Lecture 7 成对数据的统计分析

$x_1, x_2, \dots x_n$

回顾 一元 频数分布. $\bar{x}$, S^2 样本估计总体.

→ 二元 相关性

- 1° 数值数据之间 x的取值分布在连续区间
- 2° 二分类变量之间 {及格 → 0, 不及格 → 1} {男 → 0, 女 → 1}

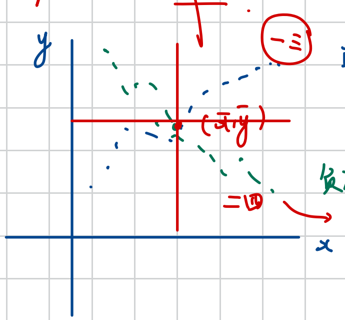
1° $(x_1, y_1), \dots (x_n, y_n)$

2° 分类 2×2列联表.

Y \ X	0	1
0	n_{11}	n_{12}
1	n_{21}	n_{22}

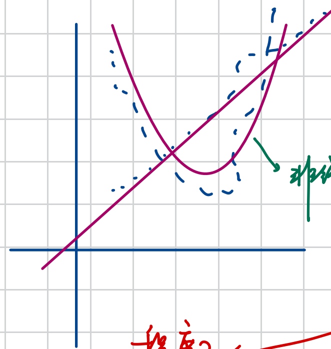
样本估计总体 → 误差

1° Pearson (线性) 相关系数



正相关. $x \cdot y > 0$

负相关. $xy < 0$



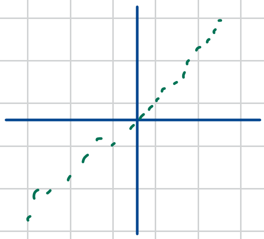
拟合 曲线

线性相关

非线性(曲线)相关.

程度?

平移 $(x_i - \bar{x}, y_i - \bar{y})$



体重 vs 身高 m → cm 同一样本 L_{xy} ↑100倍 程度仍

$$L_{xy} = (x_1-\bar{x})(y_1-\bar{y}) + \dots + (x_n-\bar{x})(y_n-\bar{y})$$

$L_{xy} > 0$ ⇝ 正相关　　$L_{xy} < 0$ ⇝ 负相关.

相关程度 → 数值大小.

$\vec a=(x_1,x_2)$ $\vec b=(y_1,y_2)$ 定义

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar x)^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar y)^2}}$$

量纲 ×

$$\cos\theta=\frac{\vec a\cdot\vec b}{|\vec a||\vec b|}=\frac{|x_1y_1+x_2y_2|}{\sqrt{x_1^2+x_2^2}\cdot\sqrt{y_1^2+y_2^2}}\le 1$$

类似.

↓ 推广

Q: $|\vec a\cdot\vec b|\le|\vec a||\vec b|$? ⇒ $(\vec a\cdot\vec b)^2\le|\vec a|^2|\vec b|^2$?

$$=\frac{\vec a\cdot\vec b}{|\vec a||\vec b|}\in[-1,1]$$

$\vec a=(x_1-\bar x,x_2-\bar x,\dots,x_n-\bar x)$ $\vec b=(y_1-\bar y,y_2-\bar y,\dots,y_n-\bar y)$.

取到 -1 或 1 说明 $\vec a\parallel\vec b$ ☆ → 同时落在一条直线上.

即 ∃λ∈R, $\vec a=\lambda\vec b$

⇒ $x_i-\bar x=\lambda(y_i-\bar y)$ r=-1, k<0; r=1, k>0

⇒ (x_i,y_i) 都在 l: $x-\bar x=\lambda(y-\bar y)$ 上.

定义 $\vec a\cdot\vec b=(x_1-\bar x)(y_1-\bar y)+\dots+(x_n-\bar x)(y_n-\bar y)$

↓

$|\vec a|=\sqrt{\vec a^2}$.

命题 (Cauchy-Schwarz) 在 $\mathbb{R}^n$ 空间中, 定义内积 $\vec a\cdot\vec b=x_1y_1+\dots+x_ny_n$.

求证: $(\vec a\cdot\vec b)^2\le|\vec a|^2|\vec b|^2$.

pf: $|\vec a|^2=x_1^2+\dots+x_n^2\ge0$ 取"=" ⇔ $\vec a=\vec 0$ ⇔ $\vec a\cdot\vec a=0$ 正定性

构造 $\vec a-t\vec b$ ∀t∈R.

$(\vec a-t\vec b)\cdot(\vec a-t\vec b)\ge0$ 恒成立.

$\vec a^2-2t\,\vec a\cdot\vec b+t^2\vec b^2\ge0$, ∀t∈R ⇒ $\Delta=4(\vec a\cdot\vec b)^2-4\vec a^2\vec b^2\le0$

⇒ $(\vec a\cdot\vec b)^2\le\vec a^2\vec b^2$.

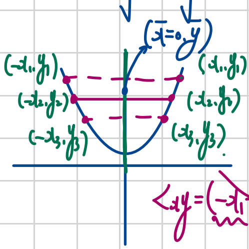
⇒ $(x_1y_1+\dots+x_ny_n)^2\le(x_1^2+\dots+x_n^2)(y_1^2+\dots+y_n^2)$ ☆

取等: $\vec a-t_0\vec b=0$, ∃t_0∈R. ⇒ $\vec a\parallel\vec b$

L_{xy} >0 正; <0 负

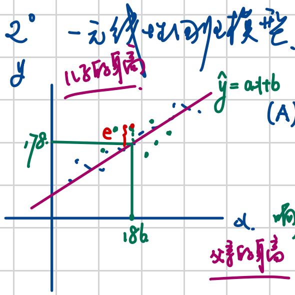
线性 → $r=\frac{L_{xy}}{\sqrt{\ }\sqrt{\ }}\in[-1,1]$

- $|r|\to1$ 线强.
- $|r|=1$ 直线
- $|r|=0$ 没有线性相关性. 但可能曲线相关.



$L_{xy}=(-x_1-0)(y_1-\bar y)+(x_1-0)(y_1-\bar y)+\dots=0$ ⇒ $r=\frac{L_{xy}}{\sqrt{\ }\sqrt{\ }}=0$

2° 一元线性回归模型



(A) 假设模型

$$\begin{cases} Y = bx + a + e \\ E(e)=0,\quad D(e)=\sigma^2 \end{cases}$$

解释变量（x）；响应变量（Y）；随机误差（e）；$\hat{y}$ ~ 预测值

$D(e)=\sigma^2$ 与x无关常量.

实际 - 预测 = e.

样本e图

1° 总体中一个x → 一个y. 同组中一个x → 总体中的多个y → $\bar{y}$ 的均值 = 预测值 ax_i+b

2° σ^2 = 随机误差e的方差 = 同一组中代入x后得到的总体的方差.

3° σ^2 与x无关

(B) 用样本数据估计未知参数 $\hat{a}$ ✓, $\hat{b}$ ✓, σ^2 ✗

$$Y = \hat{y} + e = bx + a + e$$

$(x_1, y_1), \dots, (x_n, y_n)$

$$\sum_{i=1}^{n}[y_i-(bx_i+a)] = \sum_{i=1}^{n}\hat{e}_i \text{ 最小的 } a, b.$$

正负抵消

$$\sum_{i=1}^{n}|y_i-(bx_i+a)|$$ 去绝对值 ⇒ 运算麻烦；最小一乘估计

$$\sum_{i=1}^{n}e_i^2 = \sum_{i=1}^{n}(y_i-(bx_i+a))^2 \ ✓ = Q(a,b) \quad \min.$$

$$Q(a,b)=\sum_{i=1}^{n}\left(y_i-bx_i-(\bar{y}-b\bar{x})+(\bar{y}-b\bar{x})-a\right)^2$$

$$=\sum_{i=1}^{n}\left[(y_i-\bar{y})-b(x_i-\bar{x})+(\bar{y}-b\bar{x})-a\right]^2.$$

$$=\sum_{i=1}^{n}\left[(y_i-\bar{y})-b(x_i-\bar{x})\right]^2+2\sum_{i=1}^{n}\left[(y_i-\bar{y})-b(x_i-\bar{x})\right](\bar{y}-b\bar{x}-a)+n(\bar{y}-b\bar{x}-a)^2.$$

（中间项 $=0$；$n(\bar{y}-b\bar{x}-a)^2 \geq 0$）

$$0=2\left\{\left(\sum_{i=1}^{n}y_i-n\bar{y}\right)-b\left(\sum_{i=1}^{n}x_i-n\bar{x}\right)\right\}$$ （$=0$，$=0$）

估计

☆ 取 $\hat{a}=\bar{y}-\hat{b}\bar{x}$

所谓样本中心点

$\hat{y}=\hat{a}+\hat{b}x$ 上 但 $(\bar{x},\bar{y})$ 一定在 ☆

$$= \sum_{i=1}^{n}\left[(y_i-\bar{y})^2 - 2b(x_i-\bar{x})(y_i-\bar{y}) + b^2(x_i-\bar{x})^2\right]$$

$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 - \left(2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)b + b^2\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\hat{b} = \frac{-2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{-2\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$Q(a,b) \geq Q(\hat{a},\hat{b})$

$\hat{a},\hat{b}$ 称为 a,b 的最小二乘估计值

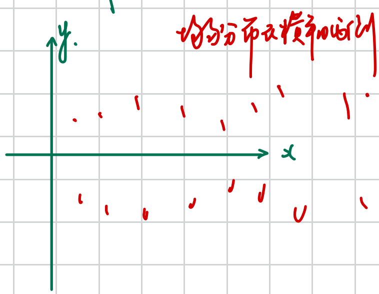
$\hat{y} = \hat{b}x + \hat{a}$ 模拟的经验回归方程.

(c) 效果检验 残差

y 真实值样本

$$\begin{cases} y = \hat{y} + e = \hat{a} + \hat{b}x + e \\ Var(e) = \sigma^2,\quad E(e) = 0 \end{cases}$$

1° 残差图



2° 决定系数 R^2.

$y_i = bx_i + a + e_i$ （e的体现）

$\hat{y}_i = bx_i + a$

$\bar{y}$ = 所有真实值的均值.

1° 总偏差平方和（方差） $S^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2$. 波动

2° $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}e_i^2$. 残差平方和.

$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$　x_i 对 y_i 的影响（回归直线）

$\hat{y}_i = \hat{a} + \hat{b}x_i$

$$S^2 = \sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 + 2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.$$

$(y_i-\hat{b}x_i-\hat{a})(\hat{b}x_i+\hat{a}-\bar{y})$

$\sum(y_i-\hat{b}x_i-\hat{a})[\hat{b}(x_i-\bar{x})]$

$= \sum[(y_i-\bar{y})-\hat{b}(x_i-\bar{x})]\hat{b}(x_i-\bar{x})$

$= \{\sum(y_i-\bar{y})(x_i-\bar{x}) - \hat{b}\sum(x_i-\bar{x})^2\}\hat{b}$

$= 0\cdot\hat{b} = 0$

$$\hat{b} = \frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \in [0,1]$$

$R^2 \to 1$ 模型拟合越好

(d) 非线性回归模型

1°

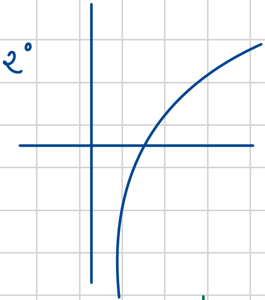
$y = \alpha e^{\beta x}$ ($\alpha > 0$) 估计 $\alpha . \beta$

$\ln y = \ln\alpha + \beta \cdot x$

$\ln\alpha$ 看成.

$(x_1, \ln y_1), (x_2, \ln y_2) \text{——} (x_n, \ln y_n)$.

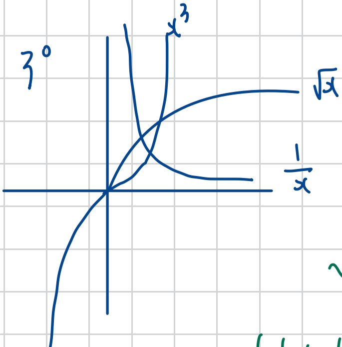
2°



$y = \alpha + \beta \log_a x$

$(\log_a x_1, y_1) \cdots (\log_a x_n, y_n)$

3°

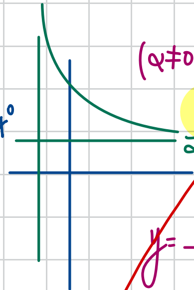


$y = \alpha x^{\beta}$ ($\alpha > 0$)

$\Rightarrow \ln y = \ln\alpha + \beta \ln x$.

$(\ln x_1, \ln y_1) \text{——} (\ln x_n, \ln y_n)$.

4°



$y = \frac{x}{\alpha x + \beta}$ ($\alpha \neq 0$)

$y = \frac{x + \frac{\beta}{\alpha} - \frac{\beta}{\alpha}}{\alpha x + \beta} = \frac{1}{\alpha} - \frac{\beta}{\alpha} \cdot \frac{1}{\alpha x + \beta}$

$\frac{1}{y} = \alpha + \beta \cdot \frac{1}{x}$

$(\frac{1}{x_1}, \frac{1}{y_1}) \cdots (\frac{1}{x_n}, \frac{1}{y_n})$ ★.

3° 分类变量的 χ^2 检验

X、Y是否相关? ⇝ 数学成绩是否优秀与甲乙校类别是否有关?

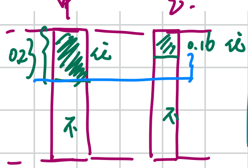
Y \ X	0	1
0	n_{11}	n_{12}
1	n_{21}	n_{22}

样本 → 频数统计

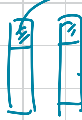
学校	数学成绩		合计
	不优秀 (Y = 0)	优秀 (Y = 1)	
甲校 (X = 0)	33	10	43
乙校 (X = 1)	38	7	45
合计	71	17	88

1° $\frac{\text{甲优秀}}{\text{甲总}} = \frac{10}{43} = 0.23 \neq \frac{\text{乙优秀}}{\text{乙总}} = \frac{7}{45} = 0.16$ 有关.

等高堆积图



样本!



取样 样本不合理 数量 ★

误差=? 犯错误p=?

χ^2（卡方）检验

1° 零假设 H_0：假设 X 与 Y 无关 ⇒ 独立.

$$P\{X=0 \mid Y=0\} = P\{X=0 \mid Y=1\} \Rightarrow P(B|A) = P(B|\bar{A})$$

（$\{X=0\}$ 记为 B，$\{Y=0\}$ 记为 A，$\{Y=1\}$ 记为 $\bar{A}$）

⇔ $\{X=0\}$ 与 $\{Y=0\}$ 独立.

$\{A, B\}$，$\{A, \bar{B}\}$，$\{\bar{A}, B\}$，$\{\bar{A}, \bar{B}\}$ 独立.

⇒ 随机变量 X、Y 独立.

2°

Y \ X	甲 $X=0$	乙 $X=1$	合计
$Y=0$	a	b	a+b
$Y=1$	c	d	c+d
合计	a+c	b+d	n=a+b+c+d

频率稳定于概率

$$P\{Y=0 \mid X=0\} = \frac{a}{a+c}$$

$$P\{Y=0 \mid X=1\} = \frac{b}{b+d}$$

⇔ $P\{Y=0 \mid X=0\} \stackrel{?}{=} P\{Y=0 \mid X=1\}$

$$P\{X=0, Y=0\} = P\{X=0\}P\{Y=0\}$$

差距很小

$$\frac{a}{a+b+c+d} - \frac{(a+c)(b+d)}{[a+c+b+d]^2} = \frac{a}{n} - \frac{(a+c)(b+d)}{n^2}$$

$$= \frac{1}{n}\left|a - \frac{(a+c)(b+d)}{n}\right|^2 \quad \text{很小}$$

四个概率差很小

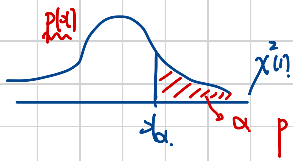
$$\chi^2 = \frac{\left(a - \frac{(a+c)(b+d)}{n}\right)^2}{\frac{(a+c)(b+d)}{n}} + \frac{\left(b - \frac{(a+b)(b+d)}{n}\right)^2}{\frac{(a+b)(b+d)}{n}} + \frac{\left(c - \frac{(a+c)(c+d)}{n}\right)^2}{\frac{(a+c)(c+d)}{n}} + \frac{\left(d - \frac{(c+d)(b+d)}{n}\right)^2}{\frac{(c+d)(b+d)}{n}}$$

$$= \frac{n(ad-bc)^2}{(a+c)(a+b)(b+d)(b+c)}.$$

在可能的在 H_0 成立情况下，不能太大

3° $\chi^2 \xrightarrow{(n\to\infty)} \chi^2(1)$ 具体分布　抽样很大　数值很小　(H_0 成立时)

记 $P\{\chi^2 \geq x_\alpha\} \approx P\{\chi^2(1) \geq x_\alpha\} = \alpha$.



$P\{X \geq x_\alpha\} = \alpha$.

4° (b) χ^2 独立性检验中可能用到的小概率值 α 与对应的临界值 x_α 如下，即 $P\{\chi^2 \geq x_\alpha\} = \alpha$:

α	0.1	0.05	0.01	0.005	0.001
x_α	2.706	3.841	6.635	7.879	10.828

$\chi^2 = 12 > x_{0.001}$　　$P\{\chi^2 > x_{0.001}\} = 0.001$

小概率事件　几乎不可能发生!　但它发生了

假设有问题! ⇒ 拒绝 H_0

⇒ X, Y 有关，犯错误概率不超过 0.001

即 99.99% 的把握 X, Y 有关 ✓

$\chi^2 = 2.6 < 2.706 = x_{0.1}$　　$\alpha = 0.1$

没有充分证据说明 H_0 为错误的.

接受 H_0，即认为 X 与 Y 独立（无关）

注：此处不可用 $0.1 = \alpha$ 来刻画 X 与 Y 独立的准确性!