内 容 简 介

本书记录了作者在山大数院的三年所学的数学知识，简明扼要地列举出每个学科所必须掌握的定理等重要知识点。此书不宜作为初学某一科目的参考资料，而适合已学完部分内容者查漏补缺。

根据山大本科开课顺序以及个人自学进度，大致收录下列学科：

大一上：数学分析 1，高等代数 1，解析几何

大一下：数学分析 2，高等代数 2

大二上：数学分析 3，复变函数，常微分方程

大二下：实变函数，偏微分方程，概率论

大三上：机器学习、数字图像处理、数理统计、矩阵论

大三下：时间序列分析、数据库系统、数据结构

目　　录

第一章　概率论与数理统计

Mathematical Analysis

概率论与数理统计是由数分高代派生出来的应用学科, 用于刻画日常生活中随机发生的事件, 具有很高的应用价值. 其中, 概率论主要研究随机变量的分布与特征, 而数理统计主要研究通过样本对未知分布进行估计.

概率论的重点: 概率的定义, 条件概率与独立性, 一元或多元随机变量分布, 常用分布函数, 随机变量的特征数, 大数定律和中心极限定理

数理统计的重点: 基本概念与三大分布, 参数估计, 假设检验, 方差分析, 回归分析

1.1 随机事件与概率

之前数学分析研究的内容都是具有确定解析式或约束条件的函数, 但概率论引进了随机因素, 即实验和结果并不是一一对应的, 一次实验可能会出多种结果. 这一部分的任务是使用概率这一量化方式, 将随机性规范化.

1.1.1 随机事件

1. 随机现象: 重复实验会出现不同结果的现象.

2. 样本空间: 随机现象可能出现的结果组成的集合.

3. 随机事件: 样本空间的子集. 当实验结果属于此子集时, 称随机事件发生.

4. 随机变量: 用于描述随机事件的人为设定变量 (非正式定义).

5. 事件的运算: 和集合一致, 有交并补余四大运算. 有两个公式很重要.

(1) 集合减法公式: $A - B = A \cap \bar{B}$.

(2) 德摩根律: $\overline{A \cup B} = \bar{A} \cap \bar{B}$; $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

6. 事件域: 令 Ω 为样本空间, 定义事件域 $\mathscr{F}$ 符合下列性质:

(1) $\Omega \in \mathscr{F}$; (2)$A \in \mathscr{F} \Rightarrow \bar{A} \in \mathscr{F}$; (3)$A_n \in \mathscr{F} \Rightarrow \bigcup\limits_{n=1}^{\infty} A_n \in \mathscr{F}$.

1.1.2 概率

1. 公理化定义: 在事件域 $(\Omega, \mathscr{F})$ 上定义可测函数 $P(A)$ 满足:

(1) 非负性: $P(A) \geq 0$; (2) 正则性: $P(\Omega) = 1$;

(3) 可列可加性: 事件 $A_1, \cdots, A_n$ 互不相容时, $P(\bigcup\limits_{n=1}^{\infty} A_n) = \sum\limits_{n=1}^{\infty} P(A_i)$.

2. 用频率定义概率: 令 $n(A)$ 为事件 A 发生的频数, 则可用大量重复事件的频率表示概率: $P(A) = \lim\limits_{n\to\infty} \frac{n(A)}{n}$.

3. 古典概型: 若样本空间有 n 个等可能发生的样本点, 则事件 A 包含 k 个样本点时, $P(A)=\frac{k}{n}$.

4. 几何概型: 若样本空间 Ω 的面积测度为 S_n, 事件 A 包含其中面积为 S_A 的一部分, 则 $P(A)=\frac{S_A}{S_n}$. (蒙特卡罗法的理论依据)

5. 贝叶斯概率: 对事件发生可能性的主观预测, 在机器学习中使用频率很高.

1.1.3 概率的性质

1. 有限可加性: 若 $A_1,\cdots,A_n$ 互不相容, 则 $P(\bigcup\limits_{i=1}^{n}A_i)=\sum\limits_{i=1}^{n}P(A_i)$.

2. 单调性: 若 $A\subset B$, 则 $P(A)\leq P(B)$.

3. 加法公式: $P(A\bigcup B)=P(A)+P(B)-P(AB)$.

1.1.4 条件概率

1. 定义: $P(A|B)$ 表示已知 B 发生的条件下 A 发生的概率. $P(A|B)=\frac{P(AB)}{P(B)}$.

2. 乘法公式: $P(AB)=P(B)P(A|B)$, 即定义式的变种.

3. 全概率公式: 若 B_i 互不相容, 且 $\bigcup\limits_{i=1^n}B_i=\Omega$, 则

$$P(A)=\sum_{i=1}^{n}P(A|B_i)P(B_i)$$

4. 贝叶斯公式: 用先验概率推后验概率. 若 B_i 互不相容, 且 $\bigcup\limits_{i=1^n}B_i=\Omega$, 则

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum\limits_{k=1}^{n}P(A|B_k)P(B_k)}$$

5. 独立性: 若 $P(A|B) = P(A)$, 即 $P(AB) = P(A)P(B)$, 则称事件 A 和 B 相互独立.

1.2 随机变量及其分布

用概率描述随机事件发生可能性的大小后, 为了更充分认识随机事件, 我们引入随机变量来刻画随机事件, 如抽奖是随机事件, 在此基础上可以定义随机变量" 是否中奖", 这是一个二值随机变量 (0/1).

使用随机变量来描述随机事件, 能更方便地研究随机事件中我们感兴趣的性质, 比如随机变量" 灯泡坏掉的个数" 能帮助我们衡量灯泡的寿命. 这些随机变量取值的规律可以用分布来描述, 离散随机变量和连续随机变量的刻画方式略有区别.

1.2.1 随机变量

1. 定义: 样本空间 Ω 上的实值函数 $X(\omega)$.

2. 离散随机变量的确定: 使用分布列描述.

X	X_1	X_2	$\cdots$	X_n
P	p_1	p_2	$\cdots$	p_n

其中 p_i 表示随机变量 X 取值 X_i 的概率, $\sum_{i=1}^{n} p_i = 1$.

3. 连续随机变量的描述: 使用分布函数与概率密度函数.

(1) 分布函数 $F(x)$: $F(x) = P(X \leq x)$, 是单调递增的右连续函数, 且 $F(+\infty) = 1, F(-\infty) = 0$.

(2) p.d.f 概率密度函数 $p(x)$: $p(x) = F'(x)$, 是非负函数且 $\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$.

1.2.2 常用分布及概率密度函数

1. 离散分布

(1) 泊松分布: $P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$, 用于计数过程, 记作 $X\sim P(\lambda)$.

(2) 伯努利分布: $P(X=1)=p, P(X=0)=1-p$, 又称两点分布.

(3) 二项分布: $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$, 即 n 重伯努利分布中事件发生的次数, 记作 $X\sim b(n,p)$.

(4) 几何分布: $P(X=k)=(1-p)^{k-1}p$, 具有无记忆性.

2. 连续分布

(1) 正态分布: $p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{1\frac{(x-\mu)^2}{2\sigma^2}}$, 是最常用的分布. 记作 $X\sim N(\mu,\sigma^2)$, 标准正态分布即 $N(0,1)$.

(2) 均匀分布: $p(x)=\frac{1}{b-a}$, 其中 $x\in(a,b)$, 记作 $X\sim U(a,b)$.

(3) 指数分布: $p(x)=\lambda e^{-\lambda x}$, 其中 $x\geq 0$, 记作 $X\sim\epsilon(\lambda)$, 具有无记忆性.

(4) 伽马分布: $p(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$, 其中 $x\geq 0$, 记作 $X\sim Ga(\alpha,\lambda)$. 特殊地, $Ga(\frac{n}{2},\frac{1}{2})=\chi^2(n)$ 为卡方分布, 统计中常用.

1.2.3 数字特征

1. 数学期望: X 在不同取值数按概率的加权平均数, 是消除随机性的主要手段, 记作 Ex. 在离散场合, $EX=\sum_{n=1}^{\infty}p_ix_i$. 在连续场合, $EX=\int_{-\infty}^{+\infty}xp(x)\mathrm{d}x$.

2. 方差: $DX=E[(X-EX)^2]$, 也记作 $\mathrm{Var}(X)$, 用于衡量数据的集中程度. 常用的计算公式为

$$DX=E(X^2)-(EX)^2$$

3. 标准差: $\sigma(x)=\sqrt{DX}$, 也记作 $Std(X)$, 好处是与 X 的量纲一致.

4. 切比雪夫不等式: $P(|X-EX|\geq\varepsilon)\leq\frac{DX}{\varepsilon^2}$.

1.2.4 随机变量函数的分布

1. 离散情形: 先求各项的像 $g(x_1), \cdots, g(x_n)$, $g(x_i$ 对应概率仍为 p_i, 再合并相同项.

2. 连续情形: 若 $Y = g(x)$ 严格单调, 反函数为 $x = h(y)$, X 的概率密度函数为 $p(x)$, 则 Y 的概率密度函数为 $p_Y(y) = p_X(h(y)) \cdot |h'(y)|$. 一般情况下, 需要根据 $P(g(x) \leq y)$ 反解出 x 的范围, 再利用 X 的分布函数求解.

1.3 多元随机变量及其分布

若样本点含有不止一个我们感兴趣的属性, 如身体指标包含身高和体重, 则可定义多元随机变量来刻画这些指标的分布. 研究多元随机变量, 除了明确各分量的分布外, 还需要研究各分量间的相关关系, 以及给定某条件后的分布情况.

事实上, 只要给定多元随机变量的联合分布, 就能得到所有信息, 该部分的目的就是掌握将信息从联合分布中提取出来的方法.

1.3.1 多元随机变量

1. 定义: 样本空间 Ω 上的向量值函数 $X(\omega) = (X_1(\omega), \cdots, X_n(\omega))$.

2. 联合分布函数: $F(x_1, \cdots, x_n) = P(X_1 \leq x_1, \cdots, X_n \leq x_n)$

3. 离散情形的联合分布列: 仅用于二元分布 (X, Y), 用 i 行 j 列元素 p_{ij} 表示 $X = X_i, Y = Y_j$ 的概率, 其中 $\sum\limits_{i,j} p_{ij} = 1$.

4. 连续情形的联合密度函数: $p(x_1, \cdots, x_n) = \frac{\partial^n F(x_1, \cdots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}$.

5. 多元正态分布: 最重要的多元连续分布. 令 $x = (x_1, \cdots, x_n)$, 均值向量为

μ, 协方差矩阵为 Γ, 则 n 元正态分布的联合密度函数

$$p(x_1,\cdots,x_n)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Gamma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)^T\Gamma^{-1}(x-\mu)}$$

特殊地, 当 $n=2$ 时, 二元正态分布为

$$p(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+(\frac{y-\mu_2}{\sigma_2})^2]}$$

记作 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.

1.3.2 边缘分布

1. 边缘分布函数: $F_x(x)=F(x,\infty)$, $F_y(y)=F(\infty,y)$.

2. 离散情形的边缘分布列: $P(X=X_i)=\sum\limits_j P(X=X_i,Y=Y_j)$; $P(Y=Y_j)=\sum\limits_i P(X=X_i,Y=Y_j)$

3. 连续情形的边缘密度函数: $p_X(x)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}y$; $p_Y(y)=\int_{-\infty}^{\infty}p(x,y)\mathrm{d}x$.

4. 随机变量的独立性: 若 $\prod\limits_{i=1}^{n}p_i(X_i)=p(x_1,\cdots,x_n)$, 即联合密度函数为边缘密度函数之积, 则称 $X_1,\cdots,X_n$ 相互独立.

1.3.3 多元随机变量函数的分布

1. $Z=X+Y$ 的分布: 可用后面提到的特征函数法, 也可用卷积公式, 即 $p_Z(z)=\int_{-\infty}^{\infty}p_X(x)p_Y(z-x)\mathrm{d}x$.

2. 次序统计量分布: 若 $X_{(1)},\cdots,X_{(n)}$ 独立同分布且升序排列, 则第 k 个次

序统计量 $X_{(k)}$ 的概率密度函数为

$$p_{(k)}(x)=\frac{n!}{(k-1)!(n-k)!}[F(x)]^{k-1}p(x)[1-F(x)]^{n-k}$$

特殊地, $\min X$ 即 $X_{(1)}$ 的概率密度函数为 $n[1-F(x)]^{n-1}p(x)$;

$\max X$ 即 $X_{(n)}$ 的概率密度函数为 $n[F(x)]^{n-1}p(x)$.

3. 变量变换法: 令 $u=u(x,y),v=v(x,y)$, 从中反解出 $x=x(u,v),y=y(u,v)$, 则 $p(u,v)=p(x,y)\left|\frac{\partial(x,y)}{\partial(u,v)}\right|$.

1.3.4 多元随机变量的特征数

1. 数学期望: $g(x,y)$ 的期望为 $\int_R g(x,y)p(x,y)\mathrm{d}x\mathrm{d}y$.

2. 方差: 定义不变, 仍有 $\mathrm{Var}(x)=E(X-EX),\mathrm{Var}(y)=E(Y-EY)$.

3. 协方差: $\mathrm{Cov}(X,Y)=E[(X-EX)(Y-EY)]=E(XY)-EXEY$, 用于刻画两变量的相关程度.

4. 相关系数: $\mathrm{Corr}(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$. 当 $\mathrm{Corr}(x)\in(0,1]$ 时, 称 X 和 Y 正相关; $\mathrm{Corr}(x)\in[-1,0)$ 时, 称 X 和 Y 负相关; $\mathrm{Corr}(x)=0$ 时, 称 X 和 Y 不相关.

5. 方差运算性质: $\mathrm{Var}(X\pm Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)\pm 2\mathrm{Cov}(X,Y)$

6. n 元随机变量的协方差矩阵:

$$\Gamma=\begin{pmatrix}\mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \cdots & \mathrm{Cov}(X_2,X_n)\\ \vdots & \vdots & \vdots & \vdots\\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Cov}(X_n,X_n)\end{pmatrix}$$

用于刻画各分量之间的总体相关性.

1.3.5 条件分布与条件期望

1. 离散条件分布: $P_{i|j} = P(X = X_i|Y = y_j) = \frac{p_i j}{\sum\limits_i p_i j}$.

2. 连续条件分布: $p(y|x) = \frac{p(x,y)}{p_X(x)}$; $p(x|y) = \frac{p(x,y)}{p_Y(y)}$

3. 全概率公式: $p_Y(y) = \int_{-\infty}^{\infty} p_X(x)p(y|x)\mathrm{d}x$.

4. 贝叶斯公式:

$$p(x|y) = \frac{p(y|x)p_X(x)}{\int_{-\infty}^{\infty} p(y|x)p_X(x)\mathrm{d}x}$$

5. 条件数学期望: $E(X|Y = y) = \int_{-\infty}^{\infty} xp(x|y)\mathrm{d}x$.

6. 重期望公式: $E[E(X|Y)] = EX$.

1.4 大数定律和中心极限定理

这部分首先将傅里叶变换引入概率密度函数的求解中, 得到特征函数这个很好用的工具, 再借助特征函数推导大数定律和中心极限定理的一般结论, 为数理统计的展开做好铺垫.

大数定律的内容很简单, 就是抽样次数足够大时, 频率近似于概率, 均值近似于数学期望, 这给大样本统计提供了理论依据. 中心极限定理说明多个独立同分布随机变量之和近似于正态分布, 这鼓励我们在大样本统计中使用正态分布进行统计推断.

1.4.1 随机变量的特征函数

1. 定义: $\varphi(t) = E(e^{itX})$. 离散情形下, $\varphi(t) = \sum\limits_{k=1}^{\infty} p_k e^{itX_k}$; 连续情形下, $\varphi(t) = \int_{-\infty}^{\infty} p(x)e^{itx}\mathrm{d}x$.

2. 性质: (1) 若 X 与 Y 独立, $Z = X + Y$, 则 $\varphi_Z(t) = \varphi_X(t) \cdot \varphi_Y(t)$.

(2) 求各阶矩的方式: $\varphi^{(k)}(0) = i^k E(X^k)$.

(3) 唯一性定理: 分布函数由特征函数唯一确定.

3. 逆转公式: $F(x_2) - F(x_1) = \lim\limits_{T\to+\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itx_1} - e^{-itx_2}}{it} \varphi(t)\mathrm{d}t$.

4. 连续随机变量的逆变换公式: $p(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi(t)\mathrm{d}t$.

1.4.2 大数定律

1. 一般形式: 对任意 $\varepsilon > 0$, 有

$$\lim_{n\to+\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} EX_i\right| < \varepsilon\right) = 1$$

2. 伯努利大数定律: 令 S_n 为 n 重伯努利试验中事件发生的次数, p 为事件发生的概率, 则对任意 $\varepsilon > 0$, 有

$$\lim_{n\to+\infty} P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) = 1$$

3. 切比雪夫大数定律: 当 $\{X_n\}$ 两两不相关且 $Var(X_i)$ 有界时, 大数定律成立.

4. 辛钦大数定律: 当 $X_1, \cdots, X_n$ 独立同分布且 EX_i 存在时, 大数定律成立.

5. 辛钦大数定律的证明: 令 $\varphi(t)$ 为 X_i 共同的特征函数, 数学期望为 a, 将 $\varphi(t)$ 在 $t = 0$ 处泰勒展开: $\varphi(t) = 1 + iat + o(t)$. 故 $\varphi_{\frac{1}{n}\sum\limits_{i=1}^{n} X_i}(t) = \left[\varphi\left(\frac{t}{n}\right)\right]^n \sim e^{iat}$, 恰是退化分布的特征函数.

1.4.3 中心极限定理

1. 林德伯格-莱维中心极限定理: 令 X_n 独立同分布, $EX_i = \mu$, $DX_i = \sigma^2$, 则 $n \to \infty$ 时, $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ 的分布弱收敛于标准正态分布.

2. 棣莫弗-拉普拉斯中心极限定理: 令 S_n 为 n 重伯努利试验中事件发生的次数, p 为事件发生的概率, 则 $\frac{S_n - np}{\sqrt{np(1-p)}}$ 在 $n \to \infty$ 时的分布弱收敛于标准正态分布.

3. 中心极限定理的证明: 将 X_i 标准化: $Y_i = \frac{X_i - \mu}{\sigma}$, 则 $EY_i = 0$ 且 $DY_i = 1$. 令 Y_N 的特征函数为 $\varphi(t)$, 故 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$ 的特征函数为 $\left[\varphi\left(\frac{t}{\sqrt{n}}\right)\right]^n$. 由泰勒展开: $\varphi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o(t^2)$, 在 $n \to \infty$ 时, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i$ 的特征函数趋近于 $e^{-\frac{t^2}{2}}$, 恰为 $N(0,1)$ 的特征函数.

1.5 数理统计基本概念

概率论研究的随机变量都有确定的总体, 而现实生活中, 我们通常需要推断某一总体服从何种分布. 这就是数理统计的核心任务: 推断总体服从的分布族, 以及通过样本估计分布中的未知参数. 由于总体的性质只能通过抽样反馈, 因此需要研究由样本推断总体的方法.

在研究过程中, 通常假设各样本与总体同分布且相互独立, 并利用统计量的分布进行推断, 这一思想贯穿了后面的所有章节. 而这部分的任务是打好基础, 理清数理统计的基本概念, 并初步介绍最常用的统计量及其抽样分布, 为参数估计和假设检验打好基础.

1.5.1 统计学基本思想

1. 任务: 收集受随机因素影响的数据, 并根据样本推断总体分布.

2. 总体: 研究对象的全体. 具体分布未知, 一般认为分布族已知, 即推断分布中的未知参数.

3. 样本: 从总体中随机抽取的 n 个数据, 记作 $x_1, x_2, \cdots, x_n$. 若这些样本独立同分布 (i.i.d.), 则称为简单随机样本.

4. 统计量: 当总体分布族已知而参数未知时, 可以构造只与样本有关而与未知参数无关的函数 $T = T(x_1, \cdots, x_n)$, 利用统计量的特征估计未知参数.

1.5.2 常用统计量

1. 样本均值: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. 样本方差: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. 计算时常用公式

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

3. 样本标准差: $s = \sqrt{s^2}$.

4. 样本 k 阶矩: $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$.

1.5.3 抽样分布

1. 定义: 统计量的分布称为抽样分布.

2. χ^2 分布: 若简单随机样本 $x_1, \cdots, x_n \sim N(0,1)$, 则 $\sum_{i=1}^{n} x_i^2$ 服从自由度为 n 的 χ^2 分布, 记作

$$\sum_{i=1}^{n} x_i^2 \sim \chi^2(n)$$

其概率密度函数只在第一象限定义且非对称. 不用刻意记忆其具体 p.d.f., 但需注意 $\chi^2(2)$ 的概率密度函数为 $\frac{1}{2}e^{-\frac{1}{2}x}$ $(x>0)$.

3. F 分布: 令独立随机变量 $X\sim\chi^2(m), Y\sim\chi^2(n)$, 则 $\frac{X/m}{Y/n}$ 服从自由度为 m 与 n 的 F 分布, 记作

$$\frac{X/m}{Y/n}\sim F(m,n)$$

F 分布具有的特殊性质: 若 $X\sim F(m,n)$, 则 $\frac{1}{X}\sim F(n,m)$.

4. t 分布: $X\sim N(0,1), Y\sim\chi^2(n)$, 则 $\frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 记作

$$\frac{X}{\sqrt{Y/n}}\sim t(n)$$

t 分布与 F 分布间存在联系: 若 $X\sim t(n)$, 则 $X^2\sim F(1,n)$.

1.5.4 利用抽样分布统计推断

1. 前置条件: x 服从正态分布, 即 $x\sim N(\mu,\sigma^2)$.

2. σ^2 已知, 对 μ 统计推断: 构造统计量

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

3. σ^2 未知, 对 μ 统计推断: 由 $\bar{x}$ 与 s^2 相互独立, 可构造统计量

$$\frac{\bar{x}-\mu}{s/\sqrt{n}}\sim t(n-1)$$

4. 对 σ^2 统计推断:

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

5. 双正态总体方差比推断: 令 X $N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 构造统计量

$$\frac{s_X^2}{s_Y^2} \sim F(m-1, n-1)$$

1.5.5 充分统计量

1. 定义: 若统计量 T 包含了样本的全部信息, 即给定 T 的取值后, $x_1, \cdots, x_n$ 的分布与未知参数 θ 无关, 则称 T 为 θ 的充分统计量.

2. 因子分解定理: 若总体分布为 $f(x;\theta)$, 存在函数 $g(T,\theta)$ 与 $h(x_1, \cdots, x_n)$ 使得 $f(x_1, \cdots, x_n; \theta) = g(T,\theta) \cdot h(x_1, \cdots, x_n)$, 则 T 为 θ 的充分统计量.

1.6 参数估计

参数估计的目的是对分布族已知, 但含有未知参数的总体, 通过样本估计其中的未知参数. 一种思路为点估计, 即给出参数的确切估计值; 另一种思路为区间估计, 即给出一个大致范围, 有很大的可能性包含参数的真实值. 点估计的评价标准为无偏性和有效性, 即样本越多, 估计值越接近真实值, 且波动尽可能小; 区间估计的手段是通过抽样分布的分位数, 划定统计量所处的范围以包含分布中比例为 $1-\alpha$ 的部分, 再解出参数所处的范围.

考虑到与机器学习接轨, 这一部分列举了很多超纲的内容, 如用先验推后验的贝叶斯估计, 以及求 ML 估计的 EM 算法, 可视自身需要加以取舍.

1.6.1 矩估计

1. 思想: 点估计的一种, 另一种即下面讨论的最大似然估计. 令总体分布为 $X(\theta)$. 用样本矩 $a_k = \sum_{i=1}^{n} X_i^k$ 代替总体矩 $EX^k(\theta)$, 列方程求解未知参数.

2. 以正态分布的矩估计为例: 令 $X \sim N(\mu, \sigma^2)$, 则

$$\begin{cases} EX = \quad \mu \quad\quad = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i \\ EX^2 = \sigma^2 + \mu^2 \quad = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2 \end{cases}$$

从中解出 $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2 = s_n^2$.

3. 矩估计的相合性: 若 $\lim\limits_{n\to\infty} E(\hat{\theta}) = \theta$, $\lim\limits_{n\to\infty} D(\hat{\theta}) = 0$, 则称 $\hat{\theta}$ 为 θ 的相合估计. 矩估计通常为相合估计.

1.6.2 最大似然估计

1. 思想: 选取参数 θ, 使得样本概率 $f(x_1, \cdots, x_n; \theta) = \prod\limits_{i=1}^{n} f(x_i; \theta)$ 最大.

2. 求解方式: 令似然函数 $L(\theta) = \prod\limits_{i=1}^{n} f(x_i; \theta)$, 求解方程以解出 $\hat{\theta}$:

$$\frac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta} = 0$$

3. 以正态分布的 ML 估计为例: 令 $X \sim N(\mu, \sigma^2)$, 则未知参数 θ 由 μ 和 σ^2

构成. 求解下列方程组:

$$\begin{cases} \dfrac{\partial \ln L(\mu,\sigma^2)}{\partial \mu} = -\displaystyle\sum_{i=1}^{n} \frac{\mu - x_i}{\sigma^2} = 0 \\ \dfrac{\partial \ln L(\mu,\sigma^2)}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \displaystyle\sum_{i=1}^{n} \frac{(\mu - x_i)^2}{2\sigma^4} = 0 \end{cases}$$

解得 $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = s_n^2$, 恰好与矩估计结果一致.

4. EM 算法: Expectation Maximization, 针对似然函数中存在不可观测的隐变量 z 时的局部 ML 优化.

(1) E 步: 构造似然函数 $Q(\theta|x,\theta^{(i)}) = E_z[\ln L(\theta;x,z)]$, 目的是消除隐变量 z 的随机性;

(2) M 步: 在已知上轮迭代值 $\theta^{(i)}$ 和样本 x 的情况下, 寻找使似然函数最大的局部最优解:

$$\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|x,\theta^{(i)})$$

(3) 迭代: 设定初始值 $\theta^{(0)}$, 重复 E 步和 M 步直至收敛.

1.6.3 点估计的评价标准

1. 无偏性: 若 $E(\hat{\theta}) = \theta$, 则称 $\hat{\theta}$ 为 θ 的无偏估计.

2. 有效性: 若 $\hat{\theta}_1, \hat{\theta}_2$ 均为 θ 的无偏估计, 且 $D\hat{\theta}_1 > D\hat{\theta}_2$, 则估计 $\hat{\theta}_2$ 比估计 $\hat{\theta}_1$ 更有效.

3. Fisher 信息量: $I(\theta) = E\left[\frac{\partial}{\partial\theta} \ln f(x;\theta)\right]^2$. $I(\theta)$ 越大, 表示总体分布中包含未知参数 θ 的信息越多.

4. Cramer-Rao 不等式: 若 T 为 $g(\theta)$ 的无偏估计, 则

$$DT \geq \frac{[g'(\theta)]^2}{nI(\theta)}$$

若 DT 取到 C-R 下界, 则称 T 为 $g(\theta)$ 的有效估计.

1.6.4 贝叶斯估计

1. 思想: 在抽样之前, 便有关于 θ 的先验信息, 即 θ 服从先验分布 $\pi(\theta)$. 以后验信息

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)\mathrm{d}\theta}$$

以后验分布的最大值点作为 θ 的点估计.

2. 朴素贝叶斯分类器: 假设各属性 $A_1, \cdots, A_n$ 相互独立, 在得到新的样本 $A_1 = a_1, \cdots, A_n = a_n$ 后, 尝试将样本归类: $Y \in y_1, \cdots, y_m$.

朴素贝叶斯分类器的最大化目标为: 以样本信息为先验, 寻找可能性最大的分类结果, 即

$$k = \arg\max_k P(Y = y_k | A_1 = a_1, \cdots, A_n = a_n)$$

由贝叶斯公式以及朴素假设, 最终优化目标为

$$k = \arg\max_k \prod_{i=1}^{n} P(A_i = a_i | Y = y_k)$$

每一项都可以通过现有样本点在 $Y = y_k$ 时 $A_i = a_i$ 的占比求出, 选出使优化目标最大的 k, 便可作出最优决策 $Y = y_k$.

3. 共轭先验: 若 $\pi(\theta)$ 与 $\pi(\theta|x)$ 同属一个分布族, 则称该分布族为 θ 的共轭先验分布族, 此时样本的作用仅是将分布族中的未知参数作调整.

1.6.5 区间估计

1. 思想: 区别于点估计, 区间估计的目标是给出 θ 可能的所在区间 $[\hat{\theta}_1, \hat{\theta}_2]$, 使 θ 有 $1-\alpha$ 的概率落入该区间. 通常使用枢轴量法, 即构造合适的统计量, 利用抽样分布的分位数划定置信限.

2. 单侧区间估计: 若给出 $\hat{\theta}$, 使得 $P(\theta \geq \hat{\theta}) \geq 1-\alpha$, 则 $\hat{\theta}$ 称为单侧置信下限; 若给出 $\hat{\theta}$, 使得 $P(\theta \leq \hat{\theta}) \geq 1-\alpha$, 则 $\hat{\theta}$ 称为单侧置信上限. 求解方法与双侧区间估计类似.

3. 单正态分布总体区间估计: 设 $X \sim N(\mu, \sigma^2)$. 主要依据为 1.5.4 节给出的统计量.

(1) σ^2 已知, 对 μ 区间估计: 构造统计量

$$u = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

则通过 $|u| \leq u_{1-\frac{\alpha}{2}}$ 反解出置信度为 $1-\alpha$ 时 μ 的置信区间.

(2) σ^2 未知, 对 μ 区间估计: 构造统计量

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

则通过 $|t| \leq t_{1-\frac{\alpha}{2}}(n-1)$ 反解出置信度为 $1-\alpha$ 时 μ 的置信区间.

(3) 对 σ^2 区间估计: 构造统计量

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

则通过 $\chi^2 \in [\chi^2_{\frac{\alpha}{2}}(n-1), \chi^2_{1-\frac{\alpha}{2}}(n-1)]$ 反解置信度为 $1-\alpha$ 时 σ^2 的置信区间.

4. 双独立正态分布总体区间估计: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$. X 有 m

个样本, 且 Y 有 n 个样本.

(1) σ_1^2 与 σ_2^2 已知, 对 $\mu_1-\mu_2$ 估计: 构造统计量

$$\frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}\sim N(0,1)$$

(2) $\sigma_1^2=\sigma_2^2$ 未知, 对 $\mu_1-\mu_2$ 估计: 构造统计量

$$\sqrt{\frac{m+n-2}{\frac{1}{m}+\frac{1}{n}}}\frac{\bar{x}-\bar{y}-(\mu_1-\mu_2)}{\sqrt{(m-1)s_X^2+(n-1)s_Y^2}}\sim t(m+n-2)$$

(3) 对方差比 $\frac{\sigma_1^2}{\sigma_2^2}$ 估计: 构造统计量

$$\frac{s_X^2/\sigma_1^2}{s_Y^2/\sigma_2^2}\sim F(m-1,n-1)$$

1.7 假设检验

统计学中会有很多假设, 最常见的是假设总体服从正态分布. 但这些假设是不是准确呢? 对这些假设作检验的本质就是反证法: 如果假设是准确的, 则选定统计量应该服从某分布, 但统计量结果在该分布中出现可能性很小, 就推翻假设的正确性.

最常见的假设检验即检验分布的参数是否为某一定值, 如灯泡的寿命是否维持原状 (指数分布的参数是否为某一定值), 是该部分的主要内容. 也有其它的检验目标, 如总体分布的假设是否合理, 样本是否为简单随机样本等, 也将作一定介绍.

1.7.1 基本思想

1. 假设检验的基本问题: 选定原假设 H_0 与备择假设 H_1, 若 H_0 发生的可能性非常小, 则拒绝原假设而接受备择假设; 若不能拒绝原假设, 则接受原假设. 假设检验的一般问题记作

$$H_0 : \_\_\_\_ \text{ vs } H_1 : \_\_\_\_$$

2. 检验方法: 先假设 H_0 成立, 结合某一检验统计量的分布给出拒绝域. 若该统计量落入拒绝域, 则拒绝 H_0, 否则接受 H_0.

3. 两类错误: 若 H_0 为真, 但统计量落入拒绝域, 则犯了第一类错误 α; 若 H_0 为假, 但接受了 H_0, 则犯第二类错误 β.

4. 显著性水平 α: 控制犯第一类错误的可能性 $\leq \alpha$, 即假设 H_0 为真, 检验统计量落入拒绝域的概率应小于等于 α.

1.7.2 正态总体假设检验

1. 单正态总体假设检验: $X \sim N(\mu, \sigma^2)$, 样本量为 n, 构造的统计量依然如 1.5.4 节所述. 由于等式假设和不等式假设仅涉及双侧置信区间和单侧置信区间, 处理手法类似, 故仅以等式假设为例.

(1) σ^2 已知, 检验 $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$: 构造统计量

$$u = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

当 H_0 为真时, $u \sim N(0, 1)$, 即拒绝域为 $\{|u| \geq u_{1-\frac{\alpha}{2}}\}$.

(2) σ^2 未知, 检验 $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$: 构造统计量

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

当 H_0 为真时, $t \sim N(0,1)$, 即拒绝域为 $\{|t| \geq t_{1-\frac{\alpha}{2}}\}$.

(3) 检验 $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$: 构造统计量

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

当 H_0 为真时, $\chi^2 \sim \chi^2(n-1)$, 即拒绝域为 $\{\chi^2 \leq \chi^2_{\frac{\alpha}{2}}(n-1), \text{ or } \chi^2 \geq \chi^2_{1-\frac{\alpha}{2}(n-1)}\}$.

2. 双正态总体假设检验: 设 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$, 且 X 的样本数为 m, Y 的样本数为 n.

(1) σ_1^2 和 σ_2^2 已知, 检验 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$: 取检验统计量

$$u = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \sim N(0,1)$$

(2) $\sigma_1^2 = \sigma_2^2$ 未知, 检验 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$: 取检验统计量

$$t = \frac{m+n-2}{\sqrt{\frac{1}{m} + \frac{1}{n}}} \frac{\bar{x} - \bar{y}}{(m-1)s_X^2 + (n-1)s_Y^2} \sim t(m+n-2)$$

(3) $m = n$ 且方差未知的成对样本检验, 检验 $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$: 取检验统计量

$$t = \frac{\bar{x} - \bar{y}}{(s_X^2 + s_Y^2)/\sqrt{2n}} \sim t(2n-2)$$

(4) 方差比检验, 检验 $H_0 : \sigma_1^2 = \sigma_2^2$ vs $H_1 : \sigma_1^2 \neq \sigma_2^2$: 取检验统计量

$$F = \frac{s_X^2}{s_y^2} \sim F(m-1, n-1)$$

1.7.3 广义似然比检验

1. 思想: 检验 $H_0:\theta\in\Theta$ vs $H_1:\theta\notin\Theta$ 时, 取统计量

$$\Lambda(x_1,\cdots,x_n)=\frac{\sup\limits_{\theta\in\Theta}f(x_1,\cdots,x_n;\theta)}{\sup\limits_{\theta\notin\Theta}f(x_1,\cdots,x_n;\theta)}$$

Λ 越大, 说明 H_0 越有可能成立.

2. 拒绝域: 尚未有统一的形式. 但可以用渐近分布 $2\Lambda\sim\chi^2(n)$, 其中 n 为独立参数的个数.

1.7.4 拟合优度检验

1. 分布拟合检验: 设总体被分为 r 个类 $A_1,\cdots,A_r$, A_i 类中有 n_i 个样本, 检验原假设 $H_0:A_i$ 所占比率为 p_i, 其中 $\sum\limits_{i=1}^{r}p_i=1$, $\sum\limits_{i=1}^{r}n_i=n$. 构造统计量

$$\chi^2=\sum_{i=1}^{r}\frac{(n_i-np_i)^2}{np_i}\sim\chi^2(r-1)$$

拒绝域为 $\{\chi^2\geq\chi^2_{1-\alpha}(r-1)\}$.

2. χ^2 拟合优度检验: 若 X 的分布函数为 $F(x)$, 将样本归为 r 类: $(-\infty,a_1]$, $(a_1,a_2],\cdots,(a_{r-1},+\infty)$, 每一类理论占比 $p_i=F(a_i)-F(a_{i-1})$, 实际有 n_i 个样本落入第 i 类, 在此基础上应用分布拟合检验.

1.7.5 正态性检验

1. 目的: 检验总体是否服从正态分布.

2. 概率图纸法: 令样本从小到大排列为 $x_{(1)}, \cdots, x_{(n)}$, 将点 $\left(x_{(i)}, \frac{i-0.375}{n+0.25}\right)$ 描在图纸上, 若近似成一条直线, 则认为总体 X 服从正态分布.

1.7.6 游程检验

1. 目的: 检验样本是否随机抽取.

2. 游程检验: 设样本中位数为 M_e, 将样本按抽样时间顺序排列, 并将 $\geq M_e$ 的值替换为 1, $< M_e$ 的值替换为 0, 得到一串 0-1 序列.

3. 判断依据: 把以 0 为界的连续 1 串称为 1 游程, 以 1 为界的连续 0 串称为 0 游程. 若 0 游程数和 1 游程数之和过大或过小, 则拒绝采样随机性, 拒绝域通过查表得出.

1.8 方差分析

单因素方差分析用于解决这一类问题: 控制其它因素都一样, 就改变一个因素, 会不会造成很显著的影响? 换用统计语言来说, 设一个因素有不同的各个水平, 这些水平的均值是否相等? 若相等, 则因素 A 对实验结果没啥影响; 若不相等, 则因素 A 的影响显著. 方差分析作出一个假设: 各水平服从方差相等的正态分布.

样本的波动可由两部分构成: 一是随机性导致同一水平内的数据波动, 即组内误差; 二是因素 A 的作用使不同水平的样本发生了质变, 即组间误差. 显然组间误差占比越高, 因素 A 的影响越显著, 方差分析表也基于此思想得出.

1.8.1 基本思想

1. 检验问题: 设因素 A 有 r 个水平, 各水平均为正态总体 $N(\mu_i, \sigma^2)$ 且方差相等, 检验因素 A 对均值的影响是否显著, 即检验

$$H_0 : \mu_0 = \mu_1 = \cdots = \mu_r \quad \text{vs} \quad H_1 : \mu_0, \mu_1, \cdots, \mu_r \text{不全相等}$$

2. 统计模型: 令 y_{ij} 表第 i 个总体的第 j 次试验结果, m_i 为水平 A_i 的样本数, 总样本数 $n = \sum_{i=1}^{r} m_i$. 记 $\varepsilon_{ij} = y_{ij} - \mu_i$ 为随机误差, 则 ε_{ij} 相互独立, 且 $\epsilon_{ij} \sim N(0, \sigma^2)$.

3. 组内偏差: 令 $\bar{y}_i$ 表示第 i 个总体的组内均值, 则用 $S_e = \sum_{i=1}^{r} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$ 表示第 i 个总体的组内偏差.

4. 组间偏差: 令 $\bar{y}$ 为所有样本的均值, 则用 $S_A = \sum_{i=1}^{r} m_i(\bar{y}_i - y)^2$ 表示因素 A 导致的组间偏差.

5. 平方和分解公式: 令总偏差为 $S_T = \sum_{i=1}^{r} \sum_{j=1}^{m_i} (y_{ij} - \bar{y})^2$, 由 $y_{ij} - \bar{y} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})$ 可推导如下重要公式:

$$S_T = S_A + S_e$$

1.8.2 单因素方差分析

1. 基本思想: 由平方和分解公式, 若 S_A 远大于 S_e, 即偏差大部分由因素 A 导致, 则认为因素 A 影响显著. 构造检验统计量

$$F = \frac{S_A/(r-1)}{S_e/(n-r)} \sim F(r-1, n-r)$$

当 F 大于临界值 $F_{1-\alpha}(r-1, n-r)$ 时, 拒绝原假设, 认为因素 A 显著.

2. 单因素方差分析表

	平方和	自由度	均方	F 比	临界值
因素 A	S_A	$r-1$	$S_A/(r-1)$	$F=\frac{S_A/(r-1)}{S_e/(n-r)}$	$F_{1-\alpha}(r-1,n-r)$
误差 e	S_e	$n-r$	$S_e/(n-r)$		
总和	S_T	$n-1$			

3. 填表方法: 先计算 $S_T=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y})^2$, 再计算 $S_e=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m_i}(y_{ij}-\bar{y}_i)^2$, 由平方和分解公式计算 $S_A=S_T-S_e$, 再从左到右填写剩下内容. 最后比较 F 比和临界值, 若 F 比大于临界值, 则拒绝 H_0, 认为因素 A 显著.

1.8.3 方差齐性检验

1. 目的: 检验 r 个总体是否符合方差相等的假设条件.

2. 哈特利检验: 令 $H=\frac{\max\{s_1^2,\cdots,s_r^2\}}{\min\{s_1^2,\cdots,s_r^2\}}$, 则 H 越接近 1, 越有可能认为方差相等. 查表得 H 分布的分位数和拒绝域.

1.9 回归分析

现实生活中, 很难有自变量和因变量的关系能和数学分析中研究的函数一样, 具有良好的性质. 但是我们可以用性质好的函数去拟合变量间的相关关系, 并综合运用前述统计方法, 评价这种拟合到底合不合理, 这就是回归分析的最基本思想.

对单变量关系的情形, 若将样本点 (x,y) 描在图纸上, 仅有散点图很像一条直线时, 我们才能猜测变量间存在线性关系, 其它形状的散点图都不能直接得出结论. 因此一元线性回归是回归分析中最重要的一环, 即用线性函数 $y=\beta_0+\beta_1 x$ 拟合 x 和 y 间的相关关系.

1.9.1 基本思想

1. 目的: 令 x 为自变量, y 为因变量, 用函数关系 $y = f(x)$ 拟合 x 与 y 间的相关关系, 并要求误差尽可能小.

2. 一元线性回归: 用 $y = \beta_0 + \beta_1 x$ 拟合相关关系, 统计模型为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

1.9.2 回归系数的最小二乘估计

1. 目的: 令 n 组样本对为 (x_i, y_i), 求 $\hat{\beta}_0, \hat{\beta}_1$, 使误差和 $Q = \sum\limits_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ 最小.

2. 求解: 由 $\frac{\partial Q}{\partial \beta_0} = 0$, 得 $2\sum\limits_{i=1}^{n}(\beta_0 + \beta_1 x_i - y_i) = 0$.

由 $\frac{\partial Q}{\partial \beta_1} = 0$, 得 $2\sum\limits_{i=1}^{n} x_i(\beta_0 + \beta_1 x_i - y_i) = 0$. 联立解得

$$\hat{\beta}_1 = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - (\sum\limits_{i=1}^{n} x_i)^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

3. 估计的性质: $\hat{\beta}_0, \hat{\beta}_1$ 均为无偏估计.

1.9.3 区间估计与预测

1. 区间估计目的: 给定 x_0, 求 $E(y_0) = \beta_0 + \beta_1 x_0$ 的区间估计.

2. 区间估计方法: 令 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, 构造统计量

$$\frac{\hat{y}_0 - (\beta_0 + \beta_1 x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}}} \sim t(n-2)$$

其中 $l_{xx}=\sum\limits_{i=1}^{n}x_i^2-\frac{1}{n}(\sum\limits_{i=1}^{n}x_i)^2$.

3. 预测: 给定 x_0 的条件下, 求 y_0 的区间估计. 构造统计量

$$\frac{\hat{y}_0-y_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}}}\sim t(n-2)$$

从中反解出 y_0 所处的区间.

1.9.4 显著性检验

1. 目的: 检验 y 和 x 的相关性是否显著. 若 y 与 x 无关, 则 $\beta_1=0$, 即检验假设 $H_0:\beta_1=0$ vs $H_1:\beta_1\neq 0$.

2. t 检验: 取检验统计量

$$t=\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{l_{xx}}}\sim t(n-2)$$

3. F 检验: 取 $S_R=\sum\limits_{i=1}^{n}(\hat{y}_i-\bar{y})^2$, $S_e=\sum\limits_{i=1}^{n}(y_i-\hat{y}_i)^2$, $S_e=\sum\limits_{i=1}^{n}(y_i-\bar{y})^2$, 则有平方和分解公式 $S_T=S_R+S_e$. 列方差分析表:

	平方和	自由度	均方	F 比	临界值
回归 R	S_R	1	S_R	$F=\frac{S_R}{S_e/(n-2)}$	$F_{1-\alpha}(1,n-2)$
误差 e	S_e	$n-2$	$S_e/(n-2)$		
总和	S_T	$n-1$			

与方差分析流程一致. 实际上, F 检验与 t 检验等价.

1.9.5 多元线性回归

1. 目的: 用 $y=\omega^Tx+b$ 拟合向量 y 与 x 之间的相关关系.

2. 求解: 最小二乘法. 解为 $\theta = (X^TX)^{-1}X^TY$, 其中 $(x1, y1), \cdots, (xm, ym)$ 为样本点,

$$\theta = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \\ b \end{pmatrix}, \quad X = \begin{pmatrix} x1_1 & x1_2 & \cdots & x1_n & 1 \\ x2_1 & x2_2 & \cdots & x2_n & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ xm_1 & xm_2 & \cdots & xm_n & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} y1 \\ y2 \\ \vdots \\ ym \end{pmatrix}$$

1.9.6 非线性回归

1. 确定函数形式: 根据散点图形状, 确定变换 $z = \varphi(y)$.

2. 作变换: 重新绘制 (x, z) 的散点图.

3. 线性回归: 若 z 与 x 的散点图近似直线, 则对 x 和 z 作一元线性回归.

第二章　数据结构

Data Structure

数据结构是计院三大核心课之一，在设计离散类型算法时，决定数据结构通常是第一步。

学习和设计数据结构时，需要掌握数据结构两大核心：逻辑结构和存储结构。前者定义了数据间的关系和需要的数据操作，即确定抽象数据类型 ADT；后者决定数据在程序中的存储方式和具体实现，一般就分顺序和链表两种。在此基础上，需要掌握数据结构对应的典型应用，也就是由数据结构派生出的算法。

数据结构的内容：线性表，栈，队列，字符串，树，二叉树，图，查找技术。

2.1 数据结构基本概念

2.1.1 数据结构

1. 定义: 非数值问题的数据组织和处理。

2. 包含内容: 逻辑结构，存储结构，算法，数据处理技术。

3. 逻辑结构: 描述数据间的关联关系。常见的有集合，线性结构，树形结构和图结构。

4. 存储结构: 描述数据在计算机内的存储方式，分为顺序和链表两种。

5. 数据操作: 即数据的增加，删除，修改，查找，排序等

6. 抽象数据类型 (ADT): 数据结构以及定义在其上的数据操作的总称。确定数据的逻辑结构和操作后可设计 ADT，确定存储结构后可具体实现 ADT。

2.1.2 面向对象程序设计

1. 类: 计算机用于创建对象的模板，含有属性，方法和事件等。

2. 封装: 将数据成员与方法嵌入类中，外部无法访问私有成员，只能访问类的给定接口以实现对应功能。

3. 声明: 类需要声明数据类型，成员变量，成员函数与成员权限，不占用存储空间；声明对象即将类实例化，可用”。” 访问对象公有成员。

4. 构造函数: 声明对象时首先调用的函数，作初始化工作。

5. 析构函数: 对象被释放时调用的函数，作清理工作。

6. C++ 的动态内存分配: new 申请分配内存，并返回指向该内存的指针；delete 释放指针指向的内存，指针本身不被删除。

7. 模板: template<typename T>，使变量或函数可取任意预设类型。

8. 参数传递: 有传值，传指针 (*)，传引用 (&) 三种. 传值改形参时实参不

变，后两种改形参时实参跟着改变。

2.2 线性表

2.2.1 逻辑结构

1. 定义: n 个具有相同数据类型的元素构成的序列，n 称为线性表的长度。

2. C++ 中的 STL: vector。

```
#include<vector>
```

3. 数据操作:

```
vector<int> V; //定义与初始化
V.push_back(int); //在线性表末尾增加元素
V.insert(int p, int); //在第p个位置插入元素
V.pop_back(); //删除线性表末尾的元素
V.erase(int p); //删除第p个位置的元素
V.clear(); // 清空线性表
V.size(); // 查询线性表元素个数
V.begin(); //线性表首元素位置
V.end(); // 线性表末尾元素位置
V[p]; //访问线性表第p个位置的元素，也可V.at(int p);
V.empty(); //判断线性表是否为空
```

4. 抽象数据类型定义:

(1) private: 数据主体，线性表长度；

(2) public: 提供数据增删改查，求长度，判断是否为空的接口

2.2.2 存储结构

1. 顺序表: 静态存储结构。数据主体用数组存储，适用于事先知道最大长度的情形。

2. 链表: 动态存储结构。用结点存储单元素及其后继元素指针，插入与删除很方便。定义如下:

```
struct node
{
    int x;        //数据主体
    node *p;      //指向后继结点
};
```

2.3 栈

2.3.1 逻辑结构

1. 定义: 仅在顶端插入与删除元素的线性表。越先进栈的元素越后出栈，称为 FILO。

2. C++ 中的 STL: stack。

```
#include<stack>
```

3. 数据操作:

```
stack<int> S; //定义与初始化
S.push(int); //在栈顶插入元素
S.top(); //查询栈顶元素
S.pop(); //弹出栈顶元素
S.size(); //查询栈内元素个数
S.empty(); //判断栈是否为空
```

4. 抽象数据类型定义:

(1) private: 数据主体，栈的长度；

(2) public: 提供数据的插入和删除，查询栈顶元素，求长度，判断是否为空的接口。

2.3.2 存储结构

1. 顺序栈: 同顺序表，在此基础上仅支持栈顶元素的插入与删除。

2. 链栈: 同链表，在此基础上仅支持栈顶的插入与删除。

2.3.3 算法与应用

1. 表达式求值: 所有的计算按照从左到右的顺序执行。用两个栈分别存储操作数和运算符:

(1) 遇到操作数: 压入操作数栈；

(2) 遇到运算符: 若优先级比运算符栈顶元素低，则将操作数栈的两个顶端元素弹出并根据运算符栈顶元素运算，将运算结果压入操作数栈；否则压入运算符栈。

2. 深度优先搜索: 以栈的形式存储搜索顺序。搜索某一步时，依次选择下一步方案并压入栈；当前步结束后，将其从栈顶弹出。

2.4 队列

2.4.1 逻辑结构

1. 定义: 仅允许一端插入另一端删除的线性表。越先进栈的元素越先出栈，称为 FIFO。

2. C++ 中的 STL: queue。

```
#include<queue>
```

3. 数据操作:

```
queue<int> Q;   //定义与初始化
Q.push(int);   //队列头部插入元素
Q.front();   //查询队列头部元素
Q.back();   //查询队列尾部元素
Q.pop();   //弹出队列头部元素
Q.size();   //  查询队列元素个数
Q.empty();   //判断队列是否为空
```

4. 抽象数据类型定义:

(1) private: 数据主体，队列的长度；

(2) public: 提供数据的插入和删除，查询队列头部/尾部，求长度，判断是否为空的接口。

2.4.2 存储结构

1. 顺序队列: 同顺序表，在此基础上仅支持队尾插入，队头删除。

2. 链队列: 同链表，在此基础上仅支持队尾插入，队头删除。

2.4.3 算法与应用

1. 广度优先搜索: 以队列的形式存储搜索顺序。搜索某一步时, 将所有可能的下一步方案入队；当前步结束后，将其从队列头部弹出。

2. 优先队列: 在队列中按照某种优先级顺序出队，可用大根堆或二叉搜索树维护。C++ 中的 STL 为 priority_queue，头文件和队列一致。

2.5 字符串

2.5.1 逻辑结构

1. 定义: 零个或多个字符组成的有限序列。

2. C++ 中的 STL: string。

```
#include<string>
```

3. 数据操作: 由于字符串和 vector<char> 没啥区别，因此支持的主要操作也没啥区别。

```
string s;  //定义并初始化
s.push_back(char);  //在字符串末尾增加字符
s.insert(int p,char);  //在第p个位置插入字符
s.pop_back();  //删除字符串末尾的元素
s.erase(int p);  //删除第p个位置的元素
s.clear();  // 清空字符串
s.size();  // 查询字符个数
s.find(char);  //查找字符并返回其位置
s.substr(int n, int l);  // 返回从第n位开始长度为l的字串
s.begin();  //字符串首元素位置
s.end();  // 字符串末尾元素位置
s[p];  //访问字符串第p个位置的元素，也可s.at(int p);
s.empty();  //判断字符串是否为空
```

4. 抽象数据类型:

(1) private: 字符串主体，字符串长度；

(2) public: 提供字符增删改查，求长度，判断是否为空的接口

P. S. 字符串通常采用顺序存储，故不讨论其存储结构。

2.5.2 算法与应用

1. KMP 模式匹配算法: 应用于在字符串 T 中匹配模式子串 S。

(1) 计算模式串的 next 数组: 令 next[j] = k，其中 k 为使得 T[0]~T[k - 1] = S[j - k]~S[j - 1] 的最大值。

(2) 正常匹配: 逐一检查是否有 S[i] = T[j]。

(3) 回退: 若某一位 S[i] != T[j]，则 i 不动，j 回退至 next[j]，继续匹配。

2.6 树

2.6.1 逻辑结构

1. 定义: n 个结点的集合。仅能有一个根节点 (最上层)，以此基础上产生由上层到下层的一对多映射，除根结点外的所有点称为子结点。

2. 结点的度: 结点拥有的子结点个数。所有结点度的最大值称为这棵树的度。

3. 路径: $n_1 \rightarrow n_2 \cdots \rightarrow n_k$。其中 n_i 是 n_{i-1} 的子节点。

4. 数据操作: 前序遍历，中序遍历，后序遍历和层序遍历。

(1) 前序遍历: 先根序遍历。遍历顺序为根-其它结点。

(3) 后序遍历: 后根序遍历。遍历顺序为其它结点-根。

(4) 层序遍历: 从上到下遍历层，每一层按从左到右的顺序遍历结点。

2.6.2 存储结构

1. 双亲表示法: 用一维数组存储树中各个结点的数据信息，以及结点的父亲在数组中的下标。利用了子结点只能有一个父亲的特性。

2. 孩子表示法: 用一维数组存储树中各个结点的数据信息，并以挂链表的形式存储结点的孩子在数组中的下标。

2.7 二叉树

2.7.1 逻辑结构

1. 定义: 二叉树是特殊的树形结构，每个结点顶多有两个子结点，即左儿子与右儿子。

2. 满二叉树: 除叶子结点外，其它结点都有两个子结点，且所有叶子在同一层。

3. 完全二叉树: 在满二叉树中，从最后一个结点开始连续去掉任意个结点后得到的二叉树。

4. 数据操作: 前序遍历，中序遍历，后序遍历和层序遍历。

(1) 前序遍历: 先根序遍历。遍历顺序为中-左-右。

(2) 中序遍历: 中根序遍历。遍历顺序为左-中-右。

(3) 后序遍历: 后根序遍历。遍历顺序为左-右-中。

(4) 层序遍历: 从上到下的遍历层，每一层按从左到右的顺序遍历结点。

2.7.2 存储结构

1. 顺序存储: 按层序遍历顺序，用数组存储二叉树中的所有结点。若从 0 开始编号，则 i 号结点的两个儿子为 2i+1 和 2i+2。

2. 二叉链表: 用一个 node 结构体存储单个结点，分别存储该结点的两个儿子。结构体定义如下:

```
struct node{
```

```
    int data;
    node *left, *right;   // 指向两个儿子结点的指针
};
```

3. 三叉链表: 在二叉链表的基础上，加一个指针指向父结点。

2.7.3 算法与应用

1. Huffman 算法: 最优二叉树构造算法

(1) 带权路径长度: 对每个叶子结点，带权路径长度为叶子权值 × 层数(根结点算第 0 层)。树的带权路径长度为所有叶子带权路径长度之和。

(2) 最优二叉树: 给定一组具有确定权值的叶子结点，带权路径长度最小的二叉树。

(3) Huffman 算法: 贪心算法。每次将当前权值最小的两结点作为同一结点的两个儿子合并，把两个结点权值之和作为新的结点，直至只剩一个结点即根结点。

(4) Huffman 编码: 将数据出现频率作为叶子结点权重，按左 0 右 1 层序编码。显然出现频率越高的数据编码越短，且任何一个编码都不是其它编码的前缀。

2.8 图

2.8.1 逻辑结构

1. 定义: 由顶点集合 V 和边集合 E 组成的数据结构，即 G=(V,E)。

2. 图的方向: 若顶点之间的边为有序对，则图为有向图；若边无首尾顺序，则图为无向图。

3. 边的权: 对边赋予的数值量，可表示距离等具体含义。

4. 顶点的度: 以某点为顶点的边数称为该点的度。有向图中，以该点为起点的边数称为出度，以该点为终点的边数称为入度。

5. 邻接: 若两个顶点之间存在边，则称这两个顶点邻接。

6. 连通图: 任意两顶点间均有路径。在有向图中，任意两顶点均有路径的极大子图称为强连通分量。

7. 数据操作: 图的遍历。分 DFS 和 BFS 两种，分别用栈和队列规划搜索顺序。

2.8.2 存储结构

1. 邻接矩阵: 设顶点个数为 n，用 n * n 的矩阵 a 存储边信息，a[i][j] 代表从 i 点和 j 点间的边权，若 i 点到 j 点无边，则 a[i][j] = inf。

2. 邻接表: 用一维数组存储顶点信息，同时挂链表存储顶点的邻接点。

3. 三元组: 设边数为 k，用 k*3 的矩阵存储边信息，三个数据分别为边权和连接的两个点。三元组适合储存点比边多得多的情形，即稀疏图。

2.8.3 算法与应用

1. 最短路径

(1) Floyd 算法: 多源最短路径算法，时间复杂度 O(n^3)。对任意两个点，枚举中间点以更新两点间的最短距离。

(2) Dijkstra 算法: 单源最短路径算法，时间复杂度 O(n^2)。初始化所有点为白点，从起点开始，每一轮将当前与起点距离最小的点设为蓝点，并用该点更新其它距离。

2. 最小生成树

(1) Prim 算法: 和 Dijkstra 一模一样的蓝白点算法。

(2) Kruskal 算法: 贪心算法。每次寻找当前的最短边，若不和现有边构成回路，则选用该边。

3. 有向无环图与 AOV 网

(1) 拓扑排序: 对有向图构造拓扑序列。每次选一个没有前驱的顶点，并删去该顶点以及所有以该点为起点的边。

(2) 关键路径: 计算活动的最早开始时间和最晚开始时间，两者相等的活动为关键活动，关键活动构成的路径为关键路径，此路径上的所有活动必须严格按期开工。

4. 求强连通分量: 首先从某一顶点开始作 DFS，得到遍历序列，再按此顺序逆向 DFS，得出各强连通分量。

2.9 查找技术

2.9.1 线性表的查找技术

1. 顺序查找: 从头到尾逐一与目标值比较。

2. 折半查找: 在有序线性表中, 每次选取区间中间值与目标比较，将搜索区域减半.

2.9.2 树表的查找技术

1. 二叉排序树: 一棵特殊的二叉树，左儿子小于根节点，右儿子大于根节点，且每棵子树均是二叉排序树。

2. 平衡二叉树: 左右子树的深度至多相差 1。将二叉排序树的不平衡点调整为平衡二叉树，能优化平均查找效率。

3. B 树: 一棵特殊的 m 叉树，根结点至少有 2 棵子树，其它结点至少有 m/2 棵，所有叶子结点出现在同一层。

2.9.3 散列表的查找技术

1. 基本思想: 构建数据到关键码的映射，根据关键码定位数据。

2. 散列函数设计: 通常使用除留余数法，选定适当的素数 p，用数据模 p 的余数作为关键码。

3. 处理冲突的方法: 开放定址法，即在冲突位置的下一个空地址处存储；挂链表法，即散列表存储每个同义词链表的头指针，遇到冲突就把数据挂到对应链表尾部。

2.10 排序技术

2.10.1 插入型排序

1. 直接插入排序: 对每一个元素，在已排序序列中找到合适的位置并插入。

2. 希尔排序: 将整个序列分为若干子序列执行插入排序，最后对所有子序列执行插入排序。

2.10.2 交换型排序

1. 冒泡排序: 两两比较相邻记录，若顺序不对则交换。

2. 快速排序: 将序列对半分，在左右区间中选择顺序不对的两个元素交换。

2.10.3 选择型排序

1. **选择排序:** 从未排序序列中依次选择最大/最小值，组成有序序列。
2. **堆排序:** 维护一个大根堆，每次将堆顶元素作为最大值导出。

2.10.4 归并型排序

1. **二路归并:** 将序列划分为两个序列，分别排序后合并为有序序列。

第三章　数据库系统

3.1　数据库基本概念

1. 数据：存储的基本对象，记录是计算机表示的和存储数据的一种格式或方法

2. 数据库：长期存储在计算机中有组织、可共享的大量数据的集合。特点是永久存储、有组织、可共享

3. 数据库系统：计算机引入数据库后的系统。特点：相互关联的数据集合，较少的数据冗余（关系规范化，程序与数据相互独立，数据安全可靠，保证数据正确性

4. 数据独立性：应用程序不依赖于特定物理表示。逻辑独立性是信息内容变化不影响应用程序，物理独立性是物理存储位置/结构不影响应用程序

3.2　数据模型

1. 数据模型：对现实数据的模拟。分为概念层数据模型和组织层数据模型。

2. 概念层数据模型：抽象实体及实体间的关联关系，常见的有（实体-联系）E-R 模型。

3. E-R 模型：把实体写在框内，属性写在圆角矩形内，联系写在菱形内。用连线将联系与关联的实体连接。

4. 组织层数据模型：以数据的组织方式描述信息。常见的为关系模型。

5. 关系模型：用二维表（关系）组织数据，关系数据库是关系的集合。关系中的行称为记录（元组），列称为属性。关系模型中的主要操作是增删改查，用行、列唯一确定数据。

6. 主码：关系中用于唯一确定一个元组的最小属性组。可作为关系主码的所有属性称为候选码。

7. 数据完整性约束：实体完整性、参照完整性和用户定义完整性。

（1）实体完整性：所有关系必须有主码，不允许记录无主码或主码相同。

（2）参照完整性：关系中某属性的取值受其它关系约束，用外码实现。

（3）用户定义完整性：针对实际应用定义的数据约束条件

8. 数据库系统的结构：三级模式结构，分为外模式、概念模式和内模式

（1）内模式：数据的物理存储方式

（2）概念模式：数据的逻辑结构和关联关系

（3）外模式：特定用户感兴趣部分的局部描述

3.3 SQL 语言

1. SQL：关系数据库管理系统的标准语言

（1）数据定义：CREATE，DROP，ALTER

（2）数据查询：SELECT，FROM，WHERE，HAVING 等

（3）数据操作：INSERT，UPDATE，DELETE

（4）数据控制：GRANT，REVOKE，DENY

2. 数据类型：int，numeric，decimal，float，char，varchar，text，datetime

3. 关系定义：

```
CREATE TABLE 表名（
列名 数据类型 完整性约束，
列名 数据类型 完整性约束...）
```

4. 完整性约束：常见的有 NOT NULL（限定非空），DEFAULT（默认值），UNIQUE（不能重复），CHECK（不等式约束），主码约束和外码约束

（1）主码约束：PRIMARY KEY，指定当前属性为主码；或单独列出，PRIMARY KEY（列名）

（2）外码约束：FOREIGN KEY（列）只能单独列出，FOREIGN KEY（列名）REFERENCES 参照表名（参照列名）。参照列名必为主键或有 UNIQUE 约束

（3）不等式约束：CHECK(逻辑表达式)

5. 关系删除： DROP TABLE 表名

6. 关系修改：ALTER TABLE 表名，在其后加入要改的内容。

（1）修改列定义：ALTER COLUMN 列名新数据类型

（2）加一列：ADD 列名数据类型完整性约束

（3）删一列：DROP 列名

3.4 数据查询

1. 基本结构： SELECT FROM WHERE GROUP BY HAVING ORDER BY

2. 单表查询：

```
SELECT 列名 AS 别名
FROM 表名
WHERE 筛选条件
```

（1）去重：SELECT DISTINCT · · ·

（2）区间筛选：WHERE 列名 BETWEEN 上限 AND 下限

（3）集合筛选：WHERE 列名 IN · · ·

（4）字符串匹配筛选：WHERE 列名 LIKE '· · ·'。_ 匹配 1 个字符，

（5）筛选空值：WHERE 列名 IS NULL

（6）多重条件筛选：WHERE (· · · OR · · ·) AND · · ·

（7）查询结果排序：ORDER BY 列名（ASC/DESC)

（8）聚合函数：SELECT f(列名) AS 别名。可选用 COUNT，SUM，AVG，MAX，MIN 等。

（9）分组统计：GROUP BY 列名 HAVING 组筛选条件

（10）取前 n 行数据：SELECT TOP n WITH TIES 列名

（11）CASE 表达式：SELECT 表名 CASE 值 1 THEN 别名 1 CASE 值 2 THEN 别名 2· · ·

3. 多表连接查询：SELECT 列名 FROM 表 1 INNER JOIN 表 2 ON 筛选条件。为表取别名可实现自连接，即表 1、表 2 为同一张表

4. 子查询：将另一个查询结果嵌套在查询条件中。SELECT 列名 FROM 表名 WHERE 表达式 IN（子查询）。存在性测试中，WHERE EXISTS（子查询）

3.5 其它数据操作

1. 插入数据： INSERT INTO 表名 VALUES (属性值)

2. 更新数据： UPDATE 表名 SET 列名 = 表达式 WHERE 筛选条件

3. 删除数据： DELETE FROM 表名 WHERE 筛选条件

4. 查询结果保存： SELECT 列名 INTO 新表名

3.6 视图

1. 概念：查询语句产生的结果，有别名和列名。

2. 创建视图：CREATE VIEW 视图名 AS SELECT···

3. 修改视图：ALTER VIEW 视图名 AS SELECT···

4. 删除视图：DROP VIEW 视图名

5. 索引：表中所包含值的列表，用来加快数据的查询速度。CREATE CLUSTERD INDEX 索引名 ON 表名 (列名)

3.7 规范化理论

1. 函数依赖：某一属性的取值依赖于另一属性。若两条记录的 X 相同能推出 Y 相同，则称 X 决定 Y，Y 依赖于 X。

（1）完全依赖：X 中任意去掉一个属性都不能唯一决定 Y

（2）部分依赖：X 中某一子集可决定 Y

（3）传递依赖：若 Z 依赖于 Y 且 Y 依赖于 X，则称 Z 传递依赖于 X

2. 关系规范化：将有错误函数依赖的关系模式转化为良好关系模式

（1）候选码：决定关系中全部属性值的最小属性组

（2）主码：关系中用于唯一表示记录的属性组

（3）外码：用于建立关系表间关联关系的属性组

3. 范式：关系所满足的不同程度要求，包括 1NF、2NF 和 3NF。

（1）第一范式：不包含重复组的关系。把所有数据项都拆成不可再分的最小属性，可将关系转化为 1NF

（2）第二范式：在 1NF 基础上，每个非主属性都完全依赖于主码。去掉部分依赖的方法是将主码的所有子集单独拎出作为新关系的主码，将属性以完

全依赖某一主码为标准归类，最后去掉只有主码的关系

（3）第三范式：在 2NF 基础上，所有非主属性都不传递依赖于主码。去掉传递依赖的方法是将传递的中间属性单独取出，作为新关系的主码，再将依赖于它的所有属性从原关系移至新关系

3.8 数据库保护

1. 事务：完整的工作单元，全部执行或都不执行。

（1）原子性：事务是数据库的逻辑工作单位，不可部分执行

（2）一致性：事务执行的不能使数据库变为不一致状态

（3）隔离性：事务的执行不能被其它事务干扰

（4）持久性：事务成功执行后，数据库被永久改变

2. 并发操作：同时运行的多个事务。

（1）丢失修改：另一个事务覆盖了当前事务提交结果

（2）脏数据：读取了另一事务撤销前的数据

（3）不可重复读：由于另一事务修改，前后读取数据不一致

（4）幽灵数据：由于另一数据修改，读取时多了/少了部分记录

3. 并发控制：合理调度并发操作，使一个事务的执行不受其它事务干扰

（1）共享锁（S 锁）：将数据标记为只读。其它事务可以继续加 S 锁，但不能加 X 锁，直到当前 S 锁被释放

（2）排它锁（X 锁）：允许当前事务读取和修改数据，但其它事务不能加任何锁或进行任何操作，只能进入等待状态

4. 封锁协议：为数据加锁的规则

（1）一级封锁协议：对数据加 X 锁，直到当前事务结束才释放。可防止修改，且保证事务可恢复

（2）二级封锁协议：一级封锁协议基础上，对读取的数据加 S 锁，读完后即释放。可防止读“脏数据”

（3）三级封锁协议：一级封锁协议基础上，对读取数据加 S 锁，事务结束后才释放。可保证数据可重复读取。

5. 数据库备份：对数据进行复制，保证故障后数据可恢复。应备份表、用户、全部数据和数据库日志。

6. 数据库恢复：将数据库从错误描述状态恢复到最近的正确状态。恢复方法有数据备份、事务日志、镜像技术。

3.9 数据库设计

1. 步骤：需求分析、结构设计（概念、逻辑、物理）、行为设计（功能、事务、程序）、数据库实施、运行和维护阶段。

2. 结构设计：概念结构设计 + 逻辑结构设计。

（1）概念结构设计：自底向上。先将数据抽象为实体和属性，用 E-R 图表达实体间的关联关系，再集成为全局 E-R 图。

（2）逻辑结构设计：将 E-R 图转化为关系表。之后根据函数依赖关系，对数据模型优化，并设计用户的外模式。

3. 行为设计：功能分析、功能设计、事务设计。

（1）功能分析：指出对实体的操作，并给出语义约束

（2）功能设计：设计各功能模块

（3）事务设计：输入设计、输出设计等