多元统计与矩阵分析

张锋　*8125345@qq.com*

中国地质大学, 计算机学院, 武汉

第1章
多元正态分布

内容

(1) 随机向量

(2) 多元正态分布概述

(3) 多元正态分布的参数估计

(4) 常用分布与抽样分布

内容

(1) 随机向量

(2) 多元正态分布概述

(3) 多元正态分布的参数估计

(4) 常用分布与抽样分布

随机向量

- 样本数据库（样本资料阵）

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

无特殊说明，本书所有变量均为列变量

- 样品：每一个个体的p个变量成为一个样品
- 样本：全体n个样品形成一个样本
- 总体
- 随机向量：指标(变量)排在一起所构成的向量
- 横看，第i个样品的观测值,p维随机向量
- 纵看，第j个变量的n次观测，n维随机向量

多“元”数据

学生成绩表											
学号	姓名	班级	性别	政治	语文	英语	数学	物理	化学	总分	平均分
20060001	陈文章	1	男	93	89	87	85	82	86		
20060002	张强	1	男	84	86	89	92	90	88		
20060003	李芬	1	女	95	90	93	89	91	89		
20060004	陆洋	2	男	83	78	76	86	80	84		
20060005	姚舒	3	女	90	92	94	95	92	90		
20060006	丁琼玉	2	女	81	84	83	78	84	67		
20060007	钟达峰	2	男	78	83	75	82	85	77		
20060008	朱桐	3	男	92	94	96	96	95	93		
20060009	吴燕	1	女	87	91	90	94	96	91		
20060010	曹斌	3	男	81	80	82	87	88	83		
20060011	宋令文	3	男	88	96	97	95	94	93		
20060012	杨华	2	女	95	93	92	92	87	85		

多“元”数据

表 3-11　指标数据(1)

地区	人均地区生产总值	人均可支配收入	人均拥有公共图书馆藏量	人均教育文化娱乐消费支出	人均交通通讯消费支出	每百户计算机拥有量
北京	129 041.64	57 229.8	1.27	3 916.72	5 033.98	101.08
天津	119 134.17	37 022.3	1.07	2 691.52	3 744.54	74.21
河北	45 234.47	21 484.1	0.34	1 578.29	2 290.30	59.48
山西	41 946.03	20 420.0	0.47	1 879.25	1 884.04	55.57
内蒙古	63 646.54	26 212.2	0.71	2 227.80	2 914.92	47.35
辽宁	53 580.32	27 835.4	0.91	2 534.52	3 088.44	60.66
吉林	55 003.79	21 368.3	0.73	1 928.51	2 217.96	52.72
黑龙江	41 970.65	21 205.8	0.57	1 897.99	2 185.53	47.35
上海	126 687.30	58 988.0	3.21	4 685.92	4 057.65	131.07
江苏	106 949.51	35 024.1	1.07	2 747.59	3 496.40	78.75
浙江	91 511.86	42 045.7	1.38	2 844.91	4 306.54	80.75
安徽	43 194.24	21 863.3	0.41	1 700.51	2 102.26	46.23
福建	82 286.09	30 047.7	0.85	1 966.44	2 642.78	67.50
江西	43 284.96	22 031.4	0.53	1 606.79	1 600.74	50.81

多“元”数据

表 3-10　全国各地区居民生活质量测度指标体系

指标名称	变量名	单位
人均地区生产总值	X_1	元
人均可支配收入	X_2	元
人均拥有公共图书馆藏量	X_3	册
人均教育文化娱乐消费支出	X_4	元
人均交通通讯消费支出	X_5	元
每百户计算机拥有量	X_6	台
每百户家用汽车拥有量	X_7	辆
每百户照相机拥有量	X_8	台
每万人公共车辆数	X_9	辆
每千人口医疗卫生机构床位数	X_{10}	个
生活垃圾无害化处理率	X_{11}	%

随机向量的分布

- 定义1.1　p个随机变量 $X_1, X_2, \ldots, X_p$ 所组成的向量 $\boldsymbol{X} = (X_1, X_2, \ldots, X_P)'$ 称为随机向量。

- 定义1.2　设 $\boldsymbol{X} = (X_1, X_2, \ldots, X_P)'$ 为p维随机向量，其联合分布函数为

$$F(x_1, \ldots, x_p) = P(X_1 \le x_1, \ldots, X_p \le x_p)$$

记为$\boldsymbol{X \sim F}$。

随机向量的分布

定义1.3 如果存在非负函数$f(x_1,\ldots,x_p)$，使得对一切$(x_1,\ldots,x_p)\in R^p$，联合分布函数可表示为

$$F(x_1,\ldots,x_p)=\int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_p}f(t_1,\ldots,t_p)\,dt_1\ldots dt_p$$

则称X为连续型随机向量，称$f(x_1,\ldots,x_p)$为***X***的联合概率密度函数，简称为密度函数或者分布密度。

密度函数有以下两条重要性质:

(1) $\forall(x_1,\ldots,x_p)\in R^p$，$f(x_1,\ldots,x_p)\geq 0$

(2) $\int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty}f(t_1,\ldots,t_p)\,dt_1\ldots dt_p=1$

事实上，一个p维变量的函数$f(x_1,\ldots,x_p)$能作为p中某个随机向量的分布密度当且仅当以上两条性质成立时。

www.cug.edu.cn

随机向量的边际分布

定义1.4　　设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_p)'$ 为 p 维随机向量，其联合分布函数 $F(x_1,\cdots,x_p)$。$\boldsymbol{X}$ 的 q 个分量所组成的子向量 $(X_{i_1},\cdots,X_{i_q})'$ 的分布称为 $\boldsymbol{X}$ 的边缘(或边际）分布。如果我们将 $\boldsymbol{X}$ 划分为 q 维子向量 $\boldsymbol{X}^{(1)}$ 与 $p-q$ 维子向量 $\boldsymbol{X}^{(2)}$，那么 $\boldsymbol{X}^{(1)}$ 的

边缘分布为

$$\begin{aligned}F^{(1)}(x_1,\ldots,x_q)&=P(X_1\le x_1,\ldots,X_q\le x_q)\\&=P(X_1\le x_1,\ldots,X_q\le x_q,X_{q+1}\le\infty,\ldots,X_p\le\infty)\\&=F(x_1,\ldots,x_q,\infty,\ldots,\infty)\end{aligned}$$

当 $\boldsymbol{X}$ 有分布密度时，$\boldsymbol{X}^{(1)}$ 也有分布密度，其边缘密度为

$$f^{(1)}(x_1,\cdots,x_q)=\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}f(x_1,\cdots,x_p)\mathrm{d}t_{q+1}\cdots\mathrm{d}t_p$$

随机向量的条件分布与独立

定义 1.5　　如果我们将 $\boldsymbol{X}$ 划分为 q 维子向量 $\boldsymbol{X}^{(1)}$ 与 $p-q$ 维子向量 $\boldsymbol{X}^{(2)}$，那么在给定 $\boldsymbol{X}^{(2)}$ 时，$\boldsymbol{X}^{(1)}$ 的分布称为条件分布。如果 $\boldsymbol{X}$ 有密度函数 $f\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right)$，那么给定 $\boldsymbol{X}^{(2)}$ 时，$\boldsymbol{X}^{(1)}$ 的密度函数为

$$f_1\left(\boldsymbol{x}^{(1)} \mid \boldsymbol{x}^{(2)}\right) = f\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) / f_2\left(\boldsymbol{x}^{(2)}\right)$$

其中，$f_2\left(x^{(2)}\right)$是 $\boldsymbol{X}^{(2)}$ 的边缘密度。

定义1.6　若 p 个随机向量 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ 的联合分布等于各自边缘分布的乘积，则称 $\boldsymbol{X}_1, \cdots$，$\boldsymbol{X}_p$ 是相互独立的。需要注意的是，如果 $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ 相互独立，那么其中任意两个随机向量两两独立，但是反之不真。

总体期望 方差 标准差

- 总体均值
  - X是离散的随机变量 $E(X)=\mu=\sum_i p_i x_i$
  - X是连续的随机变量 $E(X)=\mu=\int_{-\infty}^{+\infty} xf(x)dx$
- 总体方差

$$Var(X)=E[(X-\mu)^2]$$

$$=E[X^2-2XE[X]+(E[X])^2]=E[X^2]-2E[X]E[X]+(E[X])^2=E[X^2]-(E[X])^2$$

$$Var(X)=\sigma^2=\frac{\Sigma(X-\mu)^2}{N}$$

- 总体标准差

$$\sigma=\sqrt{\frac{\Sigma(X-\mu)^2}{N}}$$

样本均值 方差 标准差

- 样本均值

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- 样本方差

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

Why N-1??

- 样本标准差

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

随机向量的均值

设 $\boldsymbol{X}=(X_1,X_2,\ldots,X_p)',\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_q)'$ 为两个随机向量。 若 $E(X_i)=\mu_i$ 存在，则称

$$E(\boldsymbol{X})=\begin{bmatrix}E(X_1)\\ \vdots\\ E(X_p)\end{bmatrix}=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_p\end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 的均值向量。

根据定义容易验证均值向量具有以下性质：

$$E(\boldsymbol{AX})=\boldsymbol{A}E(\boldsymbol{X})$$
$$E(\boldsymbol{AXB})=\boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}$$

其中 $\boldsymbol{A}$、$\boldsymbol{B}$ 为大小适合矩阵运算的常数矩阵。

www.cug.edu.cn

协方差矩阵

若 X_i 与 X_j 的协方差存在 $(i,j=1,\cdots,p)$，则称

$$\boldsymbol{\Sigma}=Cov(\boldsymbol{X},\boldsymbol{X})=D(\boldsymbol{X})=E[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))']$$

$$=\begin{bmatrix} D(X_1) & \operatorname{cov}(X_1,X_2) & \cdots & \operatorname{cov}(X_1,X_p) \\ \operatorname{cov}(X_2,X_1) & D(X_2) & \cdots & \operatorname{cov}(X_2,X_p) \\ \vdots & \vdots & \vdots & \vdots \\ \operatorname{cov}(X_p,X_1) & \operatorname{cov}(X_p,X_2) & \cdots & D(X_p) \end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 的协方差阵。

协方差矩阵

若 X_i 与 Y_j 的协方差存在 $(i=1,\cdots,p;j=1,\cdots,q)$，则称

$$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=E[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{Y}-E(\boldsymbol{Y}))']$$

$$=\begin{bmatrix}\mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \cdots & \mathrm{Cov}(X_1,Y_q)\\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \cdots & \mathrm{Cov}(X_2,Y_q)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{Cov}(X_p,Y_1) & \mathrm{Cov}(X_p,Y_2) & \cdots & \mathrm{Cov}(X_p,Y_q)\end{bmatrix}$$

为随机向量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差阵。

当 $\boldsymbol{X}=\boldsymbol{Y}$ 时，$\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})$ 即为 $D(\boldsymbol{X})$ 。

当 $\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})=0$ 时，称 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关。

如果 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 独立，则 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 不相关。反之不真。

www.cug.edu.cn

相关阵

若 X_i 与 X_j 的协方差存在 $(i,j=1,\cdots,p)$，则X_i与X_j的相关系数

$$r_{ij}=\frac{Cov(X_i,X_j)}{\sqrt{D(X_i)D(X_j)}}$$

将这 $p\times p$ 个相关系数排列成一个方阵 $\boldsymbol{R}=(r_{ij})_{p\times p}$，称为 $\boldsymbol{X}$ 的相关阵。若记的 X_i 的方差 $D(X_i)$ 为 σ_{ii}，则我们称 $\boldsymbol{V}^{1/2}=\mathrm{diag}(\sqrt{\sigma_{11}},\cdots,\sqrt{\sigma_{pp}})$ 为标准差矩阵。协方差矩阵与相关阵有这样的关系:

$$\boldsymbol{\Sigma}=\boldsymbol{V}^{1/2}\boldsymbol{R}\boldsymbol{V}^{1/2} \text{ 或 } \boldsymbol{R}=(\boldsymbol{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\boldsymbol{V}^{1/2})^{-1} \text{ 。}$$

根据协方差阵的定义，可以验证其具有以下性质：

(1)随机向量 $\boldsymbol{X}$ 的协方差阵是对称非负定矩阵

(2)(2) $\mathrm{Cov}(\boldsymbol{AX},\boldsymbol{BY})=\boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X},\boldsymbol{Y})\,\boldsymbol{B}$

其中 $\boldsymbol{A},\boldsymbol{B}$ 为大小适合矩阵运算的常数矩阵。

内容

(1) 随机向量

(2) 多元正态分布概述

(3) 多元正态分布的参数估计

(4) 常用分布与抽样分布

多元正态分布

- 设 $X_1, \cdots, X_m$ 为 m 个相互独立标准正态变量, $\boldsymbol{X} = (X_1, \cdots, X_m)$ 为这 m 个随机变量构成的随 机向量;
- 设 $\boldsymbol{\mu}$ 为 p 维常数向量, $\boldsymbol{A}$ 为 $p \times m$ 维常数矩阵;
- 则称 $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{\mu}$ 的分布为 p 元正态分布, 或称 $\boldsymbol{Y}$ 为 p 维正态随机向量, 记为 $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{AA}')$

多元正态分布

p 元随机向量 $\boldsymbol{X}=\left(X_1,X_2,\cdots,X_p\right)'$，$\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 且 $\boldsymbol{\Sigma}$ 正定 (为了保证 $\boldsymbol{\Sigma}^{-1}$ 存在), $\boldsymbol{X}$的联合概率密度函数为:

$$f\left(x_1,x_2,\cdots,x_p\right)=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},\ \boldsymbol{\Sigma}>\boldsymbol{0}$$

则称 $\boldsymbol{X}=\left(X_1,X_2,\cdots,X_p\right)'$ 遵从 p 元正态分布, 也称 $\boldsymbol{X}$ 为 p 元正态变量, 记为:

$$\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

设 $\boldsymbol{X}\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 且 $\boldsymbol{\Sigma}$ 正定（为了保证 $\boldsymbol{\Sigma}^{-1}$ 存在）, 那么 $\boldsymbol{X}$ 的联合密度函数为

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{\Sigma})^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$f(x)=\frac{1}{(2\pi)^{1/2}|\sigma^2|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'(\sigma^2)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}}(\sigma>0,-\infty<x<\infty)$$

二元正态分布

例1.1 设$\boldsymbol{X}=\begin{bmatrix}X_1\\X_2\end{bmatrix}$服从二元正态分布，利用参数$\mu_1=E(X_1)$、$\mu_2=E(X_2)$、$\sigma_1=\sqrt{D(X_1)}$、$\sigma_2=\sqrt{D(X_2)}$、$\rho=\frac{Cov(X_1,X_2)}{\sigma_1\sigma_2}$来表示X的联合密度。

解 我们可以将协方差矩阵写作

$$\boldsymbol{\Sigma}=\begin{bmatrix}{\sigma_1}^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & {\sigma_2}^2\end{bmatrix}$$

从而其行列式为

$$|\boldsymbol{\Sigma}|={\sigma_1}^2\,{\sigma_2}^2(1-\rho^2)$$

其逆矩阵为

$$\boldsymbol{\Sigma^{-1}}=\frac{1}{{\sigma_1}^2\,{\sigma_2}^2(1-\rho^2)}\begin{bmatrix}{\sigma_2}^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & {\sigma_1}^2\end{bmatrix}$$

将其代入密度公式中可以得到X的联合密度为

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}}exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

多元正态向量的性质

设$X=(X_1,\cdots,X_p)'\sim N_p(\mu,\Sigma)$。

(1) 若Σ为对角矩阵，则$X_1,\cdots,X_p$ 独立。

(2) X的任意边缘分布仍然为正态分布。特别的，如果将X、μ、Σ作如下划分：

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}^{(1)}\\ \boldsymbol{X}^{(2)}\end{bmatrix}_{p-q}^{q} \qquad \boldsymbol{\mu}=\begin{bmatrix}\boldsymbol{\mu}^{(1)}\\ \boldsymbol{\mu}^{(2)}\end{bmatrix} \qquad \boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$$

其中$\boldsymbol{X}^{(1)}$与$\boldsymbol{\mu}^{(1)}$为p维向量，$\boldsymbol{X}^{(2)}$与$\boldsymbol{\mu}^{(2)}$为$p-q$维向量，$\boldsymbol{\Sigma}_{11}$为$q\times q$维矩阵，$\boldsymbol{\Sigma}_{12}$为$q\times(p-q)$维矩阵，$\boldsymbol{\Sigma}_{21}$为$(p-q)\times q$维矩阵，$\boldsymbol{\Sigma}_{22}$为$(p-q)\times(p-q)$维矩阵。则$\boldsymbol{X}^{(1)}\sim \boldsymbol{N_q}(\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}_{11})$, $\boldsymbol{X}^{(2)}\sim \boldsymbol{N_{p-q}}\left(\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma}_{22}\right)$。顺便指出，$\boldsymbol{X}^{(1)}$与$\boldsymbol{X}^{(2)}$相互独立当且仅当$\boldsymbol{\Sigma}_{12}$为零矩阵。

(3) 设**A**是$s\times p$阶常数矩阵，$\boldsymbol{d}$为s维常数向量，则$\boldsymbol{AX}+\boldsymbol{d}$也服从正态分布，且

$$\boldsymbol{AX}+\boldsymbol{d}\sim \boldsymbol{N}_{\mathrm{s}}(\boldsymbol{A\mu}+\boldsymbol{d},\boldsymbol{A\Sigma A}')$$

(4) 若Σ为正定阵，则$(\boldsymbol{X}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\sim\boldsymbol{\chi}^{2}(\boldsymbol{p})$

内容

(1) 随机向量

(2) 多元正态分布概述

(3) 多元正态分布的参数估计

(4) 常用分布与抽样分布

多元样本数字特征

考虑p元总体$\boldsymbol{X}\sim \boldsymbol{N_p(\mu,\Sigma)}$，设$\boldsymbol{X}_{(1)},\cdots,\boldsymbol{X}_{(n)}$为来自$p$元总体的简单随机样本，其中$\boldsymbol{X}_{(\mathrm{i})}=\left(x_{i1},\cdots,x_{ip}\right)'(i=1,\cdots,n)$。

样本均值向量$\bar{\boldsymbol{X}}$ 的定义为

$$\bar{\boldsymbol{X}}=\frac{\mathbf{1}}{\mathrm{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{X}_{(\boldsymbol{i})}=\left(\bar{x}_1,\cdots,\bar{x}_p\right)'=\frac{\mathbf{1}}{\mathrm{n}}\boldsymbol{X}'\mathbf{1}_{\boldsymbol{n}}$$

其中$\bar{x}_i=\frac{\mathbf{1}}{\mathrm{n}}\sum_{\boldsymbol{b=1}}^{\boldsymbol{n}}\boldsymbol{x_{bi}}\,(i=1,\cdots p)$，$\mathbf{1}_{\boldsymbol{n}}$是一个n维的分量全为1的向量。

$\bar{\boldsymbol{X}}$是$\boldsymbol{\mu}$的无偏估计。

多元样本数字特征

样本离差阵的定义为

$$\boldsymbol{A}=\sum_{b=1}^{n}\left(\boldsymbol{X}_{(b)}-\bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_{(b)}-\bar{\boldsymbol{X}}\right)'=\boldsymbol{X}'\boldsymbol{X}-n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}'=\boldsymbol{X}'\left[\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n{}'\right]\boldsymbol{X}=\left(a_{ij}\right)_{p\times p}$$

其中$a_{ij}=\sum_{\boldsymbol{b=1}}^{\boldsymbol{n}}(\boldsymbol{x_{bi}}-\bar{x}_i)(\boldsymbol{x_{bj}}-\bar{x}_j)\,(i,j=1,\cdots p)$。

样本协方差矩阵的定义为

$$\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{A}=\left(\boldsymbol{s_{ij}}\right)_{\boldsymbol{p}\times\boldsymbol{p}}(\text{或者}\boldsymbol{S}^*=\frac{1}{n}\boldsymbol{A})$$

此时$s_{ij}=\frac{1}{n-1}\sum_{b=1}^{n}(x_{bi}-\bar{x}_i)\left(x_{bj}-\bar{x}_j\right)(i,j=1,\cdots,p)$。

样本相关阵的定义为

$$\boldsymbol{R}=\left(\boldsymbol{r_{ij}}\right)_{\boldsymbol{p}\times\boldsymbol{p}}$$

其中$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}=\frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}},\ (i,j=1,\cdots,p)$。

$\boldsymbol{S}$是$\boldsymbol{\Sigma}$的无偏估计。

正态总体下的性质

定理 1.1 设 $\bar{\boldsymbol{X}}$ 和 $\boldsymbol{A}$ 分别为 p 元正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的样本均值向量和样本离差阵, 则

(1) $\bar{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$;

(2) 若设 $\boldsymbol{Z_1}, \cdots, \boldsymbol{Z_{n-1}}$ 独立同 $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ 分布，则 $\boldsymbol{A}$ 与 $\sum_{t=1}^{n-1} \boldsymbol{Z}_t \boldsymbol{Z}_t'$ 同分布;

(3) $\bar{\boldsymbol{X}}$ 与 $\boldsymbol{A}$ 相互独立;

(4) $\boldsymbol{A}$ 为正定阵的充要条件是 $n > p$ 。

注意到这时 $\boldsymbol{A}$ 是随机矩阵, 因此" A 为正定阵"这句话的含义事实上是" A 为正定阵"这个事件的概率为 1。

内容

(1) 随机向量

(2) 多元正态分布概述

(3) 多元正态分布的参数估计

(4) 常用分布与抽样分布

常用分布与抽样分布

- 在一元正态总体中，用于检验参数μ、σ的抽样分布有χ^2分布、t分布以及F分布。
- 在多元正态总体中，与之对应的分布为Wishart分布、Hoteling T^2分布以及Wilks分布。

卡方分布

在数理统计中, 若 $X_i \sim N(0,1)(i=1,2,\cdots,n)$, 且相互独立, 则 $\sum_{i=1} X_i^2$ 所遵从的分布为自由度为 n 的 χ^2 分布 (chi-squared distribution), 记为 $\chi^2(n)$ 。

如果从一元正态总体 $N(\mu,\sigma^2)$ 中抽取 n 个简单随机样本 $X_1,\cdots,X_n$, 我们用样本方差

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

来估计 σ^2, 此时 $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2(n-1)$ 。因此, 可以得到 $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$ 。那么对 p 元正态总体, 样本协方差阵 $\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{A}$ 又有怎样的分布呢?

Wishart分布

定义 1.7 设 $\boldsymbol{X}_{(b)} \sim N_p(\boldsymbol{\mu_b}, \boldsymbol{\Sigma})(b = 1, \cdots, n)$ 是相互独立的 n 个 p 维正态变量，记 $\boldsymbol{X} = \left(\boldsymbol{X}_{(1)}, \cdots, \boldsymbol{X}_{(n)}\right)'$ 为一个 $n \times p$ 矩阵, 则称随机阵 $\boldsymbol{W} = \sum_{b=1}^{n} \boldsymbol{X}_{(b)} \boldsymbol{X}'_{(b)} = \boldsymbol{X}'\boldsymbol{X}$ 的分布为自由度为 n 的 p 维非中心 Wishart 分布, 记为 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$ 。其中 $\boldsymbol{\Delta}$ 一般称为非中心参数, $\boldsymbol{\Delta} = \sum_{b=1}^{n} \boldsymbol{\mu_b} \boldsymbol{\mu'_b}$ 。当 $\boldsymbol{\mu_b} = 0$ 时, 我们一般称为中心 Wishart 分布, 记为 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$ 。

当 $p = 1$, $\mu_b = 0$ 时， $X_b \sim N(0, \sigma^2)$ ， 此时 $W = W_1(n, \sigma^2) = \sum_{b=1}^{n} X_{(b)}^2 \sim \sigma^2 \chi^2(n)$ 。也就是说 $W_1(n, 1)$ 就是 $\chi^2(n)$ 。因此 Wishart 分布是 χ^2 分布在多元正态情形下的推广。

Wishart分布

(1) 设 $\boldsymbol{X_{(b)}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})(b = 1, \cdots, n)$ 相互独立，则样本离差阵 $\boldsymbol{A}$ 服从 Wishart 分布, 即

$$\boldsymbol{A} = \sum_{b=1}^{n} \left(\boldsymbol{X_{(b)}} - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X_{(b)}} - \bar{\boldsymbol{X}}\right)' \sim W_p(n-1, \boldsymbol{\Sigma})$$

(2) 设 $\boldsymbol{W_i} \sim W_p(n_i, \boldsymbol{\Sigma})(i = 1, \cdots, k)$ 相互独立, 若令 $n = n_1 + \cdots + n_k$, 则有

$$\sum_{i=1}^{k} \boldsymbol{W_i} \sim W_p(n, \boldsymbol{\Sigma})$$

这个性质一般称为 Wishart 分布关于自由度 n 具有可加性，这点与 χ^2 分布类似。

(3) 设 p 阶随机阵 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma})$, $\boldsymbol{C}_{m\times p}$ 为常数矩阵, 则

$$\boldsymbol{CWC'} \sim W_m(n, \boldsymbol{C\Sigma C'})$$

特别的，如果取 $\boldsymbol{C}$ 为向量 $\boldsymbol{l} = (l_1, \cdots, l_p)'$，则有 $\boldsymbol{l'Wl} \sim W_1(n, \boldsymbol{l'\Sigma l})$，也即 $\frac{\boldsymbol{l'Wl}}{\boldsymbol{l'\Sigma}l} \sim \chi^2(n)$

t分布

在一元统计中我们学过, 若 $X \sim N(0,1), Y \sim \chi^2(n)$, 且 X 与 Y 独立, 则随机变量 $t = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布, 也称为学生分布。我们还学过, 如果将 t 平方, 就得到

$$t^2 = \frac{nX^2}{Y} \sim F(1,n)$$

即 $t^2(n)$ 服从第一自由度为 1 、第二自由度为 n 的中心 F 分布。下面仿照一元情形将 t^2 的分 布推广到 p 元总体的情形。

Hotelling T^2分布

定义 1.8 设 $\boldsymbol{W} \sim W_p(n, \boldsymbol{\Sigma}), \boldsymbol{X} \sim N_p(0, \boldsymbol{\Sigma}), n \geq p, \boldsymbol{\Sigma} > 0$, 且 $\boldsymbol{W}$ 与 $\boldsymbol{X}$ 相互独立, 则称随机变量 $T^2 = n\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X}$ 所服从的分布称为第一自由度为 p, 第二自由度为 n 的 Hotelling T^2 分布, 记为

$$T^2 \sim T^2(p, n)$$

注意我们可以证明 T^2 分布只与 n, p 有关，与 $\boldsymbol{\Sigma}$ 无关，因此在表示 T^2 分布的记号中没有 $\boldsymbol{\Sigma}$。

T^2 分布与 F 分布也有一定的关系。在一元统计中, 如果 $t = \frac{X}{\sqrt{Y/n}} \sim t(n)$，则 $t^2 = \frac{X^2}{Y/n} \sim F(1, n)$。推广到 p 元情形， 这个关系是 $\frac{n-p+1}{pn} T^2(p, n) = F(p, n-p+1)$。

Hotelling T^2分布

(1) 设 $\boldsymbol{X_{(b)}}(b=1,\cdots,n)$ 是从 p 维正态总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 中抽取的 n 个随机样本， $\bar{\boldsymbol{X}}$ 为样本均值向量, $\boldsymbol{A}$ 为样本离差阵, 则统计量

$$\begin{aligned} T^2 &= (n-1)[\sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})]'\boldsymbol{A^{-1}}[\sqrt{n}(\bar{\boldsymbol{X}}-\boldsymbol{\mu})] \\ &= n(n-1)(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\boldsymbol{A^{-1}}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \\ &\sim T^2(p,n-1) \end{aligned}$$

(2) 设有两个 p 维正态总体 $N_p(\boldsymbol{\mu_1},\boldsymbol{\Sigma}), N_p(\boldsymbol{\mu_2},\boldsymbol{\Sigma})$, 从这两个总体中抽出容量分别为 n_1 和 n_2 的两个样本。记 $\bar{\boldsymbol{X}}_{\boldsymbol{1}},\bar{\boldsymbol{X}}_{\boldsymbol{2}}$ 为两样本的均值向量, $\boldsymbol{S}_1,\boldsymbol{S}_2$ 为两样本协方差阵，并记

$$\boldsymbol{S_p} = \frac{n_1\boldsymbol{S_1}+n_2\boldsymbol{S_2}}{n_1+n_2-2}$$

若 $\mu_1=\mu_2$ ， 则

$$\frac{n_1n_2}{n_1+n_2}(\bar{\boldsymbol{X}}_{\boldsymbol{1}}-\bar{\boldsymbol{X}}_2)'\boldsymbol{S_p}(\bar{\boldsymbol{X}}_{\boldsymbol{1}}-\bar{\boldsymbol{X}}_{\boldsymbol{2}}) \sim T^2(p,n_1+n_2-2)$$

F分布

在一元统计学中, 若 $X \sim \chi^2(m), Y \sim \chi^2(n)$, 且 X 与 Y 相互独立, 则称 $F = \frac{X/m}{Y/n}$ 所遵从的分布为第一自由度为 m 、第二自由度为 n 的中心 F 分布, 记为 $F \sim F(m, n)$ 。F 分布本质上是从正态总体 $N(\mu, \sigma^2)$ 中随机抽取的两个样本方差的比。

Wilks Λ分布

定义 1.9 设 $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$， 则称协方差阵的行列式 $|\boldsymbol{\Sigma}|$ 为 $\boldsymbol{X}$ 的广义方差。再设 $\boldsymbol{A_1} \sim W_p(n_1, \boldsymbol{\Sigma}), \boldsymbol{A_2} \sim W_p(n_2, \boldsymbol{\Sigma})(\boldsymbol{\Sigma} > 0, n_1 \geq p)$, 且 $\boldsymbol{A_1}$ 与 $\boldsymbol{A_2}$ 独立, 则称

$$\Lambda = \frac{|A_1|}{|A_1 + A_2|}$$

为 Wilks 统计量或 Λ 统计量，其所遵从的分布称为 Wilks 分布, 记为

$$\Lambda \sim \Lambda(p, n_1, n_2)$$

Wilks Λ近似分布

p	n_2	统计量F	F的自由度
任意	1	$\frac{n_1-p+1}{p}\frac{1-\Lambda}{\Lambda}$	p, n_1-p-1
任意	2	$\frac{n_1-p+1}{p}\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}$	$2p, 2(n_1-p+1)$
1	任意	$\frac{1-\Lambda}{\Lambda}\frac{n_1}{n_2}$	n_2, n_1
2	任意	$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\frac{n_1-1}{n_2}$	$2n_2, 2(n_1-1)$

Wilks Λ近似分布及性质

当 $n_2 > 2, p > 2$ 时, 我们有这样的近似分布：

$$当\ n_1 \to \infty, -\left(n_1 - \frac{1}{2}(p - n_2 + 1)\right)\ln\Lambda \sim \chi^2(pn_2)\ 。$$

此外，类似于 F 分布中 $F(n,m)$ 与 $\frac{1}{F(m,n)}$ 同分布, Λ 分布也有一个类似的性质：若 $n_2 < p$, 则 $\Lambda(p, n_1, n_2) = \Lambda(n_2, p, n_1 + n_2 - p)$ 。

作业

为什么样本方差

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

分母是N-1而不是N？