第三讲. 相关系数

2014.2.25

内 容

1. 预备知识：**delta**方法和方差稳定化变换
2. **Pearson**相关系数
3. **Spearman's rho, Kendall's tau**
4. 随机向量

1. 预备知识： Delta-方法与方差稳定化变换

已知某统计量的渐近正态分布，利用**Delta**方法容易求出其变换后的极限分布

定理($Delta$方法). 若$\sqrt{n}(X_n-\theta)\xrightarrow{d}N(0,\sigma^2(\theta))$
则$\sqrt{n}(g(X_n)-g(\theta))\xrightarrow{d}N(0,[g'(\theta)]^2\sigma^2(\theta))$,
其中假设$g'(\theta)$存在且非0.

证明(不要求)：泰勒展开：$\sqrt{n}(g(X_n)-g(\theta))\approx\sqrt{n}g'(\theta)(X_n-\theta)$

通常需要对统计量做“方差稳定化变换”（变化后的渐近分布的方差与参数无关）

方差稳定化变换：若$E(X)=\mu,\mathrm{var(X)}=\sigma^2(\mu)$与均值$\mu$有关,
则$\mathrm{Y}=\int_c^X\frac{1}{\sigma(\mu)}d\mu$称为方差稳定化变换。

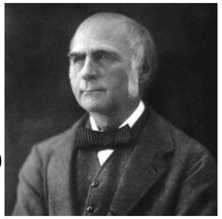
$Delta$方法：若$\sqrt{n}(X_n-\theta)\xrightarrow{d}N(0,\sigma^2)$
则$\sqrt{n}(g(X_n)-g(\theta))\xrightarrow{d}N(0,[g'(\theta)]^2\sigma^2)$.

在Delta方法中,如果X_n的渐近方差$\sigma^2=\sigma^2$（θ）与均值θ有关，
为使得$g(X_n)$的渐近方差与θ无关，只需：

$$[g'(\theta)]^2\sigma^2(\theta)=\mathrm{C}\text{（常数）}$$

解方程得：$g(\theta)\propto\int\frac{1}{\sigma(\theta)}d\theta$,称为方差稳定化变换

卡尔.皮尔逊 (Karl Pearson,1857-1936)



卡尔.皮尔逊，英国数学家，现代统计的创始人(以1900年的皮尔逊卡方检验为标志)。他是高尔顿的门徒和传记作者。论著 *The Grammar of Science* 影响了爱因斯坦。

1901年他和Galton, Weldon 一起创办了第一份统计杂志Biometrika, 1925年创办了优生学/遗传学杂志*Annals of Eugenics* (*Annals of Human Genetics*)。1911年在伦敦大学学院建立了世界上第一个(生物)统计系。

主要统计贡献：
相关系数，矩方法，Pearson分布族，P值，假设检验和决策理论，Pearson卡方检验，主成分分析..

2. Pearson 相关系数

Pearson 相关系数度量线性关联程度。概念和初始定义由**Galton**提出，但深入的研究和推广使用属于**K. Pearson**。

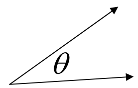
样本 : $(x_1, y_1),\dots,(x_n, y_n)$

Pearson 样本相关系数：

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \sum \frac{(x_i-\bar{x})}{\sqrt{s_{xx}}}\frac{(y_i-\bar{y})}{\sqrt{s_{yy}}}$$

$= $ 标准化向量$(x_1,\dots,x_n)'$与$(y_1,\dots,y_n)'$的内积

$=\cos(\theta)$



记号：$s_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})$，$s_{xx}=\sum(x_i-\bar{x})^2$，$s_x=\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$

正态假设下的相关性检验(精确检验/小样本检验)

随机变量x，y的(总体)相关系数 : $\rho=\rho_{xy}=\frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}}$

$$H_0:\rho=0 \leftrightarrow H_1:\rho\neq 0$$

假设$(x_1,y_1),\dots,(x_n,y_n)$ iid ~ 二元正态
(或$y|x$ ~ 正态，或更弱的条件下)：

$$t=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2}$$

(以后将证明)

r的密度函数（$\rho=0$时）

$$f(r)=\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)}\left(1-r^2\right)^{\frac{n-4}{2}}$$

$$\text{p值}=P(|t_{n-2}|\geq|t|)=2P(t_{n-2}\geq|t|)$$

作业：两样本 t-检验是特殊的相关性检验。
假设第一组$y_1,\dots,y_{n_0}$ $iid\sim N(\mu_0,\sigma^2)$，组号$x_i=0$
第二组$y_{n_0+1},\dots,y_{n_0+n_1}$ $iid\sim N(\mu_1,\sigma^2)$，组号$x_i=1$
$r=r_{xy}$为样本相关系数

则$t=\sqrt{n-2}\frac{r}{\sqrt{1-r^2}}=$两样本$t-$检验.

即,检验H0 : $\mu_0=\mu_1$等价于 检验H0 : $\rho_{xy}=0$