均值：平均水平，缺乏稳健性　　经验分布：分布与位置

中位数：中心位置，稳健性　　$X_{([np]+1)}$　np非整

三均值 = $\frac{1}{4}$(上四 + 下四) + $\frac{1}{2}$中　　$\frac{1}{2}(X_{(np)}+X_{(np+1)})$　整
（上四 Q_3，下四 Q_1，中 M）

分散性
- 方差：有量纲　　变异系数：相对分散性
- 极差，四分位极差 $R_1 = Q_3 - Q_1$　　无量纲
- 　　更稳健　　$CV = \frac{S}{\bar{x}}$　越大越波动
- 上下截断点（异常值）
- $Q_1 - 1.5R_1$，$Q_3 + 1.5R_1$
- 下四分位数　上四分位数

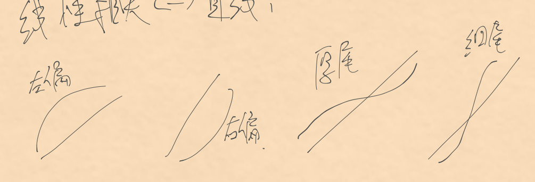
形状
- 偏度　<0左　>0右　　峰度　<0短尾　>0粗
- 哪侧更分散　　以正态为准，比较尾部分散性
- $G_1 = \frac{\mu_3}{\sigma^3}$　　$G_2 = \frac{\mu_4}{\sigma^4} - 3$
- μ_i：i阶中心矩

直方图：组距～频率　　经验分布函数　$\frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$

QQ图：(x, y)　　相同分位数

正态QQ（理，观测）　$y = \sigma x + \mu$

线性相关 ⟺ 直线



χ^2检验　　分区间

$$\chi^2 = \sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i} \sim \chi^2(l-k-1)$$，k为分布中待估参数个数　Reject for large　$H_0: F_X \equiv F_0$

m_i：观测值，np_i：理论频数

经验分布拟合检验

正态性W检验

$x_1, \dots, x_n$ 观测

$x_{(1)}, \dots, x_{(n)}$ 次序

d_i

$$W = \frac{(\sum a_i d_i)^2}{\sum (x_i - \bar{x})^2}$$

相关：相互关系密切程度。　相关 ⇐ 因果

回归：用函数表示因果。

观测数据矩阵　二维数据

$$X = \begin{bmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{bmatrix}^T$$

$$S = \begin{bmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{bmatrix}$$　$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$　$S_{xy}^2 \le S_{xx}S_{yy}$

Pearson 相关系数　$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

Spearman

R, S 为秩：第几个

$$q_{xy} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}}$$　$\bar{S} = \bar{R} = \frac{n+1}{2}$

- =0：线性不相关
- (0,1)：线性正(负)相关
- 绝对值为1：完全线性相关

多维

$(x_1, \dots, x_p)^T$　p维总体

数据为　$(x_{11}, \dots, x_{1p})^T$ x_1，$(x_{21}, \dots, x_{2p})^T$ x_2，…，$(x_{n1}, \dots, x_{np})^T$ x_n　　$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$

$$S = \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & & \ddots & \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

第j, k列的相关系数

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}}$$

Pearson 相关矩阵

$R = \{r_{jk}\}_{p\times p}$　$R = D^{-1}SD^{-1}$，$D = Diag(\sqrt{S_{11}}, \dots, \sqrt{S_{pp}})$

Spearman

$Q = \{q_{jk}\}$

标准化后

$$R = \frac{1}{n-1}\sum_{i=1}^{n} x_i^* x_i^{*T} = \frac{1}{n-1}(X^*)^T X^*$$

总体的数字特征

$\Sigma = COV(X) = E(X-\mu)(X-\mu)^T$

$\sigma_{jk} = COV(X_j, X_k) = E(X_j - \mu_j)(X_k - \mu_k)^T$

$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$　相关系数

随机向量性质

$E(AX) = AEX = A\mu$

$COV(AX) = ACOV(X)A^T = A\Sigma A^T$

$Ec^TX = c^T\mu$　$Var(c^TX) = c^T\Sigma c$

$COV(AX, BY) = ACOV(X, Y)B^T$

$COV(X, Y) = E[(X - EX)(Y - EY)^T]$

$COV(c^TX, d^TY) = c^T COV(X, Y) d$

模型

$Y = f(X_1, \dots, X_{p-1}) + \varepsilon$

$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \varepsilon$

$Y = X\beta + \varepsilon$　$\varepsilon \sim N(0, \sigma^2 I)$

估计

$\hat{\beta} = (X^TX)^{-1}X^TY$　$\hat{Y} = X(X^TX)^{-1}X^TY$（H）

$\hat{\varepsilon} = (I - H)Y$

$SSE = \varepsilon^T(I-H)\varepsilon = Y^T(I-H)Y = \sum \hat{\varepsilon}_i^2 = \hat{\varepsilon}^T\hat{\varepsilon}$

$E(SSE) = \sigma^2(n-p)$　证明取trace

$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{1}{n-p}Y^T(I-H)Y$

性质：

① $\hat{\varepsilon}$，$\hat{\beta}$ 无偏

② $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

$\hat{\beta} = (X^TX)^{-1}X^T(X\beta + \varepsilon) = \beta + (X^TX)^{-1}X^T\varepsilon$

$COV(\hat{\beta}) = E[(X^TX)^{-1}X^T\varepsilon\varepsilon^TX(X^TX)^{-1}] = \sigma^2(X^TX)^{-1}$

③ $\frac{SSE}{\sigma^2} \sim \frac{n-p}{\sigma^2}\hat{\sigma}^2 \sim \chi^2_{n-p}$

$SSE = \varepsilon^T(I-H)\varepsilon = \varepsilon^T P^T(I_r)P\varepsilon$

P正交，故 $P\varepsilon \sim N(0, \sigma^2 I)$，故 $SSE \sim \chi^2_{n-p}$

④ $\hat{\beta}$ 与 $\hat{\sigma}$ 独立

对称证

⇔ 验证 $(X^TX)^{-1}X^T(I-H) = 0$

⇔ 证 HY 与 (I-H)Y 独立

⑤ $\hat{\varepsilon} \sim N(0, \sigma^2(I-H))$

中心化，标准化

Gauss-Markov BLUE

对 $c^T\beta$，$c^T\hat{\beta}$ 是LS估计

$E(c^T\hat{\beta}) = c^T\beta$

$cov(c^T\hat{\beta}) = \sigma^2 c^T(X^TX)^{-1}c$

Gauss-Markov 定理

证：$a^TX = c^T$

$Var(a^Ty) = \sigma^2\|a\|^2$

$\|a\|^2 = \|a - X(X^TX)^{-1}c + X(X^TX)^{-1}c\|^2$

$= \|a - X(X^TX)^{-1}c\|^2 + \|X(X^TX)^{-1}c\|^2 + 2c^T(X^TX)^{-1}X^T(a - X(X^TX)^{-1}c)$

$= \|a - X(X^TX)^{-1}c\|^2 + \frac{Var(c^T\hat{\beta})}{\sigma^2}$

故 $Var(a^Ty) \ge Var(c^T\hat{\beta})$

取等条件 ⇔ $a = X(X^TX)^{-1}c$

$a^Ty = c^T\hat{\beta}$

检验 预测

SS统计量

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = Y^T(I - \frac{1}{n}J)Y$　$J = 11^T$

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = Y^T(I-H)Y$

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = Y^T(H - \frac{1}{n}J)Y$　$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i$

SST = SSE + SSR

拟合优度 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

全局显著性

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} \sim F(p-1, n-p)$$

Reject for large F.

方差分析表

	DF	SS	MS	F	p值
R	p-1	$\sum(\hat{y}_i - \bar{y})^2$	SSR/p-1	MSR/MSE	p_0
E	n-p	$\sum(y_i - \hat{y}_i)^2$	SSE/n-p		
T	n-1	$\sum(y_i - \bar{y})^2$			

对单一系数的估计与检验

从 $\hat{\beta}_k$ 估 β_k　C_{kk} 为 $(X^TX)^{-1}$ 主对角线上第k个元素

$t_k = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{C_{kk}}} \sim t(n-p)$

$p_0 = 2P(t(n-p) > |t_k|)$　H_0 成立

置信区间 α　$\hat{\beta}_k \pm \hat{\sigma}\sqrt{C_{kk}}\, t_{1-\alpha/2}(n-p)$

预测　对 y_0 预测

$\hat{y}_0 \sim N(x_0^T\beta, \sigma^2 x_0^T(X^TX)^{-1}x_0)$

Plug-in $\hat{\sigma}$，studentized

证，$y_0 = x_0^T\beta + \varepsilon_0$，$\varepsilon_0 \sim N(0, \sigma^2)$

$\hat{y}_0 = x_0^T[(X^TX)^{-1}X^T(X\beta + \varepsilon)] = x_0^T\beta + (X^TX)^{-1}X^T\varepsilon$　故 y_0，$\hat{y}_0$ 独立

$E y_0 - \hat{y}_0 = 0$，$Var(y_0 - \hat{y}_0) = \sigma^2(1 + x_0^T(X^TX)^{-1}x_0)$

$\hat{\beta}$ 与 $\hat{\sigma}$ 独立

$y \in y_0 \pm t_{1-\alpha/2}(n-p)\sqrt{\hat{\sigma}^2[1 + x_0^T(X^TX)^{-1}x_0]}$

一般检验（显著性检验）

等约简模型　SSE(F)　f_F = n-全模型中参数个数

SSE(R)　f_R = n-约简模型中参数个数

$$F = \frac{[SSE(R) - SSE(F)]/(f_R - f_F)}{SSE(F)/f_F} \sim F(f_R - f_F, f_F)$$

残差检验

学生化残差　$\hat{\varepsilon}_i \sim N(0, \sigma^2(1 - h_{ii}))$

$r_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSE(1 - h_{ii})}} \to N(0,1)$

$h_{ii} = x_i^T(X^TX)^{-1}x_i$

box-cox

$$y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \ne 0 \\ \ln Y & \lambda = 0 \end{cases}$$

变量选取

穷举法：拟合所有，p较小时可行。→ $R_a^2(p)$ 最大的 → 同数量里 $R^2(p)$ 最大 → 3个 …；→ C_p 算 C_p，作 C_p 图，选最接近 $C_p = p$ 的；→ PRESS，剔除i后预测，算误差平方和

逐步回归：拟合部分，p较大可行

① 加自变量，选SSE减少最多的变量

② 若加进去后自变量之前的统计量

③ 变量剔除

↳ 每次一步到，后一个或前一个上添加(剔除)

A 已有 l-1 个自变量回归模型

$$F_{X_k} = \frac{SSE(A) - SSE(A, X_k)}{MSE(A, X_k)/(n-l-1)} \sim F$$

LOGISTIC

因子 $(x_1, \dots, x_{p-1})$ 的 m组取值　$x_k = (x_{k1}, \dots, x_{k,p-1})^T$　k=1, …, m

对A（Bernoulli 变量Y）进行 n_k 次独立观测

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} \quad \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & & x_{2,p-1} \\ \vdots & & & & \vdots \\ 1 & x_{m1} & \cdots & & x_{m,p-1} \end{pmatrix} \quad N = \begin{bmatrix} n_1\pi(x_1) \\ n_2\pi(x_2) \\ \vdots \\ n_m\pi(x_m) \end{bmatrix}$$

$Y_k \sim b(n_k, \pi(x_k))$　自变量为 x_k 时，Y=1的概率

$\ln\left(\frac{\pi(x_k)}{1 - \pi(x_k)}\right) = \tilde{x}_k^T\beta$，k=1, …, m

$\pi(x_k) = \frac{\exp(\tilde{x}_k^T\beta)}{1 + \exp(\tilde{x}_k^T\beta)}$，k=1, 2, …, m

似然方程，$X^TN = X^TY$

$W = Diag[n_1\pi(x_1)(1 - \pi(x_1)), \dots, n_m\pi(x_m)(1 - \pi(x_m))]$

$I(\beta) = X^TWX$

$\beta^{(t+1)} = \beta^{(t)} - (I(\beta))^{-1}(X^T(Y - N^{(t)}))$

推断　$H_0: \beta_1 = \dots = \beta_r = 0$

① 全模型 $X_1, \dots, X_{p-1}$　与约减模型

$K^2 = 2\ln\frac{L(\hat{\beta}; y_1, \dots, y_m)}{L(\hat{\beta}_{H_0}; y_1, \dots, y_m)} \sim \chi^2(r)$

reject for large K^2

② Wald 统计量

$\beta_i = 0$　　对角线估 $s(\beta)$

$\hat{\beta} - \beta \sim N(0, I^{-1}(\beta))$

$W_i = \left(\frac{\hat{\beta}_i}{s(\hat{\beta}_i)}\right)^2 \sim \chi^2(1)$

方差分析

$\mu_1 = \mu_2 = \mu_3$　$MS_{组} = MS_{组内}$

不等　$MS_{组} >> MS_{组内}$

因素：对Y影响的定性变量，不同水平

因子效应：

交互作用

模型

$y_{ij} = \mu + \delta_i + \varepsilon_{ij}$　j=1, 2, …, n_i　样本；i=1, 2, …, a　水平

$\varepsilon_{ij} \sim N(0, \sigma^2)$ 独立

$\sum_{i=1}^{a} n_i\delta_i = 0$

记号　$n = \sum_{i=1}^{a} n_i$　$\bar{\varepsilon}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i}\varepsilon_{ij} \sim N(0, \frac{\sigma^2}{n_i})$

$\mu = \frac{1}{n}\sum_{i=1}^{a} n_i\mu_i$　$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{a} n_i\bar{\varepsilon}_{i\cdot} \sim N(0, \frac{\sigma^2}{n})$

$\delta_i = \mu_i - \mu$　$\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \mu + \delta_i + \bar{\varepsilon}_{i\cdot}$

$\bar{y} = \mu + \bar{\varepsilon}$

$SST = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$　总平方和

$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$　误差平方和

$SSA = \sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y})^2$　因素A

SST = SSE + SSA

$E(SSE) = \sum_{i=1}^{a} E((n_i - 1)S_i^2) = (n-a)\sigma^2$

$E(SSA) = E\left[\sum_{i=1}^{a} n_i(\delta_i + \bar{\varepsilon}_{i\cdot} - \bar{\varepsilon})^2\right] = (a-1)\sigma^2 + \sum_{i=1}^{a} n_i\delta_i^2$

故 $\frac{SSE}{n-a} \to \sigma$，但当 H_0 真 $\frac{SSA}{a-1} \to \delta$

⇒ $F = \frac{SSA/(a-1)}{SSE/(n-a)} \sim F(a-1, n-a)$

$p = P(F(a-1, n-a) > f)$

方差分析表

来源	自由度	平方和	均方	F
A	a-1	$\sum_{i=1}^{a} n_i(\bar{y}_{i\cdot} - \bar{y})^2$	SSA/a-1	$f = \frac{MSE}{MSA}$
误差	n-a	$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$	SSE/n-a	
总和	n-1			

$\bar{y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{n_i})$　估计 μ_i

$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \sim N(\mu_i - \mu_j, (\frac{1}{n_i} + \frac{1}{n_j})\sigma^2)$　估计

用MSE替代构造区间

Bonferroni：$t_{1-\frac{\alpha}{2}}(n-a) \to t_{1-\frac{\alpha}{2m}}(n-a)$　$m = C_a^2$

两因素　先交互，后单个

$y_{ijk} \sim N(\mu_{ij}, \sigma^2)$　k=1, 2, …, c

$\mu = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$　总平均

$\mu_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$，$\alpha_i = \mu_{i\cdot} - \mu$ → A效应

$\mu_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$　$\beta_j = \mu_{\cdot j} - \mu$ → B效应

$\gamma_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu$

$= \mu_{ij} - \mu - (\mu_{i\cdot} - \mu) - (\mu_{\cdot j} - \mu)$

$= (\mu_{ij} - \mu) - (\alpha_i + \beta_j)$　A与B交互效应

$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$（$\mu_{ij}$）

i=1, 2, …, a；j=1, 2, …, b；k=1, 2, …, c

$\sum_{i=1}^{a}\alpha_i = 0$　$\sum_{j=1}^{b}\beta_j = 0$

$\sum_{i=1}^{a}\gamma_{ij} = \sum_{j=1}^{b}\gamma_{ij} = 0$

表：

来源	自由度	平方和	均方	F值	P值
A	a-1	SS_A	MS_A	MS_A/MS_E	p_A
B	b-1	SS_B	MS_B	MS_B/MS_E	p_B
交互	(a-1)(b-1)	SS_{AB}	MS_{AB}	MS_{AB}/MS_E	p_{AB}
误差	ab(c-1)	SS_E	MS_E		
总和	abc-1	SS_T			

$SS_A = bc\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y})^2$

$SS_B = ac\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y})^2$

$SS_{AB} = c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y})^2$

$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(y_{ijk} - \bar{y}_{ij\cdot})^2$

主成分分析

$COV(X) = \Sigma$　$Y = a^TX$

$\max_{a_1} Var(Y_1) = a_1^T\Sigma a_1$，s.t. $a_1^Ta_1 = 1$ ⇒ $Y_1 = a_1^TX$

$\max_{a_k} a_k^T\Sigma a_k$，s.t. $a_k^Ta_k = 1$，$a_k^T\Sigma a_i = 0$ (i=1, 2, …, k-1) ⇒ $Y_k = a_k^TX$

求 Σ 的特征值与相应正交单位化特征向量

$\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_p \ge 0$

$e_1, e_2, \dots, e_p$

$Y_k = e_k^TX$，$Var(Y_k) = e_k^T\Sigma e_k = \lambda_k$

$COV(Y_j, Y_k) = e_j^T\Sigma e_k = 0$　$COV(Y) = Diag(\lambda_1, \dots, \lambda_p)$

总方差　$\sum_{k=1}^{p} Var(Y_k) = \sum_{k=1}^{p}\lambda_k = tr(\Sigma) = \sum_{k=1}^{p} Var(X_k)$

贡献率　$\frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i} = \frac{Var(Y_k)}{\sum_{i=1}^{p} Var(X_i)}$　前m个累计贡献率

主成分与 X_j 的相关系数

$Y = P^TX$ ⇒ $X = PY$

$X_j = e_{1j}Y_1 + e_{2j}Y_2 + \dots + e_{pj}Y_p$ ⇒ $COV(Y_i, X_j) = \lambda_i e_{ij}$

主成分负荷

$\rho(Y_i, X_j) = \frac{COV(Y_i, X_j)}{\sqrt{Var(Y_i)}\sqrt{Var(X_j)}} = \frac{\lambda_i e_{ij}}{\sqrt{\lambda_i}\sqrt{\sigma_{jj}}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}e_{ij}$

$\sum_{i=1}^{p}\rho^2(Y_i, X_j) = \sum_{i=1}^{p}\frac{\lambda_i e_{ij}^2}{\sigma_{jj}} = 1$

因 $P^T\Sigma P = \Lambda$ ⇒ $\Sigma = P\Lambda P^T$　$P = e_1, e_2, \dots, e_p$

$\sigma_{jj} = (e_{1j}, e_{2j}, \dots, e_{pj})\Lambda(e_{1j}, e_{2j}, \dots, e_{pj})^T$ ⇒ $\sum_{i=1}^{p}\lambda_i e_{ij}^2 = \sigma_{jj}$

主成分对于原始变量的贡献率

前m个主成分对于原始变量 X_j 的贡献率 $v_j^{(m)}$ 定义为 X_j 与 $Y_1, \dots, Y_m$ 相关系数的平方和

$$v_j^{(m)} = \sum_{i=1}^{m}\rho^2(Y_i, X_j) = \sum_{i=1}^{m}\frac{\lambda_i e_{ij}^2}{\sigma_{jj}}$$

标准化

$X_k^* = \frac{X_k - \mu_k}{\sqrt{Var(X_k)}}$

$\rho = cov(X^*)$　X的相关系数阵

标准化后主成分 ⇔ 求ρ特征值分解

$\sum_{k=1}^{p} Var(Y_k^*) = \sum_{k=1}^{p}\lambda_k^* = \sum_{k=1}^{p} Var(X_k^*) = p$

贡献率 $\frac{\lambda_k^*}{p}$　累计 $\sum \frac{\lambda_k^*}{p}$

样本主成分

$S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$　代替Σ

$\rho = (r_{jk})_{p\times p} = \left(\frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}\right)_{p\times p}$　代替ρ

第k个主成分

$y_k = \hat{e}_k^Tx$，k=1, 2, …, p

得分　将 $X = (X_1, \dots, X_p)^T$ 的 n组观测值 $(x_1, \dots, x_n)$，$x_i = (x_{i1}, \dots, x_{ip})^T$（向量，样本）(i=1, …, n)

依次代入

$y_k = \hat{e}_k^Tx$，k=1, 2, …, p

得 $y_{1k}, \dots, y_{nk}$　第k个样本主成分的得分

$y_{1k} = \hat{e}_k^Tx_1, \dots, y_{nk} = \hat{e}_k^Tx_n$

得分的样本方差

$\hat{e}_k^TS\hat{e}_k = \hat{\lambda}_k$　k=1, …, p

$\hat{e}_j^TS\hat{e}_k = 0$　$1 \le j \ne k \le p$

样本总方差

$\sum_{k=1}^{p} S_{kk} = \sum_{k=1}^{p}\hat{\lambda}_k$

贡献率　累计：对m求和

$\hat{\lambda}_k / \sum_{j=1}^{p}\hat{\lambda}_j$

原数据	主成分
1　$x_{11}, x_{12}, \dots, x_{1p}$	$y_{11}, y_{12}, \dots, y_{1p}$
2　$x_{21}, x_{22}, \dots, x_{2p}$	$y_{21}, y_{22}, \dots, y_{2p}$
⋮　⋮	⋮
n.　$x_{n1}, x_{n2}, \dots, x_{np}$	$y_{n1}, \dots, y_{np}$

典型相关分析

$Y_1, \dots, Y_q$ 与 $X_1, \dots, X_p$ 间的相关

找 $U_i = a_i^TX$，$V_i = b_i^TY$　s.t.

U_i, V_i 最大可能提取相关性，各对典型变量提取的相关性不重叠

模型

$X = (X_1, \dots, X_p)^T$　$Y = (Y_1, \dots, Y_q)^T$　$p \le q$

$\binom{X}{Y} = (X_1, \dots, X_p, Y_1, \dots, Y_q)^T$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$\Sigma_{11} = COV(X)$　$\Sigma_{22} = COV(Y)$

$\Sigma_{21}^T = \Sigma_{12} = COV(X, Y)$　满秩（假设）

$U_1 = a_1^TX$　$Var(U_1) = a_1^T\Sigma_{11}a_1$

$V_1 = b_1^TY$　$Var(B_1) = b_1^T\Sigma_{22}b_1$

$COV(U_1, V_1) = a_1^T\Sigma_{12}b_1$

$\rho_{U_1,V_1} = \frac{a_1^T\Sigma_{12}b_1}{\sqrt{a_1^T\Sigma_{11}a_1}\sqrt{b_1^T\Sigma_{22}b_1}}$

转化为优化问题

$\max_{U_1,V_1} \rho_{U_1,V_1} = a_1^T\Sigma_{12}b_1$，s.t. $a_1^T\Sigma_{11}a_1 = 1$，$b_1^T\Sigma_{22}b_1 = 1$

$\max_{U_k,V_k} \rho_{U_k,V_k} = a_k^T\Sigma_{12}b_k$，s.t. $a_k^T\Sigma_{11}a_k = 1$，$b_k^T\Sigma_{22}b_k = 1$，$cov(U_k, U_j) = cov(V_k, V_j) = 0$，$cov(U_k, V_j) = cov(V_k, U_j) = 0$　$\forall 1 \le j < k$

特征值分解

$A = \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$，$B = \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$

特征值：$\rho_1^2 \ge \rho_2^2 \ge \dots \ge \rho_p^2$　A的p个　　$\rho_1^2 \ge \rho_2^2 \ge \dots \ge \rho_p^2$　前p个最大特征值

$e_1, e_2, \dots, e_p$　　$f_1, f_2, \dots, f_p$

$U_k = e_k^T\Sigma_{11}^{-1/2}X$　　$V_k = f_k^T\Sigma_{22}^{-1/2}Y$

$\rho_{U_k,V_k} = \rho_k$　k=1, 2, …, p

判别分析

k个总体 $G_1, \dots, G_k$　p维

$X = (X_1, \dots, X_p)^T$

G_i 分布 $F_i(x) = F_i(x_1, \dots, x_p)$

$f_i(x)$

找规则，使得 某种意义下 判别 最优

距离：

欧氏　没考虑波动性

马氏：
- x, y　来自 μ, Σ 的总体G的样品
- x与y的：$d^2(x, y) = (x - y)^T\Sigma^{-1}(x - y)$
- x与G的：$d^2(x, G) = (x - \mu)^T\Sigma^{-1}(x - \mu)$

G_1 与 G_2，μ_1, μ_2，Σ相同

$d^2(G_1, G_2) = (\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2)$

距离判别：

$$\begin{cases} x \in G_1 & d^2(x, G_1) < d^2(x, G_2) \\ x \in G_2 & d^2(x, G_1) > d^2(x, G_2) \end{cases}$$

归到最近的马氏距离总体

$\Sigma_1 = \dots = \Sigma_k$ 时：

$\hat{\Sigma} = \frac{1}{n-k}\left[\sum_{i=1}^{k}(n_i - 1)S_i\right]$　$n = \sum_{i=1}^{k} n_i$

$W_i = a_i^Tx + b$，$a_i = \Sigma^{-1}\mu_i$，$b = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i$

二次判别　$d^2(x, G_i)$ 取最小的 i

判别准则：

p_i：样本属于 G_i 的概率（先验）

$p^* = P$(将x判错) $= p_1P(2|1, R) + p_2P(1|2, R)$

$P(j|i, R)$：R将属于 G_i 的样品误判为 G_j 的条件概率

1. 回代法

$n = n_1 + n_2$ 训练 → 判别

实际 \ 结果	G_1	G_2	合计
G_1	n_{11}	n_{12}	n_1
G_2	n_{21}	n_{22}	n_2
合计	$\hat{n}_1$	$\hat{n}_2$	n

故 $\hat{p}^* = \frac{n_{12} + n_{21}}{n_1 + n_2} < p$

2. 交叉确认估计法：

1. 剔除一个，剩余的 $n_1 + n_2 - 1$ 建立准则
2. 对剔除的进行判别
3. 对每个样品重复，误判比例作为误判率

n_{12}^*：对 G_1 误判数

n_{21}^*：对 G_2 误判数

$\hat{p}^* = \frac{n_{12}^* + n_{21}^*}{n_1 + n_2}$

Bayes

先验：G_1 和 G_2 出现的先验

$p_1 = P(G_1)$，$p_2 = P(G_2)$

$p_1 + p_2 = 1$　无信息 $p_1 = p_2 = 1/2$；比例 $p_1 = n_1/n$　$p_2 = n_2/n$

准则R：空间的划分 $R^p = \{R_1, R_2\}$

$P(2|1, R) = \int_{R_2} f_1(x)dx$　损失为 c(2|1)

$P(1|2, R) = \int_{R_1} f_2(x)dx$　损失为 c(1|2)

G_1 和 G_2 后验：$P(G_1|x) = \frac{p_1f_1(x)}{p_1f_1(x) + p_2f_2(x)}$ ∝ $p_1f_1(x)$

c(1|2) = c(2|1) 时，

$R_1 = \{x: p_1f_1(x) \ge p_2f_2(x)\}$

$R_2 = \{x: p_2f_2(x) > p_1f_1(x)\}$

$p^* = p_1P(2|1, R) + p_2P(1|2, R)$　证明时结合一项

当 c(1|2) ≠ c(2|1) 时，

$R_1 = \{x: c(2|1)p_1f_1(x) \ge c(1|2)p_2f_2(x)\}$

$R_2 = \{x: c(1|2)p_2f_2(x) > c(2|1)p_1f_1(x)\}$

对正态：

损失相同

(1) 当 $\Sigma_1 = \Sigma_2 = \Sigma$ 时。

$R_1 = \{x: W_1(x) \ge W_2(x)\}$，

$R_2 = \{x: W_1(x) < W_2(x)\}$。

$W_1(x) = a_1^Tx + b$，$a_1 = \Sigma^{-1}\mu_1$，$b = -\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \ln p_1$

$W_2(x) = a_2^Tx + b$　$a_2 = \Sigma^{-1}\mu_2$　$b = -\frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \ln p_2$

μ_1, μ_2, Σ 估计如前

(2) 当 $\Sigma_1 \ne \Sigma_2$ 时，c(1|2) ≠ c(2|1) 时。

$R_1 = \{x: -\frac{1}{2}(x - \mu_1)^T\Sigma_1^{-1}(x - \mu_1) - \frac{1}{2}\ln|\Sigma_1| + \ln(c(2|1)p_1)$

$\ge -\frac{1}{2}(x - \mu_2)^T\Sigma_2^{-1}(x - \mu_2) - \frac{1}{2}\ln|\Sigma_2| + \ln(c(1|2)p_2)\}$

$R_2 = \{\dots\}$

聚类分析

样品的相似性度量　$x_i = (x_{i1}, \dots, x_{ip})^T$，$x_j = (x_{j1}, x_{j2}, \dots, x_{jp})^T$

1. 欧氏距离

$d(x_i, x_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2}$

2. 绝对距离

$d(x_i, x_j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}|$

3. Minkowski

$d(x_i, x_j) = \left[\sum_{k=1}^{p}|x_{ik} - x_{jk}|^m\right]^{1/m}$，m ≥ 1.

4. Chebyshev：∞范数

5. 马氏距离

6. 方差加权：标准化后欧氏

7. 兰氏距离　$x_{ij} > 0$

$\frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}$

变量间的相似性度量

相关系数

夹角余弦

$\frac{\langle x_1, x_2 \rangle}{\|x_1\|\|x_2\|}$

类间距离：

G_p 的重心：$\bar{x}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_i^{(p)}$，$x_1^{(p)}, \dots, x_{n_p}^{(p)}$ 为样品

① 最短距离

$D_{pq} = \min_{i \in G_p, j \in G_q}\{d_{ij}\}$　$D_{rk} = \min\{D_{pk}, D_{qk}\}$

② 最长距离

$D_{pq} = \max_{i \in G_p, j \in G_q}\{d_{ij}\}$　$D_{rk} = \max\{D_{pk}, D_{qk}\}$

③ 类平均距离

$D_{pq} = \frac{1}{n_pn_q}\sum_{i \in G_p}\sum_{j \in G_q} d_{ij}$　$D_{rk} = \frac{n_p}{n_r}D_{pk} + \frac{n_q}{n_r}D_{qk}$

$D_{pq}^2 = \frac{1}{n_pn_q}\sum_{i \in G_p}\sum_{j \in G_q} d_{ij}^2$　$D_{rk} = \frac{n_p}{n_r}D_{pk}^2 + \frac{n_q}{n_r}D_{qk}^2$

④ 重心距离

$D_{pq} = d(\bar{x}_p, \bar{x}_q) = \sqrt{(\bar{x}_p - \bar{x}_q)^T(\bar{x}_p - \bar{x}_q)}$　$D_{rk}^2 = \frac{n_p}{n_r}D_{pk}^2 + \frac{n_q}{n_r}D_{qk}^2 - \frac{n_pn_q}{n_rn_r}D_{pq}^2$

⑤ 离差

$D_{pq}^2 = W_r - W_q - W_p = \frac{n_pn_q}{n_p + n_q}(\bar{x}_p - \bar{x}_q)^T(\bar{x}_p - \bar{x}_q)$

$D_{rk}^2 = \frac{n_p + n_k}{n_r + n_k}D_{pk}^2 + \frac{n_q + n_k}{n_r + n_k}D_{qk}^2 - \frac{n_k}{n_r + n_k}D_{pq}^2$

递推

$G_p, G_q, G_k \xrightarrow{合并} G_r, G_k$

$\{D_{pk}, D_{qk}\} \to D_{rk}$

$D_{kr}^2 = \alpha_pD_{kp}^2 + \alpha_qD_{kq}^2 + \beta D_{pq}^2 + \gamma|D_{kp}^2 - D_{kq}^2|$

表

Q聚类

① n个样品 n类

$$D_{(0)} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix}$$　$D_{pq} = d_{pq}$

② 选 $D_{(0)}$ 最小元素 D_{pq}，$G_p, G_q \to G_r$

消去 $D_{(0)}$ 中 G_p, G_q 行列，

加入 G_r，算出 $D_{(1)}$

③ 迭代，直到聚为一大类

④ 记下编号画图