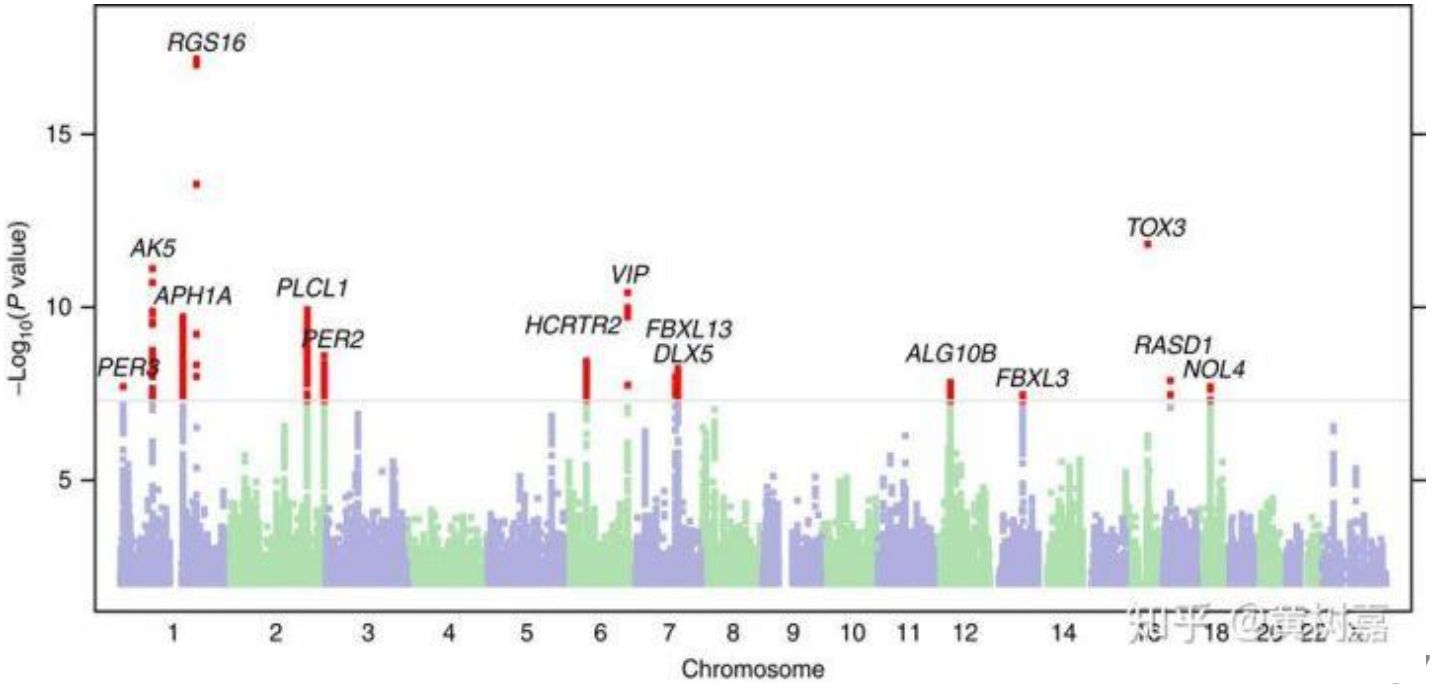
1. 正则化最小二乘
2. 变量筛选
3. 贝叶斯变量选择

GWAS

寻找与表型相关的SNP位点



$$\text{Phenotype} = \alpha_0 + \sum_{i=1}^{p} \alpha_i \times \text{SNP}_i + \eta$$



正则化最小二乘

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \Longrightarrow L(\beta|X,Y) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Z_i\beta)^2$$

$$\hat{\beta} = (Z'Z)^{-1}Z'Y, p > n\text{时}，(Z'Z)\text{不可逆} \Longrightarrow (Z'Z + \lambda I)，\ \hat{\beta} = (Z'Z + \lambda I)^{-1}Z'Y$$

1. 如何选择参数λ
2. 岭回归的实际意义是什么？
3. 岭回归的解是否具有解释性？
4. 实际问题是什么特征？

$$L(\beta|X,Y) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Z_i\beta)^2 + \lambda|\beta|^2$$

正则化最小二乘：岭回归

正则化最小二乘

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

$$L(\beta|X,Y) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Z_i\beta)^2 + \lambda|\beta|^2$$

$\hat{\beta} = (Z'Z)^{-1}Z'Y, p > n$时，$(Z'Z)$不可逆 ⟹ $(Z'Z + \lambda I)$，$\hat{\beta} = (Z'Z + \lambda I)^{-1}Z'Y$

岭回归得到的结果缺乏实际意义的可解释性

$$L(\beta|X,Y) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Z_i\beta)^2$$

未知变量个数太多无法求解

$$\|\beta\|_0 = \sum_{j=1}^{p} I(\beta_j \neq 0)$$

控制有用的变量个数

在条件 $\|\beta\|_0 \leq t$ **下，最小化** $L(\beta|X,Y) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - Z_i\beta)^2$

第四节 高维统计方法

在条件 $||\beta||_0 \le t$ **下，最小化** $L(\beta|X,Y)=\frac{1}{2n}\sum_{i=1}^{n}(Y_i-Z_i\beta)^2$

即：在实际问题中，只有部分变量有作用。

在条件 $||\beta||_1 \le t$ **下，最小化** $L(\beta|X,Y)=\frac{1}{2n}\sum_{i=1}^{n}(Y_i-Z_i\beta)^2$

著名的LASSO方法

$$Y \sim N(\mu + X\beta, \sigma^2) \qquad \pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \qquad \pi(\sigma^2) \propto 1/\sigma^2$$

Bayes LASSO

变量筛选

$\{X_i : i = 1, 2, \cdots, p\}, Y$

Step 1: 选择一种相关性度量 (DM)

Step 2: 计算相关性 $\mathrm{DM}(X_i, Y)$

Step 3: Choose X_i if $\mathrm{DM}(X_i, Y) > c$

如何选择参数c和DM?

Sure Screening Property

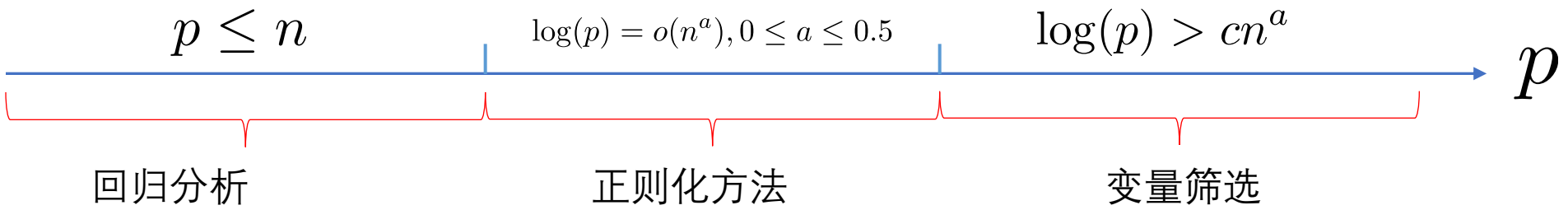
$$\min_{X_j \in M_*} |\mathrm{DM}(X_j, Y)| \geq 2wn^{-k}$$

$$P(\max_{1 \leq j \leq p} \left|\widehat{\mathrm{DM}}(X_j, Y) - \mathrm{DM}(X_j, Y)\right| \geq wn^{-k}) \to 0, n \to +\infty$$

$$M_{v_n} = \left\{1 \leq j \leq p : \left|\widehat{\mathrm{DM}}(X_j, Y)\right| \geq v_n\right\}$$

$$\text{if } v_n \leq wn^{-k}, P(M_* \subseteq M_{v_n}) \to 1$$

DM - SIS: DM是一种相关性度量方法 (Cor, dcor, scor, HHG, MIC, MI etc.)

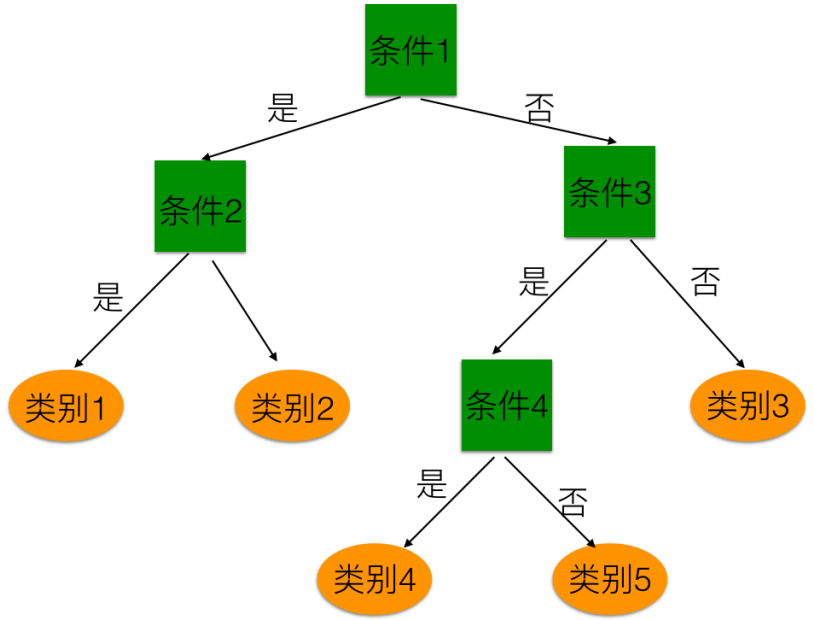


第五节 统计学习基础

1. 有监督学习
   - ❖ 回归树和决策树
   - ❖ 模型平均
   - ❖ 随机森林
   - ❖ 模型评估和模型选择
   - ❖ 支持向量机
2. 无监督学习
   - ❖ 降维(PCA,低维空间嵌入：MDS,ISOMAP,LLE,LLP等)
   - ❖ 聚类(K-means,层次聚类)

决策树

$$f(x) = \sum_k c_k I(a_{k-1} < x \leq a_k)$$



其中$a_0, c_k, a_k, k = 1,2,\ldots,$是未知参数。

模型平均(Bagging算法)

$$f(x) = \sum_{k=1}^{m} w_k f_k(x)$$

其中w_k是第k的模型的权重。

在样本$\{(x_i, y_i): i = 1, 2, \ldots, N\}$上利用Bootstrap重抽样技术得到不同的样本，并拟合一个模型$f^b(x)$。

$$f(x) = \frac{1}{B}\sum_{b=1}^{B} f^b(x).$$

随机森林=决策树+Bagging算法

随机森林算法

1. 对于$b = 1$到B

(a) 从训练集中重抽样本数为N的样本Z^*.

(b) 从重抽的样本中构建决策树T_b. 对树的每个叶节点，递归地重复下述步骤，直到达到最小的节点数。

(i)随机地从p个变量中选择m个变量。

(ii)在m个变量中选择最好的可作为分裂点的变量。

(iii)将该节点分裂为两个子节点。

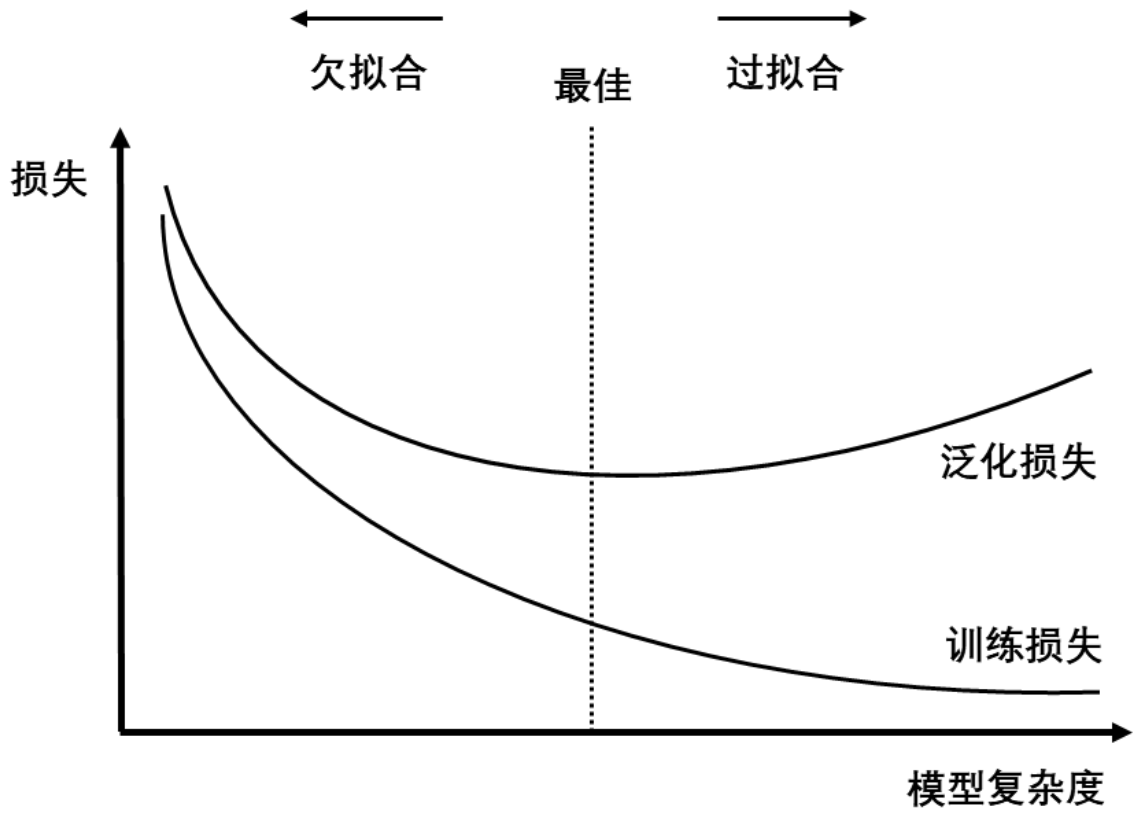
2. 输出决策树$\{T_b\}_{b=1}^{B}$所构成的随机森林。

对于新的样本点x做预测时：

回归问题：$\hat{f}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

分类问题：令$\hat{C}_b(x)$为第b个决策树的预测，则随机森林的预测结果由B个决策树通过多数表决决定。

模型评估和模型选择



K折验证(K fold Cross-validation)：
将训练集分为K份大小相同的数据集，选取其中K-1份作为训练集，将剩下的一份作为验证集，估计误差。如此重复K次，用K份验证集上的平均误差或方差等信息来估计模型的泛化能力，进而进行模型选择

支持向量机

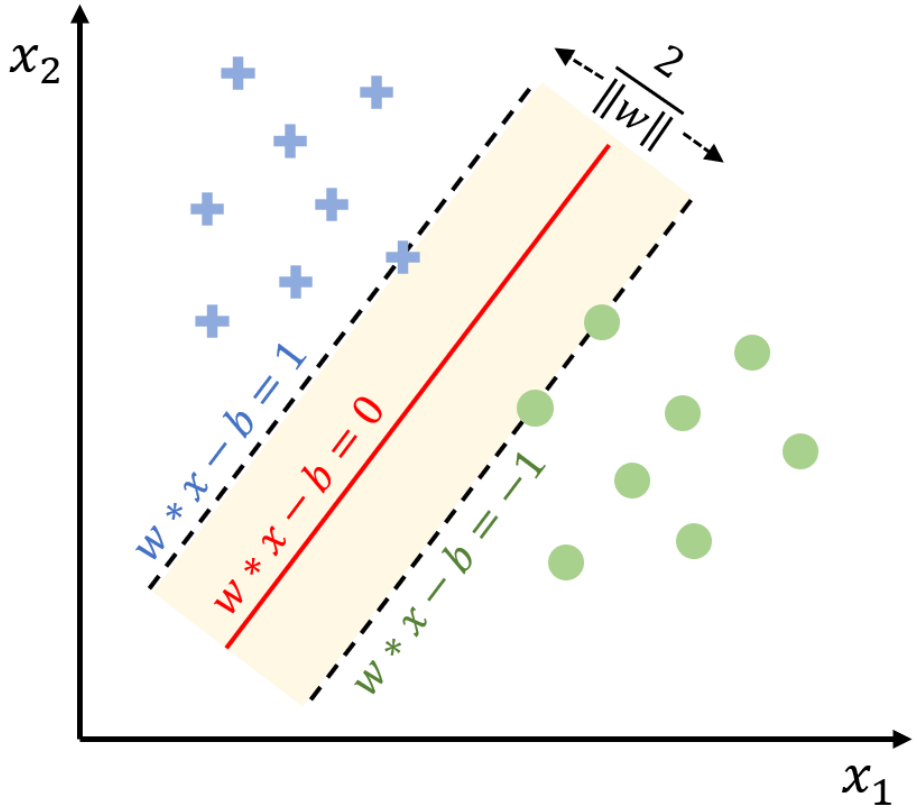
给定数据集$\{(X_i, y_i), i = 1, 2, \ldots, n\}$，其中$X_i \in R^d, y_i \in \{-1, 1\}$。

分类器是超平面$w \cdot x - b$=0，若$w \cdot x_i - b > 0$，则$y_i = 1$，否则$y_i = -1$。

$$y_i = \begin{cases} 1, & w \cdot x_i - b \geq \eta \\ -1, & w \cdot x_i - b \leq -\eta \end{cases}$$

$$\max_{w,b} \frac{2\eta}{|w|^2}$$

$$s.t. \quad y_i(wx_i - b) \geq \eta$$



无监督学习: 降维

假设数据集 $\{X_i \in R^m, i = 1, 2, \ldots, n\}$ 是n个细胞中m个基因的表达数据。

希望构造新的变量Z来表示原数据,而且还满足以下性质:
1）Z的维度更低，最好是2维，或者三维;
2）Z保留了原始数据的信息。

PCA

$$Z = X'\beta$$

$$\max Var(X'\beta) = \beta'\Sigma\beta, s.t. \ |\beta| = 1,$$

$$p_k = \frac{\Sigma_{i \le k} \lambda_i}{\Sigma_{i \le m} \lambda_i}$$

低维空间嵌入

假设$X = (X_1, \ldots, X_n)$为高维空间R^p中的n个观测点，$Y = (Y_1, \ldots, Y_n)$是它在低维空间R^k中的嵌入。

$d_{ij} = d(X_i, X_j)$是原空间中任意两个点之间的距离，$d_{ij}' = d(Y_i, Y_j)$为低维空间中任意两点之间的距离

低维空间嵌入就是要在一个低维空间中重构原数据之间的相对位置关系，即：$d_{ij} = d_{ij}'$。

- 高维尺度分析（Multidimensional Scaling，MDS）

$$L(Y) = \sum \left(d'_{ij} - d_{ij} \right)^2$$

- 等距特征映射（ISOMAP）

- 局部线性嵌入（Local linear embedding）

$$L(Y) = \sum (Y_i - \sum_{j \in R_i} w_{ij} Y_j)^2,$$

其中，R_i是X_i的k阶邻居，即：$R_i = \{j \mid d(X_j, X_i) \leq d_k\}$, 权重$w_{ij}$表示用$X_i$的邻居预测$X_i$的回归系数。

低维空间嵌入

假设$X = (X_1, \ldots, X_n)$为高维空间R^p中的n个观测点，$Y = (Y_1, \ldots, Y_n)$是它在低维空间R^k中的嵌入。

$d_{ij} = d(X_i, X_j)$是原空间中任意两个点之间的距离，$d_{ij}' = d(Y_i, Y_j)$为低维空间中任意两点之间的距离

低维空间嵌入就是要在一个低维空间中重构原数据之间的相对位置关系，即：$d_{ij} = d_{ij}'$。

- 局部线性映射（Local linear projection）

$$L(Y) = \sum (Y_i - Y_j)^2 w_{ij},$$

其中，w_{ij}是基于X_i, X_j得到的权重。低维空间中的两个点离得越远，权重就越大。

- t分布式随机邻居嵌入（t-distributed Stochastic Neighbor Embedding，t-SNE）和统一流形逼近与投影(Uniform Manifold Approximation and Projection，UMAP)则考虑保持低维空间中所有点之间的位置分布关系不变。

聚类分析

聚类分析是分析样本之间的相似性，把相似的样本聚集到一起，不相似的样本则尽可能分开。

方法	$\boldsymbol{a}=(a_1,a_2,\dots a_p),\boldsymbol{b}=(b_1,b_2,\dots,b_p),p\geq 1$ 之间的相似性
欧几里得距离	$\mathrm{d}(\mathbf{a},\mathbf{b})=\sqrt{\sum(a_i-b_i)^2}$
切比雪夫距离	$\mathrm{d}(\mathbf{a},\mathbf{b})=\max_{1\leq i\leq p}\lvert a_i-b_i\rvert$
曼哈顿距离	$\mathrm{d}(\mathbf{a},\mathbf{b})=\sum\lvert a_i-b_i\rvert$
堪培拉距离	$\mathrm{d}(\mathbf{a},\mathbf{b})=\sum\frac{\lvert a_i-b_i\rvert}{a_i+b_i}$
闵可夫斯基距离	$\mathrm{d}(\mathbf{a},\mathbf{b})=\sqrt[q]{\sum(a_i-b_i)^q},q\geq 1$
Cosine距离	$Cos(\mathbf{a},\mathbf{b})=\frac{\sum a_ib_i}{\sum a_i^2\sum b_i^2}$
Pearson 相关系数	$\rho(a,b)=\frac{\sum(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{Var(a)Var(b)}},\bar{a}=\frac{1}{p}\sum a_i$
Spearman 相关系数	$\rho(R_a,R_b)$, R_a,R_b 分别表示a 和b的秩向量。

K-means 算法

假设$\{x_1, x_2, \ldots, x_n\}$是p维随机变量$X \in R^p$的n个独立观测，那么聚类算法本质上是从样本点到类别的一个映射：$C: X \to k, k \in \{1, 2, \ldots, K\}$，即：$C(x_i) = k$，简写为$C(i) = k$，其中$i = 1, 2, \ldots, n$. 换句话说，聚类问题就是要求解未知的函数C。

1. 类内距离，即：属于相同类别的两个点之间的距离的总和

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(x_{i'}, x_{i'})$$

2. 类间距，即：属于不同类别的两个点之间的距离的总和

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d(x_i, x_{i'})$$

3. 总间距，即：所有点之间的距离之和

$$T = \frac{1}{2}\sum_{i}\sum_{i'} d(x_{i'}, x_{i'}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\left(\sum_{C(i')=k} d(x_i, x_{i'}) + \sum_{C(i')\neq k} d(x_i, x_{i'})\right)$$

K-means聚类算法

1. 对于数据点集D，聚类数K，初始化K个聚类中心，通常是从D中随机选择。
2. 重复(a)-(b)直到聚类中心点不再变化或达到最大迭代次数。

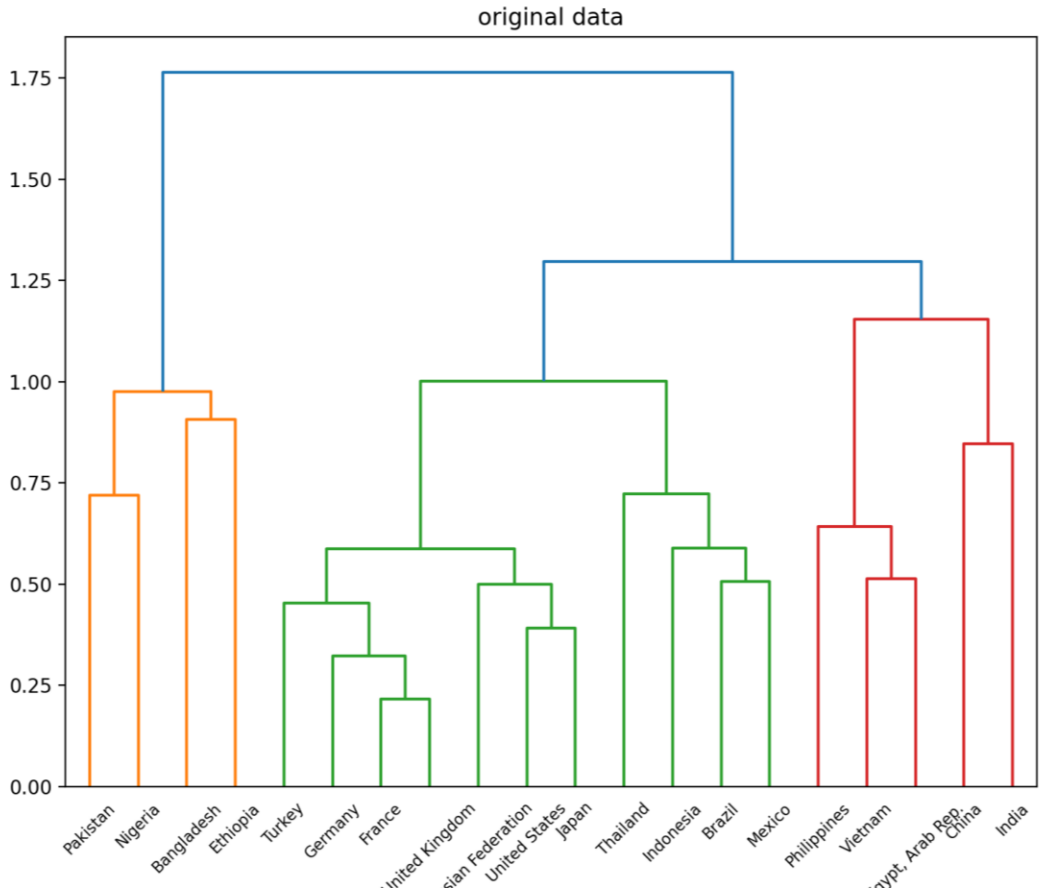
(a) 对于D中的每一个数据点x_i，为其分配最近的聚类中心。

(b) 对每一个聚类中心$C_j, j = 1, \ldots, K$，更新C_j的中心为C_j中所有点的均值。

层次聚类算法

层次聚类(Hierarchical Clustering)是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。

聚类方式	说明
最短距离法	一个簇和另外一个簇中的不同向量互相之间的距离最小值为两个簇的距离
最大距离法	一个簇和另外一个簇中的不同向量互相之间的距离最大值为两个簇的距离
平均距离/UPGMA法 (Unweighted Pair Group Method using arithmetic Averages)	一个簇和另外一个簇中不同向量两两之间距离的平均值为两个簇的距离
WPGMA法（Weighted Pair Group Method using arithmetic Averages）	一个簇和另外一个簇中不同向量两两之间距离的加权平均值为两个簇的距离
UPGMC法（Unweighted Pair Group Method using Centroids）	两个簇的质心之间的距离为两个簇的距离
WPGMC法（weighted Pair Group Method using Centroids）	两个簇的加权质心之间的距离为两个簇的距离

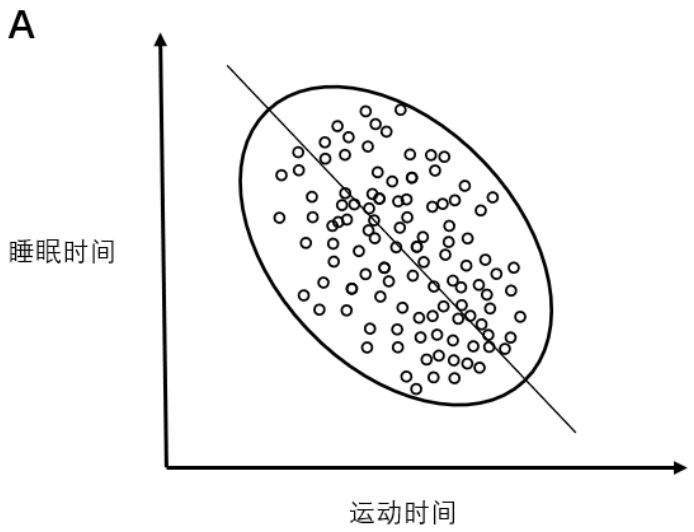


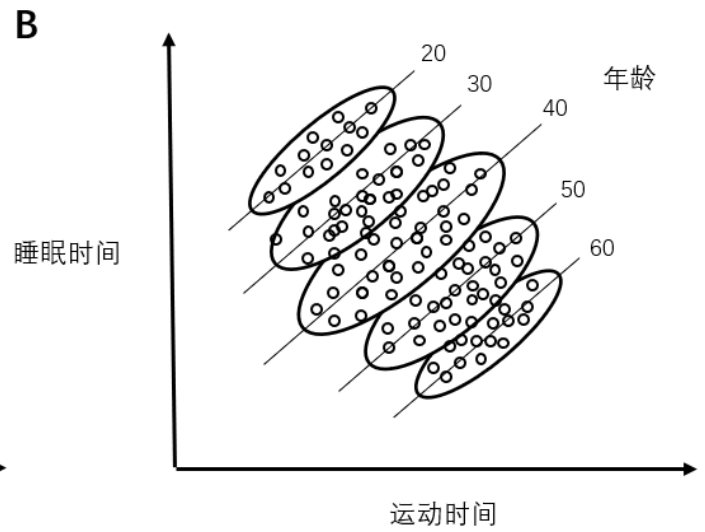
第六节 统计因果推断

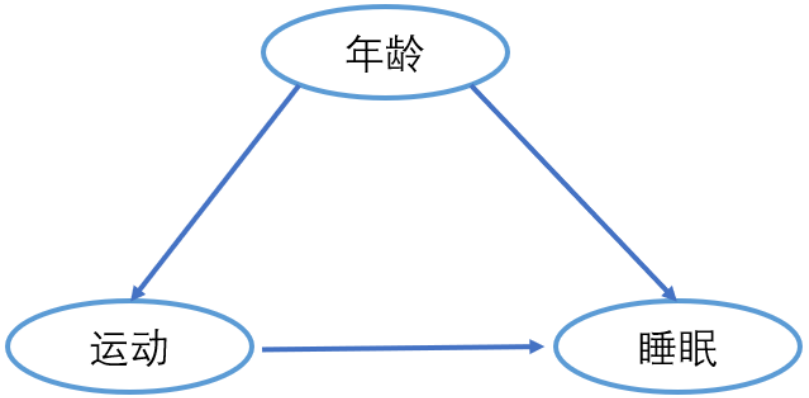
1. 辛普森悖论
2. 关联与因果
3. 因果效应的定义与识别
4. 因果效应的估计
5. 工具变量

辛普森悖论

两个变量X和Y的边缘相关性和给定Z后的条件相关性可能是完全相反的







关联与因果

由于混杂因素的干扰，关联并不等于因果

因果效应的定义与识别

潜在结果：$\boldsymbol{Y(1)}$、$\boldsymbol{Y(0)}$

Individuals	T	Y(1)	Y(0)
James	1	12	*
Sarah	1	10	*
Emily	1	9	*
Michael	0	*	13
David	0	*	16
Elizabeth	0	*	14

平均因果效应（Average Causal Effect，ACE）

$$\text{ACE} \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \mathbb{E}[\text{Y}(1)] - \mathbb{E}[\text{Y}(0)]$$

第六节 统计因果推断

由于反事实结果不可观测，ACE面临着不可识别的问题，我们通过引入三个假设来使得ACE可识别。

（1）可忽略性假设：干预***T***与潜在结果相互独立,即：试验是随机的。

$$\big(Y(1),Y(0)\big) \perp T$$

（2）正性假设：干预组与对照组都有个体参与

$$0 < P(T = 1|X = x) < 1$$

（3）个体稳定性假设：给定的个体的干预结果不会受到其他个体是否接受干预的影响

$$\mathbb{E}[\mathrm{Y}(1)|T = 1, X] = \mathbb{E}[\mathrm{Y}|T = 1, X]$$

因果效应的估计

逆倾向得分加权

$$\widehat{ACE}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i T_i}{\hat{e}(X_i)} + \frac{Y_i(1-T_i)}{1-\hat{e}(X_i)}\right)$$

结果回归

$$\mathbb{E}[\mathrm{Y}|T,X] = \beta_0 + \beta_t T + \beta_x X$$

$$\widehat{ACE}_{OR} = \hat{\beta}_t$$

双稳健估计

$$ACE = \mathbb{E}\left[\frac{YT}{e(X)} - \frac{T-e(X)}{e(X)}m_1(X)\right] - \mathbb{E}\left[\frac{Y(1-T)}{1-e(X)} + \frac{T-e(X)}{1-e(X)}m_0(X)\right]$$

其中$e(X)$为IPW中的倾向分数，$m_1(X)$和$m_0(X)$分别为结果回归中的$\mathbb{E}[\mathrm{Y}|T=1,X]$和$\mathbb{E}[\mathrm{Y}|T=0,X]$。

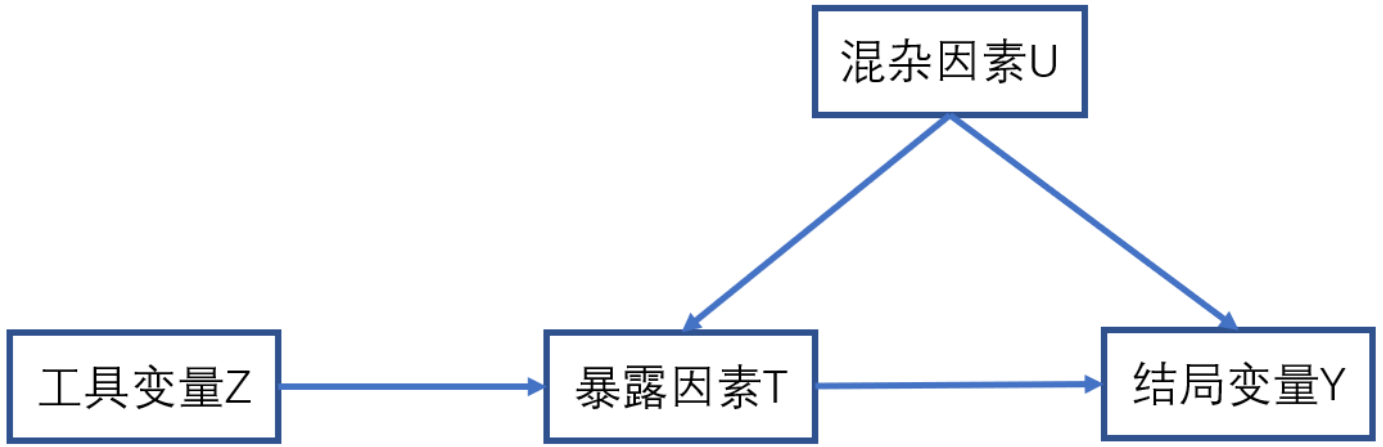
工具变量

工具变量需要满足以下三个条件

a）相关性：工具变量Z与暴露或干预T具有显著的统计关联，即$Z \perp /T$。

b）排它性限制：在给定暴露T和其他混杂U的条件下，工具变量Z对结果变量Y没有直接作用，即$Y \perp Z \mid T, U$。

c）无混杂性：工具变量Z与混杂因素 U相互独立，即$Z \perp U$。



工具变量的识别与估计

工具变量法在不做任何模型假设的情况下因果效应是无法识别，下面我们在线性模型假设下，即变量之间为线性关系，讨论工具变量的识别问题。

二分类工具变量下ACE的识别

$$\alpha = \frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(T|Z=1) - \mathbb{E}(T|Z=0)}$$

连续型工具变量下ACE的识别

$$\alpha = \frac{Cov(Y,Z)}{Cov(T,Z)}$$

两阶段最小二乘估计

第一阶段为使用暴露变量T对工具变量Z进行回归：

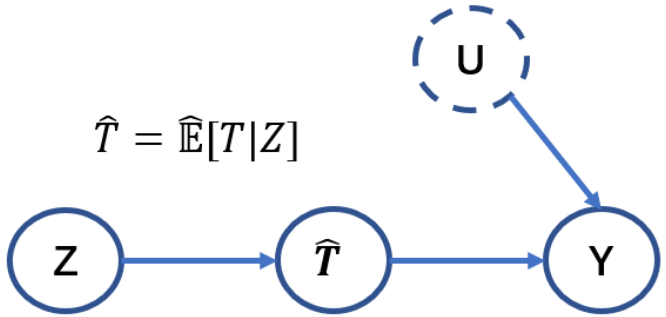
$$T = \gamma Z + \varepsilon_T$$

得到预测的暴露变量$\hat{T} = \hat{\gamma} Z$

第二阶段，研究者使用结果变量Y对第一阶段预测的$\hat{T}$进行回归，

$$\mathrm{Y} = \alpha \hat{T} + \beta \mathrm{U} + \varepsilon_Y$$

求解得到因果效应的估计值 $\hat{\alpha}$ 。这两步过程将消除混杂因素的影响，提供 T和Y 之间因果效应的无偏估计。



小结

统计量

基于生物学知识: 用于量化生物学概念
比如：基因的表达量PKFM, 转录因子的调控强度等， 甲基化水平

基于统计学习/机器学习模型或者算法
比如: LASSO, 线性回归， logistic回归

用于刻画因素之间的关系

基于统计学基本原理
比如：卡方统计量， Fisher检验方法, Wilcox符号秩检验

小结

统计学与人工智能的区别

1. 统计学主要用于结构化数据，人工智能则能够很好的处理非结构化数据
2. 统计学得到的结果具有很好的可解释性，人工智能方法有所欠缺
3. 人工智能可以自动提取特征(但并不一定是好事)，但是统计学需要根据先验知识定义特征
4. 很多好的统计思想可以与人工智能相结合，提升人工智能的计算效率、可解释性等

共享约束需知：使用条款

1. **版权声明**：本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求**：任何使用本PPT的行为均须遵守以下条件：
   1) **致谢和标注**：若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
   2) **修改和再分发**：未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用**：严禁将本PPT用于任何形式的商业化用途，包括但不限于：
   1) 通过网络或其他途径进行付费使用或分发；
   2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任**：任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
   1) 要求停止侵权行为；
   2) 追究侵权使用者的经济赔偿责任。
5. **其他规定**：
   1) 本使用条款的解释权归本PPT作者所有。
   2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。