June, 2018

数理统计期末复习

- 基本概念

- 概率论: 随机变量, 分布函数, 大数定律, 中心极限定理
- 统计量:
  - 定义: $\{x_1, ..., x_n\}$是样本, $T_n = f(x_1, ..., x_n)$, 若f为已知函数且不依赖任何未知参数, 则T_n是统计量
  - 例子: 中位数, 次序统计量, 方差, 均值, ...
- 充分统计量
  - 定义: 若$T_n = f(X_1, ..., X_n)$是充分统计量, 则$P((X_1, ..., X_n) \le (x_1, ..., x_n) \mid T_n = t)$不依赖于$\theta$
  - 因子分解定理: 设$X = (X_1, ..., X_n) \sim f_\theta(x_1, ..., x_n)$, $T(X)$是一个统计量. 则$T(X)$是充分统计量的充分必要条件是: $f_\theta(x_1, ..., x_n)$可以分解为$f_\theta(x_1, ..., x_n) = g_\theta(T(x_1, ..., x_n))h(x_1, ..., x_n)$. 其中$h(x_1, ..., x_n)$不依赖于$\theta$, 对于连续随机变量$f_\theta(x_1, ..., x_n)$是概率密度函数, 对于离散随机变量$f_\theta(x_1, ..., x_n) = P_\theta(X_1 = x_1, ..., X_n = x_n)$.
  - 例子: 高斯分布($\sum_i x_i$ $\sum_i x_i^2$), 均匀分布($\max_i x_i$), 二项分布($\sum_i x_i$)
  - (离散)因子分解定理的证明
- 统计推断:
  - 点估计: $X \sim f_\theta(x)$, 通过样本$\{X_1, ..., X_n\}$来估计θ, $\hat{\theta} = T(X_1, ..., X_n)$称为$\theta$的点估计
  - 假设检验
- 标准

- 无偏性: 样本方差
- (强/弱)相合性: (强/弱)大数定律
  - $\bar{x} \xrightarrow{a.s.} \mu$ (证明: 大数定律)
  - $\hat{\sigma}^2 \xrightarrow{a.s.} \sigma^2$ (证明: 大数定律)
  - 样本中位数: $X_{(\frac{n}{2})} \xrightarrow{a.s.} m[X]$
    - hint: 利用$\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}} \xrightarrow{a.s.} P(X \leq x)$ (两点分布大数定律), 证明 $m[X] - \epsilon \leq X_{(\frac{n}{2})} \leq m[X] + \epsilon, \forall \epsilon > 0$
  - 最大值: $X \sim U[0, \theta] \quad \max_i x_i \xrightarrow{a.s.} \theta$
    - Borel-Cantelli Lemma: 若$\sum_{i=1}^{\infty} P(|X_i - X| \geq \epsilon) < \infty, \forall \epsilon > 0$, 则$X_n \xrightarrow{a.s.} X$
- 渐进正态性: 中心极限定理, MLE的渐近正态性, 等等

- 点估计
  - 矩方法/估计
    - 定义: 总体k阶矩是参数的函数, 于是参数可以用总体k阶矩表示. 总体k阶矩可以用样本估计,于是通过方程组可以解出参数
    - 例子: 估计高斯分布的均值和方差
  - 极大似然估计
    - 已知: $X \sim f_{\theta_0}(x) \quad (X_1, \ldots, X_n) \sim f(x_1, \ldots, x_n, \theta_0)$
    - 定义: $\hat{\theta} = \arg\max_{\theta \in \Theta} f(x_1, \ldots, x_n, \theta) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_\theta(x_i)$
    - 相合性: $\hat{\theta} \xrightarrow{a.s.} \theta_0$

- 由大数定律可得: $\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(x_i) \xrightarrow{a.s.} \mathbb{E}[\log f_\theta(x)]$
- 可以证明: $\frac{\partial}{\partial\theta}\mathbb{E}[\log f_\theta(x)]\Big|_{\theta=\theta_0} = 0$ 且 $\frac{\partial^2}{\partial\theta^2}\mathbb{E}[\log f_\theta(x)]\Big|_{\theta=\theta_0} < 0$

- 有/无偏: MLE可能是无偏估计(二项分布), 也可能是有偏估计(高斯分布的方差, 均匀分布的最大值)
- 例子: 二项分布, 均匀分布($X \sim U[0,\theta]$, $\hat{\theta} = X_{(n)}$), 高斯分布
- 渐近正态性: $\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} \mathcal{N}(0,\frac{1}{I(\theta_0)})$

- *高斯发现正态分布的过程
- 一致最小方差无偏估计(UMVUE)
  - 定义: 若$E_\theta[\hat{g}(X)] = E_\theta[g(\theta)]\ \forall\theta\in\Theta$, 则称$\hat{g}(X)$是$g(\theta)$的无偏估计; $\hat{g}(X)$是无偏估计, 且对任一无偏估计$\hat{g}_1(X)$都有$Var_\theta(\hat{g}(X)) \leq Var_\theta(\hat{g}_1(X))\ \forall\theta\in\Theta$, 则称$\hat{g}(X)$是$g(\theta)$的一致最小方差无偏估计
  - 引理: $\hat{g}(X)$是无偏估计, T是充分统计量. 定义$H := E[\hat{g}(X)\,|\,T] = G(T)$, 可以证明$H$是一统计量且还是$g(\theta)$的无偏估计, 则有$Var[H] \leq Var[\hat{g}(X)]\ \forall\theta\in\Theta$. **(此引理告诉我们在有充分统计量T时, 为求UMVUE, 只需考虑能表达为T的函数的无偏估计类)**
  - 引理的证明hint: $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$, $\mathbb{E}[\hat{g}(X)\,|\,T_n] - g(\theta) = \mathbb{E}[\hat{g}(X)-g(\theta)\,|\,T_n]$, $\mathbb{E}\big[\mathbb{E}[X\,|\,Y]\big] = \mathbb{E}[X]$
- 零无偏估计法: 验证某个特定的估计量$\hat{g}(X)$为UMVUE的工具
  - 定理: 设$\hat{g}(X)$是$g(\theta)$的一个无偏估计, $Var[\hat{g}(X)] < \infty$ $(\forall\theta\in\Theta)$, 且对任何满足条件“$\mathbb{E}[l(X)] = 0$ $(\forall\theta\in\Theta)$”的统计量$l(X)$都有$Cov[\hat{g}(X), l(X)] = 0$ $(\forall\theta\in\Theta)$, 则称$\hat{g}(X)$是$g(\theta)$的UMVUE. 并且可以说明$\hat{g}(X)$是$g(\theta)$唯一的的UMVUE.

- 定理的证明hint: $Cov[\hat{g}(X), l(X)] = \mathbb{E}[\hat{g}(X) \cdot l(X)]$, $l(X) := \hat{g}_1(X) - \hat{g}(X)$

- (补充)推论: 假设$T_n = T(X_1, ..., X_n)$是充分统计量, 设$\hat{g}(T_n)$是$g(\theta)$的一个无偏估计, $Var[\hat{g}(T_n)] < \infty$ $(\forall \theta \in \Theta)$, 且对任何满足条件"$\mathbb{E}[l(T_n)] = 0$ $(\forall \theta \in \Theta)$"的统计量$l(T_n)$都有$Cov[\hat{g}(T_n), l(T_n)] = 0$ $(\forall \theta \in \Theta)$, 则称$\hat{g}(T_n)$是$g(\theta)$的UMVUE.
  - 证明hint: 结合上述的引理和定理

- 例子:
  - 二项分布($X \sim B(p)$): $\mathbb{E}[l(T_n)] = 0 \Rightarrow l(T_n)$
  - 均匀分布($X \sim U[0,\theta]$): $\hat{\theta} := \frac{n+1}{n} \max_{1 \le i \le n} X_i$, $\mathbb{E}[l(T_n)] = 0 \Rightarrow l(T_n)$

- 完备统计量
  - 定义: 假设T是一个统计量, 若对任意不依赖于θ的函数$L(\cdot)$满足$E[L(T)] = 0, \forall \theta \in \Theta \Rightarrow L(T) = 0$, 则称T是完备统计量
  - 定理: 假设T是充分且完备的统计量, 且$\hat{g}(T)$是$g(\theta)$的无偏估计, 且$Var[\hat{g}(T)] < \infty$, 则$\hat{g}(T)$是$g(\theta)$的唯一的UMVUE
    - 证明: 构造$L(T) = \hat{g}(T) - \hat{g}_1(T)$
  - 定理: 设样本X有概率密度函数$f_\theta(x) = c(\theta)\exp\{\sum_{i=1}^{k} T_i(x)Q_i(\theta)\}h(x)$, 则统计量$T = (T_1(X), ..., T_k(X))$是充分的. 若进一步地, $\left\{ \left(Q_1(\theta), ..., Q_k(\theta)\right) : \theta \in \Theta \right\}$有内点, 则$T$是完备的
    - 例子: 可以证明$T = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$是充分完备统计量, $(\bar{x}, \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2))$是$X \sim \mathcal{N}(\mu, \sigma^2)$唯一的UMVUE
  - C-R不等式

- 若$\hat{g}(X)$是$g(\theta)$的无偏估计, 那么有$Var[g(\hat{X})] \geq \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{I(\theta)}$
- 证明hint: $\int \left(\hat{g}(x) - g(\theta)\right)\frac{\partial f_\theta(x)}{\partial \theta}dx = \frac{\partial g(\theta)}{\partial \theta}$,
$\int f \cdot g dx \leq \left(\int f^2 dx\right)^{\frac{1}{2}}\left(\int g^2 dx\right)^{\frac{1}{2}}$ (Cauchy-Schwarz Inequality)
- 意义: 给出了无偏估计的方差的下界, 定义了Fisher信息量
$I(\theta) = \mathbb{E}\left[\frac{\partial \log f_\theta(x)}{\partial \theta}\right]^2$, 注意$f_\theta(x)$为样本的联合分布概率密度函数
- 例子:
  - n个IID随机变量, $I(\theta) = nI^*(\theta)$, $I^*(\theta)$为单个样本的Fisher信息量
  - 高斯分布($X \sim \mathcal{N}(\mu, \sigma_0^2)$): 可以证明$Var[\bar{x}]$达到了C-R不等式给出的下界
  - 极大似然估计的渐近正态性 $\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, \frac{1}{I^*(\theta)})$

- 假设检验
  - 错误类型:
    - 第一类: H_0正确, 但被拒绝
    - 第二类: H_0错误, 但被接受
  - Neyman-Pearson准则: 控制第一类错误的概率($\leq \alpha$), 最小化第二类错误的概率
  - 单样本t检验
    - 假设: $X \sim \mathcal{N}(\mu, \sigma^2)$, $H_0 : \mu = 0$
    - 构造t统计量: $T_n := \frac{\sqrt{n}\bar{x}}{\hat{S}_n} \xrightarrow{d} \mathcal{N}(0,1)$ where $\hat{S}_n^{\ 2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- t随机变量: $X \sim \mathcal{N}(0,1),\ Y \sim \chi^2(n),\ t(n) := \dfrac{X}{\sqrt{\frac{Y}{n}}}$
- 可以证明在H_0成立时: $T_n \sim t(n-1)$
  - 重要结论: 假设$X \sim \mathcal{N}(\mu, \sigma^2)$, 则$\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1)$ 且 $\bar{X}$与$\sum_{i=1}^{n}(X_i - \bar{X})^2$独立
- 两样本t检验:
  - 假设: $X \sim \mathcal{N}(\mu_1, \sigma^2),\ Y \sim \mathcal{N}(\mu_2, \sigma^2),\ H_0 : \mu_1 \neq \mu_2$
  - 构造两样本t统计量:

$$T_n := \frac{\sqrt{\frac{n_1 n_2}{n_1 + n_2}}(\bar{X} - \bar{Y})}{\hat{S}_n},\ \hat{S}_n^{\ 2} = \frac{1}{n_1 + n_2 - 2}\left\{\sum_{i=1}^{n_1}(x_i - \bar{x})^2 + \sum_{i=1}^{n_2}(y_i - \bar{y})^2\right\}$$

  - 可以证明在H_0成立时: $T_n \sim t(n_1 + n_2 - 2)$
    - χ^2分布的性质: $X \sim \chi^2(n), Y \sim \chi^2(m), X \perp Y \Rightarrow X + Y \sim \chi^2(m+n)$
    - $\dfrac{1}{\sigma}\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}(\bar{X} - \bar{Y}) \sim \mathcal{N}(0,1),\ \dfrac{1}{\sigma}\hat{S}_n^2 \sim \chi^2(n_1 + n_2 - 2)$
- 相关性检验
  - 假设: $X, Y \quad H_0 : \rho(X, Y) = 0$
  - 样本协方差: $\hat{\rho}_n = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad \hat{\rho}_n \xrightarrow{a.s.} \rho$
  - 在H_0成立时: $\sqrt{n}\hat{\rho}_n \xrightarrow{d} \mathcal{N}(0,1)$

- χ2检验: is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true.
  - Pearson's χ^2 test for Multinomial data
    - 不含未知参数
      - 假设: $H_0 : P_j = P_{0j}\ (0 \le j \le r) \quad \{P_{0j}\}_{j=0}^r$已知
      - 构造: $T_n = \sum_{j=0}^{r} \frac{(n_j - nP_{0j})^2}{nP_{0j}}$ where $n_j = \sum_{i=1}^{n} 1_{\{x_i = j\}}$
      - 定理: 在H_0成立时: $T_n \xrightarrow{d} \chi^2(r)$
    - 含未知参数
      - 假设: $H_0 : P_j = P_{0j}(\theta_1, \ldots, \theta_d)$
      - 构造: $T_n = \sum_{j=0}^{r} \frac{\left(n_j - nP_{0j}(\hat{\theta}_1, \ldots, \hat{\theta}_d)\right)^2}{nP_{0j}(\hat{\theta}_1, \ldots, \hat{\theta}_d)}$ where
        $(\hat{\theta}_1, \ldots, \hat{\theta}_d) = \arg\max_{\theta_1, \ldots, \theta_d} \prod_{i=1}^{n} (P_{0j}(\theta_1, \ldots, \theta_d))^{x_i} = \arg\max_{\theta_1, \ldots, \theta_d} \prod_{j=0}^{r} (P_{0j}(\theta_1, \ldots, \theta_d))^{n_j}$
      - 在H_0成立时: $T_n \xrightarrow{d} \chi^2(r-d)$
      - 例子(χ^2独立性检验):
        - $H_0 : P(A = i, B = j) = P(A = i)P(B = j)$
        - $T_n = \sum_{i=0}^{r} \sum_{j=0}^{s} \frac{(n_{ij} - n\hat{P}(A = i)\hat{P}(B = j))^2}{n\hat{P}(A = i)\hat{P}(B = j)}$
          $\hat{P}(A = i) = \frac{\sum_j n_{ij}}{n}, \quad \hat{P}(B = j) = \frac{\sum_i n_{ij}}{n}$
      - 在H_0成立时: $T_n \xrightarrow{d} \chi^2\big((r+1)(s+1) - 1 - (r+s)\big) = \chi^2(rs)$

- 连续随机变量
  - 假设: $X \sim f(x) \quad H_0 : f(x) = f_0(x),\ f_0(x)$已知
  - 另一种方法: 经验分布函数 $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}}$

- 非参数检验
  - 置换检验(两样本问题)
    - 假设: $X \sim F_X, Y \sim F_Y \quad H_0 : F_X = F_Y$
    - 样本: $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$
    - 构造: $T = \bar{X}_n - \bar{Y}_m$
    - $P(|T| \ge c | H_0) = \frac{\sum_{i=1}^{N} 1_{\{|H_i| \ge c\}}}{N}$ where $N = C_{n+m}^{n}$
    - 当n和m较大时, 置换检验不再适用
  - 秩和检验(两样本问题)
    - 假设: $X \sim F_X, Y \sim F_Y \quad H_0 : F_X = F_Y$
    - 样本: $\{X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}\}$
    - 统计量: 排序后获得$\{X_1, \ldots, X_{n_1}\}$的序$\{R_1, \ldots, R_{n_1}\}$, 构造$T = \sum_{i=1}^{n_1} R_i$
    - 根据大样本理论, 在H_0成立时: $\frac{T - \frac{1}{2}n_1(n_1 + n_2 + 1)}{\sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}} \xrightarrow{d} \mathcal{N}(0,1)$

- $\mathbb{E}[T] = \sum_{i=1}^{n_1} \mathbb{E}[R_i] = \sum_{i=1}^{n_1} \sum_{k=1}^{n_1+n_2} k \cdot P(R_i = k)$

- $Var[T] = \sum_{i=1}^{n_1} Var[R_i] + \sum_{i \neq j} Cov[R_i, R_j]$
- $Var[R_i] = \mathbb{E}[R_i^2] - \mathbb{E}^2[R_i] = \frac{1}{12}(N^2 - 1),$
  $Cov[R_i, R_j] = -\frac{N+1}{12} \ (i \neq j)$
- 证明参考: https://stats.stackexchange.com/a/316208/169345

• 似然比检验
  - 假设: $X \sim f_\theta(x), \theta \in \Theta \quad H_0 : \theta \in \Theta_0$
  - 统计量: $T_n = \frac{\max_{\theta \in \Theta} \prod_{i=1}^n f_\theta(x_i)}{\max_{\theta \in \Theta_0} \prod_{i=1}^n f_\theta(x_i)}$
  - 例1: $X \sim \mathcal{N}(\mu, \sigma^2) \quad H_0 : \mu = 0$
  - 例2: $X \sim \mathcal{N}(\mu, \sigma^2) \quad H_0 : \sigma^2 = \sigma_0^2$
  - (补充) Wilks's Theorem: If $dim(\Theta) - dim(\Theta_0) = t > 0$ and H_0 holds, then $\lim_{n \to \infty} 2 \ln T_n \sim \chi^2(t)$

• 多重检验: 同时进行多个假设检验时,需要修改拒绝域
  - FWER(组间误差): $P(\cup_{i \in H_0} \{ |T_i| \geq t \}) \leq \sum_{i \in H_0} P(|H_i| \geq t) \leq m \cdot P(|H_i| \geq t)$
    表示第一类错误至少出现一次(rejecting at least one true null hypothesis)的概率
  - Bonferroni Method(控制FWER $\leq \alpha$): $P(|H_i| \geq t) \leq \frac{\alpha}{m}$ 缺点:过于保守,几乎不会拒绝原假设
  - FDP(错误发现比例): $FDP = \frac{\sum_{i \in H_0} 1_{\{|T_i| \geq t\}}}{\max\{ \sum_{i=1}^m 1_{\{|T_i| \geq t\}}, \ 1\}}$

- FDR(错误发现率): $FDR = \mathbb{E}\left[FDP = \frac{\sum_{i \in H_0} 1_{\{|T_i| \geq t\}}}{\max\{\sum_{i=1}^{m} 1_{\{|T_i| \geq t\}}, 1\}}\right]$

• p-value

  - 定义: $p := 1 - F(|T|)$ 其中$F(\cdot)$是在原假设成立情况下的cdf (注: 该定义并非一般性定义)
  - $\hat{t} = \sup\left\{0 < t < 1 : \frac{mt}{\max\{\sum_{i=1}^{m} 1_{\{p_i \leq t\}}, 1\}} \leq \alpha\right\}$, 若$p_i \leq \hat{t}$, 则拒绝$H_{0i}$
  - 对比: $\hat{t} = \inf\left\{t > 0 : \frac{mP(|\mathcal{N}(0,1)| \geq t)}{\max\{\sum_{i=1}^{m} 1_{\{p_i \leq t\}}, 1\}} \leq \alpha\right\}$, 若$|T_i| \geq \hat{t}$, 则拒绝$H_{0i}$
  - B-H method(控制$FDR \leq \alpha$)
    • $p_{(1)} < p_{(2)} < \cdots < p_{(m)}$
    • $\hat{k} = \max\{k : p_{(k)} \leq \frac{\alpha k}{m}\}$
  - 若$p_i \leq p_{(\hat{k})}$, 则拒绝H_{0i}
  - (补充定理)不论原假设为真的数量m_0有多少, 不管原假设为假的p值的分布如何, B-H方法总可以控制 $FDR \leq \frac{m_0}{m}\alpha \leq \alpha$.

- 多元分析

  • 多元正态分布

    - 定义1: 若 $f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$ 则称 $X \sim \mathcal{N}(\mu, \Sigma)$
    - 定义2: 若对 $\forall t \in \mathfrak{R}^m$ 有 $t^T x \sim \mathcal{N}(t^T\mu, t^T\Sigma t)$, 则称 $X \sim \mathcal{N}(\mu, \Sigma)$
    - 可以证明: 定义1 $\Longleftrightarrow$ 定义2

- 证明定义2⇒定义1的步骤如下:
  - 若$X \sim \mathcal{N}(\mu, \Sigma)$, 则 $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$, 更一般地, $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$
  - 若$X_i, X_j \sim \mathcal{N}$且$cov(X_i, X_j) = 0$, 则$X_i$与$X_j$独立. 更一般地, 对于MVN, 两组变量不相关⇔两组变量独立 (证明: 随机变量的特征函数$f(t) = \mathbb{E}[e^{it^T x}]$, $\mathbb{E}[e^{it^T X}] = \mathbb{E}[e^{it_{S_1}^T X_{S_1}}]\mathbb{E}[e^{it_{S_2}^T X_{S_2}}] \iff X_{S_1} \perp X_{S_2}$)
  - 由上面两个结论可得, 若$X \sim \mathcal{N}(\mu, \Sigma)$, 则$\Sigma^{-\frac{1}{2}}(X - \mu) \sim \mathcal{N}(0,I)$
  - 概率变换公式: 已知$X \sim f_X(x), Y = g(X)$, 则 $f_Y(y) = f_X(g^{-1}(y))\left|\dfrac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)}\right|$

- 数据的分类:
  - 问题: $X \sim f_1, Y \sim f_2$, 找到$\mathfrak{R}^p$的一个划分$R_1 \cup R_2 = \mathfrak{R}^p, R_1 \cap R_2 = \varnothing$, 当$Z \in R_1$时认为$Z \in X$, 当$Z \in R_2$时认为$Z \in Y$.
  - 错误分类概率:

$$\Delta = P(Z \in R_2 | Z \sim f_1) + P(Z \in R_1 | Z \sim f_2) = \int_{R_2} f_1(x)dx + \int_{R_1} f_2(x)dx$$

$$= 1 - \int_{R_1} (f_1(x) - f_2(x))dx$$

  - 目标: $\min_{R_1} \Delta \iff \max_{R_1} \int_{R_1} (f_1(x) - f_2(x))dx$
  - 最优解: $R_1 = \{x | f_1(x) \geq f_2(x)\}$
  - Fisher线性判别式(LDA)
    - 假设: $X \sim \mathcal{N}(\mu_1, \Sigma), Y \sim \mathcal{N}(\mu_2, \Sigma) \in \mathfrak{R}^p$

- $\{x : \frac{f_1(x)}{f_2(x)} \geq 1\} = \{x : (x - \frac{\mu_1 + \mu_2}{2})^T \Sigma^{-1}(\mu_1 - \mu_2)\}$
- 判别向量: $\beta := \Sigma^{-1}(\mu_1 - \mu_2)$ 可解线性方程组$\Sigma\beta = (\mu_1 - \mu_2)$得到
- $\hat{\mu}_1 = \bar{X},\ \hat{\mu}_2 = \bar{Y},\ \hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}\{\sum_{i=1}^{n_1}(X_i - \bar{X})(X_i - \bar{X})^T + \sum_{i=1}^{n_2}(Y_i - \bar{Y})(Y_i - \bar{Y})^T\}$

- 二次判别式(QDA)
    - 假设: $X \sim \mathcal{N}(\mu_1, \Sigma - 1), Y \sim \mathcal{N}(\mu_2, \Sigma_2) \in \mathfrak{R}^p$
    - $\hat{\Sigma}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \bar{X})(X_i - \bar{X})^T,\ \hat{\Sigma}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \bar{Y})(Y_i - \bar{Y})^T$

- 最优预测 (补充: 也叫最小均方误差预测)
    - 定义: $\hat{f} = \arg\min_f \mathbb{E}\left[(Y - f(X))^2\right]$
    - 可以证明, 最优预测为 $\hat{f}(X) = \mathbb{E}[Y \mid X]$ (hint: $\mathbb{E}[Y - f(X)]^2 = \mathbb{E}\left[Y - \mathbb{E}[Y \mid X] + \mathbb{E}[Y \mid X] - f(X)\right]^2$, $\mathbb{E}[AB] = \mathbb{E}[\mathbb{E}[AB \mid X]]$)
    - 例子: 假设$X \sim \mathcal{N}(\mu = 0, \Sigma)$, 可以证明存在$\{\beta_i\}_{i=2}^p$使得 $X_1 = \sum_{i=2}^p \beta_i X_i + \epsilon_1, \quad \mathbb{E}[\epsilon_1] = 0, \quad \epsilon_1 \perp X_i\ (2 \leq i \leq p)$. 在这种情况下有 $\mathbb{E}[X_1 \mid X_2, \ldots, X_p] = \sum_{i=2}^p \beta_i X_i$
        - 证明hint: $\beta_i = -\frac{w_{1i}}{w_{11}}$ or $\beta_i = w_{1i}$, $\mathbb{E}[X_i\epsilon_1] = 0 \iff X_i \perp \epsilon_1\ (2 \leq i \leq p)$

- 线性回归
    - 模型: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$ or $Y = X\beta + \epsilon$ where $\epsilon \perp X_1, \ldots, X_p$, $\mathbb{E}[\epsilon] = 0$ and $X = \begin{bmatrix} 1 & X_1 & \cdots & X_p \end{bmatrix}$

- Y: 响应变量, X: 协变量, β: 回归系数
- 目标: 估计参数$\beta_0, \ldots, \beta_p$
- β的最小二乘估计: $\hat{\beta} = (X^T X)^{-1} X^T Y$
- 假设$\epsilon \sim \mathcal{N}(0,\sigma^2)$:
  - 则$\hat{\beta} - \beta = (X^T X)^{-1} X^T \epsilon \sim \mathcal{N}(0,\ \sigma^2 (X^T X)^{-1})$
  - 定义: 残差: $\hat{\epsilon} = \begin{bmatrix} Y_1 - X_1\hat{\beta} \\ \vdots \\ Y_n - X_n\hat{\beta} \end{bmatrix}$, $\hat{\sigma}^2 = \dfrac{1}{n-p-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \|\hat{\epsilon}\|^2$
  - 可以证明:
    - $\hat{\sigma}^2 \sim \dfrac{\sigma^2}{n-p-1} \chi^2(n-p-1)$
      - 证明hint: $\hat{\epsilon} = H\epsilon$ where $H := I_n - X(X^T X)^{-1} X^T$ is idempotent; SVD: $H = UDU^T$; singular values of idempotent matrix is 0 or 1; sum of eigenvalues of a square matrix A $\sum_i \lambda_i = tr(A)$; $tr(AB) = tr(BA)$
    - $\hat{\sigma}$与$\hat{\beta}$独立 ($cov[\hat{\epsilon}, \hat{\beta}] = \mathbb{E}[\hat{\epsilon}\hat{\beta}^T] - \mathbb{E}[\hat{\epsilon}]\mathbb{E}[\hat{\beta}]^T = 0$)
- 假设检验: $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 w_{ii})$, $H_0 : \beta_i = 0 \quad H_1 : \beta_i \neq 0$
  - 统计量: $T_i = \dfrac{|\hat{\beta}_i|}{\hat{\sigma}\sqrt{w_{ii}}}$
  - 可以证明在H_0成立时: $T_i \sim |t(n-p-1)|$
  - 多重假设检验(FWER/FDR): $H_{0i} : \beta_i = 0 \quad (1 \leq i \leq p)$
  - 似然比检验:
    - 假设: $H_0 : \beta_s = 0,\ s \subset \{1,2,\ldots,p\}$

- 统计量: $T_n = \dfrac{\max\limits_{\beta \in \Re^{p+1}} L(\beta | X, Y)}{\max\limits_{\beta_s = 0} L(\beta | X, Y)}$

- 条件相关性
  - $cov[X_1, X_2 | X_3] = \mathbb{E}[X_1 X_2 | X_3] - \mathbb{E}[X_1 | X_3]\mathbb{E}[X_2 | X_3]$
  - 设$X \sim \mathcal{N}(\mu, \Sigma)$, $\Sigma^{-1} = [w_{ij}] \in \Re^{p \times p}$, 则$cov[X_i, X_j | X_{k \neq i,j}] = 0 \iff w_{ij} = 0$
- 逻辑回归
  - 模型: $P(Y = 1 | X_1, ..., X_p) = \dfrac{1}{1 + \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\}}$
  - 概率解释: 混合高斯分布
- 分位数回归
  - 问题: 给定α, 求y, 使得$P(Y \leq y | X) = \alpha$
  - 模型(线性假设): $y = \beta_0(\alpha) + \beta(\alpha)^T X$
  - 损失函数: $\rho_\alpha(x) := (\alpha - 1_{\{x<0\}}) \cdot x$
  - 优化目标: $\hat{\beta}_0, \hat{\beta} = \arg\min\limits_{\beta_0, \beta} \dfrac{1}{n} \sum\limits_{i=1}^{n} \rho_\alpha(y_i - \beta_0 - \beta^T x_i)$
  - $\sqrt{n}\left[(\hat{\beta}, \hat{\beta}) - (\beta, \beta)\right] \xrightarrow{d} \mathcal{N}(0,A)$
  - 中位数回归: $\alpha = \dfrac{1}{2}$, 绝对值损失
- 支持向量机: $\min\limits_{\beta_0, \beta} \mathbb{E}\left(1 - Y(\beta_0 + \beta^T X)\right)_+$ and

  $$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \frac{1}{n} \sum_{i=1}^{n} \left(1 - Y(\beta_0 + \beta^T X)\right)_+$$
- 主成分分析

- 主成分等于Σ的特征向量, 方差等于Σ的特征值
- $\dfrac{Var(Y_1)}{\sum_i Var(Y_i)} = \dfrac{\lambda_1}{\sum_i \lambda_i}$
- 稀疏主成分分析: 当p很大时, Σ是奇异的, 不能作出很好的的估计. 实际上并不是每个位置都要估计, 对于$u = (u_1 \quad \cdots \quad u_p)^T$, 大部分$u_i = 0$.
  $\max\limits_v \quad v^T\Sigma v \quad s.t. \ \|v\|_2 = 1, \ \|v\|_0 \le k$

- 优化问题:
  - 目标: $\hat{\theta} = \arg\min\limits_{\theta\in\Re^p} \dfrac{1}{n}\sum_{i=1}^{n} f(x_i,\theta), f(\cdot)$是已知的凸的损失函数
  - 估计方法1:
    - $\theta_{t+1} = \theta_t - A^{-1}\dfrac{1}{n}\sum_i \dfrac{\partial}{\partial\theta}f(x_i,\theta)$ where $A = \dfrac{1}{n}\sum_i \dfrac{\partial^2}{\partial\theta^2}f(x_i,\theta)$
    - 复杂度: $O(n(p^2+p)+p^3)$
  - 估计方法2(随机梯度下降):
    - $\theta_{t+1} = \theta_t - n_t\dfrac{\partial}{\partial\theta}f(x_{I_t},\theta)$ where $I_t \sim U(\{1,2,...,n\})$, n_t是步长
    - 复杂度: $O(p)$
    - 可证明: 当n_t合理时, $|\theta_n - \hat{\theta}| = O(\dfrac{1}{\sqrt{n}})$
- 贝叶斯统计:$f(\theta|X) = \dfrac{f(X|\theta)f(\theta)}{\sum_{\theta_i\in\Theta} f(X|\theta_i)f(\theta_i)}$