

# 第5章

## 多元正态分布的参数估计

李高荣

北京师范大学统计学院

E-mail: [ligaorong@bnu.edu.cn](mailto:ligaorong@bnu.edu.cn)



## 1 多元正态分布的参数估计

- 多元正态分布样本统计量
- 极大似然估计

## 2 多元正态分布的参数估计的性质

## 3 相关系数的估计与应用

- 相关系数的极大似然估计
- 样本相关系数的精确分布
- $\rho \neq 0$ 时样本相关系数的分布
- 精确分布下 $\rho$ 的假设检验
- 精确分布下 $\rho$ 的区间估计

## 4 样本相关系数的渐近正态分布

## 5 样本偏相关系数



- 扫描二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

## 多元正态分布样本统计量

- 设  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$  为  $p$  元正态总体  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  的样本容量为  $n$  的简单随机样本矩阵, 其中  $\boldsymbol{\mu} \in \mathbb{R}^p$ ,  $\boldsymbol{\Sigma} > 0$ ,  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  和  $n > p$ 。

### 样本均值向量

设  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})' (i = 1, \dots, n)$  为一组随机样本, 则称

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}' \mathbf{1}_n = (\bar{x}_1, \dots, \bar{x}_p)'$$

为 **样本均值向量**, 其中  $\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{ik}, k = 1, \dots, p$ 。

## 样本离差阵

设  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  ( $i = 1, \dots, n$ ) 为一组随机样本, 则称

$$\begin{aligned}\mathbf{V} &= \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' = \mathbf{X}'\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}' \\ &= \mathbf{X}'\left[\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right]\mathbf{X}\end{aligned}$$

为**样本离差阵**, 其中  $\mathbf{X}$  为  $n \times p$  的样本矩阵。

$$\begin{aligned}\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' &= \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \\ &= \mathbf{V} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\end{aligned}$$

## 样本协方差阵

设  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  ( $i = 1, \dots, n$ ) 为一组随机样本, 则称

$$\mathbf{S} = \frac{1}{n-1} \mathbf{V} = (s_{ij})_{p \times p}$$

为**样本协方差阵**。其中  $s_{kk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$  ( $k = 1, \dots, p$ ) 称为第  $k$  个样本变量  $\mathbf{x}_k$  的**样本方差**;  $\sqrt{s_{kk}}$  称为变量  $\mathbf{x}_k$  的**样本标准差**。

**作用:**

- 估计多元正态总体分布的协方差阵
- 对有关总体分布均值向量和协方差阵的假设进行检验

## 样本相关阵

设  $\mathbf{S} = (s_{ij})_{p \times p}$  为样本协方差阵，则称  $\tilde{\mathbf{R}} = (r_{ij})_{p \times p}$  为样本相关阵，其中

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}, \quad i, j = 1, \dots, p.$$

若记  $\tilde{\mathbf{D}}^{1/2} = \text{diag}(\sqrt{s_{11}}, \dots, \sqrt{s_{pp}})$  为样本标准差对角阵，则

$$\tilde{\mathbf{R}} = \tilde{\mathbf{D}}^{-1/2} \mathbf{S} \tilde{\mathbf{D}}^{-1/2}, \quad \mathbf{S} = \tilde{\mathbf{D}}^{1/2} \tilde{\mathbf{R}} \tilde{\mathbf{D}}^{1/2}.$$

**例1:** 设从某小学某班随机抽取10个同学了解该班学生身高和体重的情况, 每个同学的身高为 $X_1$ (单位: 厘米), 体重为 $X_2$ (单位: 公斤), 具体数值如下:

$$\mathbf{X} = \begin{pmatrix} 142.0 & 32.7 \\ 143.0 & 33.5 \\ 143.0 & 33.0 \\ 145.0 & 34.2 \\ 148.0 & 37.4 \\ 149.0 & 43.1 \\ 138.0 & 31.5 \\ 144.0 & 43.2 \\ 150.0 & 37.5 \\ 153.0 & 51.3 \end{pmatrix}.$$



# 多元正态分布样本统计量

```
X1 = c(142.0, 143.0, 143.0, 145.0, 148.0,  
       149.0, 138.0, 144.0, 150.0, 153.0)  
X2 = c(32.7, 33.5, 33.0, 34.2, 37.4,  
       43.1, 31.5, 43.2, 37.5, 51.3)  
X = cbind(X1, X2)  
> apply(X, 2, mean)  
      X1      X2  
145.50  37.74  
> cov(X)  
      X1      X2  
X1 19.83333 22.11111  
X2 22.11111 39.98933  
> cor(X)  
      X1      X2  
X1 1.0000000 0.7851283  
X2 0.7851283 1.0000000
```

## 重要定理：定理5.1.1

### 定理5.1.1

设 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  ( $i = 1, \dots, n$ ) 为来自 $p$ 元正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一组随机样本， $\bar{\mathbf{x}}$ 为样本均值向量， $\mathbf{V}$ 为样本离差阵，则

- ①  $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ ;
- ②  $\mathbf{V} \sim W_p(n-1, \boldsymbol{\Sigma})$ ，其中 $n > p$ ，并且 $W_p(n-1, \boldsymbol{\Sigma})$ 是自由度是 $n-1$ 的Wishart分布；
- ③  $\bar{\mathbf{x}}$ 与 $\mathbf{V}$ 相互独立。
- ④  $\Pr(\mathbf{V} > 0) = 1$ 的充要条件是 $n > p$ 。

## 定理5.1.1的证明

**证明：** 设 $\Gamma$ 是第1个行向量为 $(1/\sqrt{n}, \dots, 1/\sqrt{n})$  的 $n$ 阶正交矩阵，具有如下形式：

$$\Gamma = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \gamma_{21} & \cdots & \gamma_{2n} \\ \vdots & & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{pmatrix} = (\gamma_{ij})_{n \times n}.$$

令

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_n \end{pmatrix} = \Gamma \begin{pmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix} = \Gamma \mathbf{X},$$

## 定理5.1.1的证明

证明(续): 则有

$$\mathbf{y}_i = \sum_{k=1}^n \gamma_{ik} \mathbf{x}_k, \quad (i = 1, \dots, n)$$

为 $p$ 维随机向量。

因为 $\mathbf{y}_i$ 是 $p$ 维正态随机向量 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的线性组合, 则 $\mathbf{y}_i$ 也是 $p$ 维正态随机向量。  
由 $\Gamma$ 的定义, 则有

$$E(\mathbf{y}_i) = \sum_{k=1}^n \gamma_{ik} E(\mathbf{x}_k) = \begin{cases} \sqrt{n}\boldsymbol{\mu}, & \text{当 } i = 1 \text{ 时,} \\ \mathbf{0}, & \text{当 } i \neq 1 \text{ 时;} \end{cases}$$

$$\text{Cov}(\mathbf{y}_i, \mathbf{y}_j) = \sum_{k=1}^n \gamma_{ik} \gamma_{jk} \boldsymbol{\Sigma} = \begin{cases} \boldsymbol{\Sigma}, & \text{当 } i = j \text{ 时,} \\ \mathbf{0}, & \text{当 } i \neq j \text{ 时.} \end{cases}$$

## 定理5.1.1的证明

(1) 因为  $\mathbf{y}_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i = \sqrt{n}\bar{\mathbf{x}} \sim N_p(\sqrt{n}\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 。故有：

$$\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n).$$

(2) 由于  $\mathbf{y}_1 = \sqrt{n}\bar{\mathbf{x}}$ ，则

$$\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' + n\bar{\mathbf{x}} \bar{\mathbf{x}}' = \mathbf{V} + \mathbf{y}_1 \mathbf{y}_1'.$$

因为  $\mathbf{Y}'\mathbf{Y} = (\mathbf{X}'\boldsymbol{\Gamma}')(\boldsymbol{\Gamma}\mathbf{X}) = \mathbf{X}'\mathbf{X}$ ，则

$$\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i'.$$

## 定理5.1.1的证明

因此, 有:

$$\begin{aligned}\mathbf{V} &= \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' - \mathbf{y}_1 \mathbf{y}_1' = \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i' - \mathbf{y}_1 \mathbf{y}_1' \\ &= \sum_{i=2}^n \mathbf{y}_i \mathbf{y}_i' \sim W_p(n-1, \Sigma).\end{aligned}$$

- (3) 因为  $\bar{\mathbf{x}} = \mathbf{y}_1 / \sqrt{n}$  是  $\mathbf{y}_1$  的函数, 而  $\mathbf{V} = \sum_{i=2}^n \mathbf{y}_i \mathbf{y}_i'$  是  $\mathbf{y}_2, \dots, \mathbf{y}_n$  的函数, 且  $\mathbf{y}_1$  与  $\mathbf{y}_2, \dots, \mathbf{y}_n$  相互独立, 故  $\bar{\mathbf{x}}$  与  $\mathbf{V}$  相互独立。
- (4) 令  $\mathbf{Y}_* = (\mathbf{y}_2, \dots, \mathbf{y}_n)'$ 。由  $\mathbf{V} \stackrel{d}{=} \mathbf{Y}_*' \mathbf{Y}_*$ , 知:  $\text{rank}(\mathbf{V}) = \text{rank}(\mathbf{Y}_*' \mathbf{Y}_*) = \text{rank}(\mathbf{Y}_*)$ 。而  $\mathbf{Y}_*$  为  $(n-1) \times p$  随机阵, 各行向量独立同分布。故,  $\Pr(\mathbf{V} > 0) = \Pr(\text{rank}(\mathbf{V}) = p) = \Pr(\text{rank}(\mathbf{Y}_*) = p) = 1 \Leftrightarrow n-1 \geq p$ , 即  $n > p$ 。

# 极大似然估计(MLE)

- 设 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 是来自多元正态总体 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的随机样本, 其中 $n > p, \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} > 0$
- 样本 $\mathbf{x}_i (i = 1, \dots, n)$ 的似然函数记为

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^n \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\ &= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\ &= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp \left[ -\frac{1}{2} \text{tr} \left( \boldsymbol{\Sigma}^{-1} \left\{ \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' \right\} \right) \right] \\ &= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp \left[ -\frac{1}{2} \text{tr} \left( \boldsymbol{\Sigma}^{-1} \{ \mathbf{V} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \} \right) \right]. \end{aligned}$$

- 首先给定  $\Sigma > 0$  时, 求  $\mu$  的极大似然估计, 即求对数似然函数  $\ln L(\mu, \Sigma)$  的极大值点
- 给定  $\Sigma > 0$ , 关于  $\mu$  的对数似然函数为:

$$\begin{aligned}\ln L(\mu, \Sigma) &= -\frac{np}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \text{tr} \left( \Sigma^{-1} \{ \mathbf{V} + n(\bar{\mathbf{x}} - \mu)(\bar{\mathbf{x}} - \mu)' \} \right) \\ &= -\frac{np}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \text{tr}(\Sigma^{-1} \mathbf{V}) - \frac{n}{2} \left[ (\bar{\mathbf{x}} - \mu)' \Sigma^{-1} (\bar{\mathbf{x}} - \mu) \right] \\ &\leq -\frac{np}{2} \ln 2\pi - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \text{tr}(\Sigma^{-1} \mathbf{V}).\end{aligned}$$

- 不等式中等号成立当且仅当  $\mu = \bar{\mathbf{x}}$ 。因此, 总体均值向量  $\mu$  的MLE为  $\hat{\mu} = \bar{\mathbf{x}}$ 。



- 由定理5.1.1的结论(1):  $\bar{\mathbf{x}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$ , 可知 $\bar{\mathbf{x}}$ 是 $\boldsymbol{\mu}$ 的无偏估计。
- 将 $\boldsymbol{\mu}$ 用它的MLE替换, 得到 $\boldsymbol{\Sigma}$ 的似然函数为:

$$L(\bar{\mathbf{x}}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} \exp \left[ -\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{V}) \right].$$

- 利用矩阵的性质: 存在可逆对称阵 $\mathbf{C}$ , 使得 $\mathbf{B} = \mathbf{C}\mathbf{C}$ 。记 $\tilde{\boldsymbol{\Sigma}} = \mathbf{C}^{-1}\boldsymbol{\Sigma}\mathbf{C}^{-1}$ , 则 $|\boldsymbol{\Sigma}| = |\mathbf{B}||\tilde{\boldsymbol{\Sigma}}|$ 。
- 令 $\boldsymbol{\Sigma}^{-1/2}\mathbf{V}\boldsymbol{\Sigma}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$ , 其中 $\mathbf{U}$ 是正交矩阵,  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_p)$ 是对角矩阵。

- 利用  $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ , 可知:

$$|\mathbf{V}|^{n/2} = |\boldsymbol{\Sigma}|^{n/2} \prod_{k=1}^p \lambda_k^{n/2},$$

$$\text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{V}) = \text{tr}(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} \mathbf{V}) = \sum_{k=1}^p \lambda_k.$$

- 则似然函数可简化为:

$$L(\bar{\mathbf{x}}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2} |\mathbf{V}|^{n/2}} \prod_{k=1}^p \left[ \lambda_k^{n/2} \exp \left\{ -\frac{\lambda_k}{2} \right\} \right].$$

- 由于  $f(x) = x^{n/2} \exp\{-x/2\}$  在  $x = n$  处取最大值, 可知上式在  $\lambda_1 = \dots = \lambda_p = n$  时取最大值。
- 因此,  $\Sigma$  的极大似然估计  $\hat{\Sigma}$  满足条件:

$$\hat{\Sigma}^{-1/2} \mathbf{V} \hat{\Sigma}^{-1/2} = n \mathbf{I}_p.$$

- 则,  $\Sigma$  的MLE为:  $\hat{\Sigma} = \mathbf{V}/n$ 。
- 可见, 似然函数的最大值为:

$$L(\bar{\mathbf{x}}, \hat{\Sigma}) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \frac{1}{|\hat{\Sigma}|^{n/2}}.$$

# 判断估计量好坏的准则

在统计学中，评价参数的点估计通常有一些准则：

- ① 无偏性
- ② 充分性
- ③ 相合性
- ④ 完备性
- ⑤ 有效性
- ⑥ minimax性

## 无偏性

令  $\mathbf{x}_1, \dots, \mathbf{x}_n$  是来自总体  $X \sim f(\mathbf{x}, \boldsymbol{\theta})$  的一组独立随机样本，其中  $\boldsymbol{\theta}$  是未知的参数向量。假设  $\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n)$  是  $\boldsymbol{\theta}$  的一个估计，并且它是一个统计量。对于不同的样本  $\mathbf{x}_1, \dots, \mathbf{x}_n$ ，估计  $\hat{\boldsymbol{\theta}}$  取不同的值，如果  $\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n)$  的均值等于未知参数向量  $\boldsymbol{\theta}$ ，即

$$E[\hat{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n)] = \boldsymbol{\theta}, \quad \text{对一切可能的 } \boldsymbol{\theta} \text{ 成立,}$$

则称  $\hat{\boldsymbol{\theta}}$  为  $\boldsymbol{\theta}$  的 **无偏估计**。

- 根据定理5.1.1的结论(2):  $\mathbf{V} \sim W_p(n-1, \Sigma)$ , 则  $E(\mathbf{V}) = (n-1)\Sigma$ , 所以极大似然估计  $\hat{\Sigma} = \mathbf{V}/n$  并不是  $\Sigma$  的无偏估计。
- 记样本协方差阵  $\mathbf{S} = \mathbf{V}/(n-1)$ , 则样本协方差阵  $\mathbf{S}$  才是  $\Sigma$  的无偏估计。

## 定理5.2.1

设  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$  ( $i = 1, \dots, n$ ) 为来自  $p$  元正态总体  $N_p(\mu, \Sigma)$  的一组随机样本,  $n > p$ ,  $\bar{\mathbf{x}}$  为样本均值向量,  $\mathbf{V}$  为样本离差阵,  $\mathbf{S}$  为样本协方差阵, 则  $\mu$  和  $\Sigma$  的MLE分别为  $\hat{\mu} = \bar{\mathbf{x}}$  和  $\hat{\Sigma} = \mathbf{V}/n$ 。进一步,  $\mu$  和  $\Sigma$  的无偏估计分别为  $\bar{\mathbf{x}}$  和  $\mathbf{S}$ 。

- **充分统计量**：就是包含了样本中对兴趣待估参数向量全部信息的统计量。
- 如何判断一个统计量是充分统计量？

## Neyman-Fisher因子判别法则

令 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 是来自总体 $X \sim f(\mathbf{x}, \boldsymbol{\theta})$ 的一组独立随机样本，其中 $\boldsymbol{\theta}$ 是未知的兴趣待估参数向量。设 $t \equiv t(\mathbf{x}_1, \dots, \mathbf{x}_n)$ 为一统计量，若样本的分布密度函数可以分解为：

$$\prod_{i=1}^n f(\mathbf{x}_i, \boldsymbol{\theta}) = g(t, \boldsymbol{\theta}) h(\mathbf{x}_1, \dots, \mathbf{x}_n),$$

其中 $h(\mathbf{x}_1, \dots, \mathbf{x}_n)$ 与 $\boldsymbol{\theta}$ 无关； $g(t, \boldsymbol{\theta})$ 可能与参数向量 $\boldsymbol{\theta}$ 有关，但与样本有关是通过统计量 $t$ 发生关系，则称 $t$ 是 $\boldsymbol{\theta}$ 的充分统计量。



## 定理5.2.2

设 $X \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 并且 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'(i = 1, \dots, n)$ 为来自 $p$ 元正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一组随机样本,  $n > p$ ,  $\bar{\mathbf{x}}$ 为样本均值向量,  $\mathbf{S}$ 为样本协方差阵, 则

- ① 当 $\boldsymbol{\Sigma}$ 已知时,  $\bar{\mathbf{x}}$ 是 $\boldsymbol{\mu}$ 的充分统计量;
- ② 当 $\boldsymbol{\mu}$ 已知时,  $\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$ 是 $\boldsymbol{\Sigma}$ 的充分统计量;
- ③ 当 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 未知时,  $\bar{\mathbf{x}}$ 和 $\mathbf{S}$ 分别为 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的充分统计量。

几点说明:

- 根据样本 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 的联合密度函数和Neyman-Fisher 因子判别法则, 很容易完成定理5.2.2的证明;
- 定理5.2.2说明,  $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 一切“好的”估计都是 $(\boldsymbol{\mu}, \mathbf{S})$  的函数;
- 对多元正态总体而言, 充分统计量的重要性体现在关于 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的所有信息包含在 $\bar{\mathbf{x}}$ 和 $\mathbf{S}$ 中;
- 对于非多元正态总体的情形, 一般是不正确的, 除了 $\bar{\mathbf{x}}$ 和 $\mathbf{S}$ 的信息外, 还有其他有用的样本信息。

## 相合性

令  $\mathbf{x}_1, \dots, \mathbf{x}_n$  是来自总体  $X \sim f(\mathbf{x}, \boldsymbol{\theta})$  的一组独立随机样本，其中  $\boldsymbol{\theta}$  是未知的兴趣待估参数向量， $\boldsymbol{\theta}$  的变化范围为参数空间  $\Theta$ ， $g(\boldsymbol{\theta})$  是  $\boldsymbol{\theta}$  的函数。设  $T_n \equiv T_n(\mathbf{x}_1, \dots, \mathbf{x}_n)$  为样本  $\mathbf{x}_1, \dots, \mathbf{x}_n$  的统计量。如果对任给的  $\varepsilon > 0$ ，有

$$\lim_{n \rightarrow \infty} \Pr\{|T_n - g(\boldsymbol{\theta})| > \varepsilon\} = 0,$$

则称  $T_n$  是  $g(\boldsymbol{\theta})$  的弱相合估计，记为  $T_n \xrightarrow{P} g(\boldsymbol{\theta})$ 。如果

$$\Pr\left\{\lim_{n \rightarrow \infty} T_n = g(\boldsymbol{\theta})\right\} = 1,$$

则称  $T_n$  是  $g(\boldsymbol{\theta})$  的强相合估计，记为  $T_n \xrightarrow{a.s.} g(\boldsymbol{\theta})$ 。

- 注:  $T_n \xrightarrow{a.s.} g(\boldsymbol{\theta}) \Rightarrow T_n \xrightarrow{P} g(\boldsymbol{\theta})$

## 定理5.2.3

在定理5.2.1的假设及记号下,  $\bar{\mathbf{x}}$ 和 $\mathbf{V}/n$ 分别为 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的强(弱)相合估计。

**证明：** 由于  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})' (i = 1, \dots, n)$  为来自  $p$  元正态总体  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  的一组随机样本， $n > p$ ,  $\bar{\mathbf{x}}$  为样本均值向量。从而由 **Kolmogorov 强大数律** (若  $\{X_i\}$  为相互独立同分布的随机变量, 则

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} c \iff E|X_1| < \infty, \quad c = EX_1).$$

显然  $\bar{\mathbf{x}}$  为  $\boldsymbol{\mu}$  的强相合估计。

由于  $\boldsymbol{\Sigma}$  的极大似然估计为  $\frac{1}{n} \mathbf{V}$ , 记

$$\mathbf{V}/n = (v_{ij}/n)_{1 \leq i, j \leq p}$$

且

$$\begin{aligned}\frac{1}{n}v_{ij} &= \frac{1}{n} \sum_{k=1}^n (x_{ki} - \bar{x}_i) (x_{kj} - \bar{x}_j) \\ &= \frac{1}{n} \sum_{k=1}^n x_{ki} x_{kj} - \left( \frac{1}{n} \sum_{k=1}^n x_{ki} \right) \left( \frac{1}{n} \sum_{k=1}^n x_{kj} \right).\end{aligned}$$

显然  $\frac{1}{n} \sum_{k=1}^n x_{ki} \xrightarrow{a.s.} \mu_i$ ,  $\frac{1}{n} \sum_{k=1}^n x_{kj} \xrightarrow{a.s.} \mu_j$ , 而

$$E(x_{ki} x_{kj}) = E(x_{ki} - \mu_i)(x_{kj} - \mu_j) + \mu_i \mu_j = \sigma_{ij} + \mu_i \mu_j.$$

由Cauchy-Schwarz不等式, 则有

$$E |x_{ki}x_{kj}| \leq [Ex_{ki}^2 Ex_{kj}^2]^{1/2} = [(\sigma_{ii} + \mu_i^2)(\sigma_{jj} + \mu_j^2)]^{1/2} < \infty.$$

因此由Kolmogorov 强大数律有

$$\frac{1}{n}v_{ij} \xrightarrow{a.s.} \sigma_{ij}.$$

综合上面的结果, 可证得 $\mathbf{V}/n$ 为 $\Sigma$ 的强相合估计。

## 定理5.2.4

假设  $\mathbf{x}_1, \dots, \mathbf{x}_n$  是来自  $p$  元正态分布  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  的 i.i.d. 随机样本, 记  $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ , 则

$$\mathbf{B}_n = \frac{1}{\sqrt{n}}(\mathbf{V} - n\boldsymbol{\Sigma})$$

收敛于一个正态随机矩阵  $\mathbf{B} = (b_{ij})$ , 其元素均值为 0, 协方差为

$$E(b_{ij}b_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}.$$

定理5.2.4的证明留为作业, 课下完成。



- 完备统计量的理解：
- 不管参数 $\theta$ 怎么变化，统计量 $T$ 的任何一种构造(比如函数 $g(T)$ )，都不可能为0的无偏估计，即无法得到

$$E_{\theta}(g(T)) = 0.$$

除非 $g(T)$ 本身就是0。

- **完备统计量**和**充分统计量**一样，也是为寻求参数估计的优良统计量起重要的作用。
- 考虑参数统计模型 $(\mathcal{X}, \mathcal{B}, \mathcal{F})$ ，其中
  - ①  $\mathcal{X}$ 为样本空间，为样本 $X$ 的一切可能取值；
  - ②  $\mathcal{B}$ 是 $\mathcal{X}$ 的某些子集构成的 $\sigma$ 域，并称 $(\mathcal{X}, \mathcal{B})$ 为**可测空间**；
  - ③  $\mathcal{F} = \{F_{\theta} : \theta \in \Theta\}$ 为关于参数向量 $\theta$ 的分布族。
- 设 $\phi(\mathbf{x})$ 是定义在 $\mathcal{X}$ 上的 **$\mathcal{B}$ 可测函数**
- 则 $\phi(X)$ 的期望为：

$$E_{\theta}[\phi(X)] = \int_{\mathcal{X}} \phi(\mathbf{x}) dF_{\theta}(\mathbf{x}), \quad \theta \in \Theta,$$

其中 $E_{\theta}$ 是强调期望是在参数为 $\theta$ 的分布下进行的。

- 希望 $E_{\theta}[\phi(X)]$ 对每一个固定的 $\theta$ 都有唯一确定的值。
- 这时考虑下面的命题成立：

$$\phi_1(\mathbf{x}) = \phi_2(\mathbf{x}), \quad a.s. F_{\theta} \Leftrightarrow E_{\theta}[\phi_1(X)] = E_{\theta}[\phi_2(X)].$$

- 必要性“ $\Rightarrow$ ”是显然的；
- 充分性“ $\Leftarrow$ ”：如果命题

$$E_{\theta}[\phi(X)] = 0 \Rightarrow \phi(\mathbf{x}) = 0, \quad a.s. F_{\theta}$$

成立，就可证明上面的充分性。

- 因为 $E_{\theta}[\phi_1(X)] = E_{\theta}[\phi_2(X)]$ ，则 $E_{\theta}[\phi_1(X) - \phi_2(X)] = 0$
- 故可利用上面的命题，可证得：

$$\phi_1(\mathbf{x}) = \phi_2(\mathbf{x}), \quad a.s. F_{\theta}.$$

## 完备性

设 $T$ 是一个连续的随机变量，其密度函数为 $f_T(t, \theta)$ ,  $\theta \in \Theta$ ,  $\Theta$ 为参数 $\theta$ 变化的范围。分布密度族 $\{f_T(t, \theta) : \theta \in \Theta\}$ 称为**完备的**，如果对任意的实数函数 $g(T)$ ，当

$$E_{\theta}[g(T)] = \int g(t)f_T(t, \theta)dt = 0$$

对每个 $\theta \in \Theta$ 成立时，能推出 $g(T)$ 几乎处处为0，或 $\Pr_{\theta}(g(T) = 0) = 1$ 对任何 $\theta \in \Theta$ 。如果一个充分统计量的分布密度族是完备的，则称它是**完备充分统计量**。

**注：**定理5.2.2证明了 $(\bar{x}, S)$ 是 $(\mu, \Sigma)$ 的充分统计量，同样也可以证明它们是完备的，详细的证明可见Anderson (2003)。因此称 **$(\bar{x}, S)$  是 $(\mu, \Sigma)$ 的完备充分统计量**。

## 有效性

令 $X_1, \dots, X_n$ 是来自总体 $X \sim f(x, \boldsymbol{\theta})$ 的一组独立随机样本，其中 $\boldsymbol{\theta}$ 是未知的兴趣待估参数向量， $\boldsymbol{\theta}$ 的变化范围为参数空间 $\Theta$ ， $g(\boldsymbol{\theta})$ 是 $\boldsymbol{\theta}$ 的函数。设 $T$ 为样本 $X_1, \dots, X_n$ 的统计量，且是 $g(\boldsymbol{\theta})$ 的无偏估计。如果对 $g(\boldsymbol{\theta})$ 的任一无偏估计 $\tilde{T}$ 都有

$$E[T - g(\boldsymbol{\theta})]^2 \leq E[\tilde{T} - g(\boldsymbol{\theta})]^2, \quad \text{对任意 } \boldsymbol{\theta} \in \Theta,$$

则称 $T$ 为 $g(\boldsymbol{\theta})$ 的**有效估计**。

- **推广：**当待估函数 $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \dots, g_r(\boldsymbol{\theta}))'$ 是一个向量时，设它的某个无偏估计 $\mathbf{T} = (T_1, \dots, T_r)'$ 有协方差矩阵。如果对于任意一个具有协方差矩阵的无偏估计 $\tilde{\mathbf{T}}$ ，有

$$\text{Cov}(\mathbf{T}) \leq \text{Cov}(\tilde{\mathbf{T}}), \quad \text{对任意 } \boldsymbol{\theta} \in \Theta,$$

即 $\text{Cov}(\tilde{\mathbf{T}}) - \text{Cov}(\mathbf{T})$ 是非负定矩阵，则称 $\mathbf{T}$ 为待估函数 $\mathbf{g}(\boldsymbol{\theta})$ 的有效估计。

## 定理5.2.5

在定理1的假设及记号下， $(\bar{\mathbf{x}}, \mathbf{S})$ 为 $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一致最小方差无偏估计(UMVUE)。

# 相关系数的极大似然估计

- 在有了 $\mu$ 和 $\Sigma$ 的极大似然估计 $\hat{\mu} = \bar{x}$ 和 $\hat{\Sigma} = V/n$ 后, 我们是否可以通过使用 $\hat{\mu}$ 和 $\hat{\Sigma}$ 替换前面定义过的回归系数, 相关系数, 条件协方差阵和偏相关系数等中的 $\mu$ 和 $\Sigma$ 来得到相应的MLE? 由下面引理知道这是可以的。

## 引理5.4.1

设 $\theta$ 的极大似然估计为 $\hat{\theta}$ , 若 $\theta \rightarrow \phi(\theta)$ 为一一变换, 则 $\phi(\theta)$ 的极大似然估计为 $\phi(\hat{\theta})$ 。

## 问题

求相关系数的极大似然估计?

# 相关系数的极大似然估计

解: 由相关系数矩阵  $\mathbf{R} = \mathbf{D}^{-1/2} \boldsymbol{\Sigma} \mathbf{D}^{-1/2}$ , 其中

$$\mathbf{D}^{1/2} = \text{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{pp}}),$$

$\boldsymbol{\Sigma} \rightarrow (\mathbf{D}, \mathbf{R})$  一一对应, 因此由  $\boldsymbol{\Sigma}$  的极大似然估计  $\hat{\boldsymbol{\Sigma}}$ , 可知  $\hat{\mathbf{D}} = \sqrt{\text{diag}(\hat{\boldsymbol{\Sigma}})}$ , 则  $\mathbf{R}$  的MLE为:

$$\hat{\mathbf{R}} = \hat{\mathbf{D}}^{-1/2} \hat{\boldsymbol{\Sigma}} \hat{\mathbf{D}}^{-1/2},$$

其  $(i, j)$  元素为,  $i, j = 1, \dots, p$

$$\hat{\rho}_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i) (x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2 \sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}} = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}} =: r_{ij}.$$



## 样本相关系数的精确分布

- 既然相关系数仅与两个变量有关，不失一般性，我们仅考虑 $r_{12}$ 的精确分布。
- 假设总体 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ，由多元正态分布的性质，可知 $(X_1, X_2)'$ 服从二元正态分布，即

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2 \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

其中 $-\infty < \mu_1, \mu_2 < \infty$ ,  $\sigma_1, \sigma_2 > 0$ ,  $-1 < \rho < 1$ 。

## 样本相关系数的精确分布

- 假设下面的样本来自于二元正态分布总体 $(X_1, X_2)'$ , 即

$$\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}, \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}, \dots, \begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix}.$$

- 令 $v_{kl} = \sum_{i=1}^n (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)$ ,  $k, l = 1, 2$ 。由上面的样本, 可定义样本离差矩阵为

$$\mathbf{V} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}.$$

## 样本相关系数的平移不变性

- 样本均值为  $\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{i1}$ ,  $\bar{x}_2 = \frac{1}{n} \sum_{i=1}^n x_{i2}$ 。
- $X_1$  和  $X_2$  的相关系数  $\rho$  的极大似然估计, 即样本相关系数估计定义为:

$$r = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2}} = \frac{v_{12}}{\sqrt{v_{11}}\sqrt{v_{22}}}.$$

## 样本相关系数的平移不变性

- 下面说明样本相关系数 $r$ 与分布的参数 $\mu_1, \mu_2, \sigma_1$ 和 $\sigma_2$  无关。
- 令 $u_{i1} = \frac{x_{i1} - \mu_1}{\sigma_1}$ ,  $u_{i2} = \frac{x_{i2} - \mu_2}{\sigma_2}$ ,  $i = 1, \dots, n$ .
- 则 $\begin{pmatrix} u_{11} \\ u_{12} \end{pmatrix}, \begin{pmatrix} u_{21} \\ u_{22} \end{pmatrix}, \dots, \begin{pmatrix} u_{n1} \\ u_{n2} \end{pmatrix}$  为i.i.d.的样本, 服从分布

$$N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

## 样本相关系数的平移不变性

- 把  $x_{ik} = \sigma_k u_{ik} + \mu_k$  代入样本相关系数, 则有

$$r = \frac{v_{12}}{\sqrt{v_{11}}\sqrt{v_{22}}} = \frac{\sum_{i=1}^n (u_{i1} - \bar{u}_1)(u_{i2} - \bar{u}_2)}{\sqrt{\sum_{i=1}^n (u_{i1} - \bar{u}_1)^2 \sum_{i=1}^n (u_{i2} - \bar{u}_2)^2}}.$$

- 可见,  $r$  的分布与参数  $\mu_1, \mu_2, \sigma_1$  和  $\sigma_2$  无关, 只与  $\rho$  有关。

## 样本相关系数的精确分布

- 讨论 $r$ 的分布时, 假设 $\mu_1 = \mu_2 = 0$ 和 $\sigma_1 = \sigma_2 = 1$ 。
- 由定理5.1.1(3)的证明, 可知

$$\mathbf{V} = \sum_{i=2}^n \mathbf{y}_i \mathbf{y}_i' \sim W_2(n-1, \Sigma), \quad \text{其中 } \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

- 则 $v_{11}, v_{12}, v_{22}$ 可以表示为

$$v_{ij} = \sum_{k=2}^n y_{ki} y_{kj}, \quad i, j = 1, 2,$$

其中 $(y_{k1}, y_{k2})'$ 来自于二元正态总体

$$N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

## 样本相关系数的精确分布

- 样本  $(y_{21}, y_{22}), \dots, (y_{n1}, y_{n2})$  相互独立。
- 进一步, 把  $y_{k1}$  和  $y_{k2}$  表示成

$$y_{k1} = u_k, \quad y_{k2} = \rho u_k + \sqrt{1 - \rho^2} v_k,$$

其中  $\{u_k, v_k, k = 2, \dots, n\}$  i.i.d., 并服从标准正态分布。

- 记  $\mathbf{u} = (u_2, \dots, u_n)'$ ,  $\mathbf{v} = (v_2, \dots, v_n)'$  为  $n - 1$  维向量。
- 以  $\mathbf{u}' / \|\mathbf{u}\|$  为第一行构造一个  $n - 1$  阶正交矩阵  $\mathbf{C}$ 。令  $\mathbf{w} = \mathbf{C}\mathbf{v}$ , 由性质 3.2.3 可知,  $\mathbf{w}$  仍是一个由 i.i.d. 的标准正态随机变量构成的  $n - 1$  维随机向量, 且与  $\mathbf{u}$  相互独立。

# 样本相关系数的精确分布

- 简单计算，有

$$\begin{aligned} r &= \frac{v_{12}}{\sqrt{v_{11}}\sqrt{v_{22}}} = \frac{\sum_{k=2}^n y_{k1}y_{k2}}{\sqrt{\sum_{k=2}^n y_{k1}^2} \sqrt{\sum_{k=2}^n y_{k2}^2}} = \frac{\sum_{k=2}^n u_k(\rho u_k + \sqrt{1-\rho^2}v_k)}{\sqrt{\sum_{k=2}^n u_k^2} \sqrt{\sum_{k=2}^n (\rho u_k + \sqrt{1-\rho^2}v_k)^2}} \\ &= \frac{\mathbf{u}'(\rho \mathbf{u} + \sqrt{1-\rho^2}\mathbf{v})}{\|\mathbf{u}\| \cdot \|\rho \mathbf{u} + \sqrt{1-\rho^2}\mathbf{v}\|} = \frac{\mathbf{u}'\mathbf{C}'(\rho \mathbf{C}\mathbf{u} + \sqrt{1-\rho^2}\mathbf{w})}{\|\mathbf{u}\| \cdot \|\rho \mathbf{C}\mathbf{u} + \sqrt{1-\rho^2}\mathbf{w}\|} \\ &= \frac{\rho\|\mathbf{u}\| + \sqrt{1-\rho^2}w_1}{\sqrt{(\rho\|\mathbf{u}\| + \sqrt{1-\rho^2}w_1)^2 + (1-\rho^2)\sum_{k=2}^{n-1} w_k^2}}. \end{aligned}$$



# 样本相关系数的精确分布

- 由正交矩阵 $\mathbf{C}$ 的定义, 有

$$\mathbf{C} = \begin{pmatrix} \frac{u_1}{\|\mathbf{u}\|} & \frac{u_2}{\|\mathbf{u}\|} & \cdots & \frac{u_{n-1}}{\|\mathbf{u}\|} \\ c_{21} & c_{22} & \cdots & c_{2,n-1} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n-1,1} & c_{n-1,2} & \cdots & c_{n-1,n-1} \end{pmatrix}$$

- 由 $\mathbf{C}$ 的正交性, 则有

$$\frac{u'}{\|\mathbf{u}\|} \mathbf{C}' \mathbf{w} = w_1.$$

- 进一步, 有

$$\begin{aligned}& \|\rho \mathbf{C} \mathbf{u} + \sqrt{1 - \rho^2} \mathbf{w}\|^2 \\&= (\rho \mathbf{C} \mathbf{u} + \sqrt{1 - \rho^2} \mathbf{w})' (\rho \mathbf{C} \mathbf{u} + \sqrt{1 - \rho^2} \mathbf{w}) \\&= \rho^2 \|\mathbf{u}\|^2 + 2\rho\sqrt{1 - \rho^2} \|\mathbf{u}\| \omega_1 + (1 - \rho^2) \sum_{k=1}^{n-1} \omega_k^2 \\&= (\rho \|\mathbf{u}\| + \sqrt{1 - \rho^2} \omega_1)^2 + (1 - \rho^2) \sum_{k=2}^{n-1} \omega_k^2.\end{aligned}$$

## 样本相关系数的精确分布

- 考虑一种特殊情况, 当  $\rho = 0$  时, 则有

$$r = \frac{w_1}{\sqrt{w_1^2 + \sum_{k=2}^{n-1} w_k^2}} \stackrel{d}{=} \frac{t}{\sqrt{t^2 + n - 2}},$$

- $\sum_{k=2}^{n-1} w_k^2 \sim \chi_{n-2}^2$ ,  $t \sim t_{n-2}$ , 根据  $t$  分布的定义, 可以证明上式左右两边有相同的分布。
- 因为  $t_{n-2}$  的密度函数为

$$\frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{(n-2)\pi}\Gamma\left(\frac{n-2}{2}\right)} \left(1 + \frac{t^2}{n-2}\right)^{-\frac{n-1}{2}}.$$

## 样本相关系数的精确分布

- 则样本相关系数  $r = t / \sqrt{t^2 + n - 2}$ , 其逆变换为:

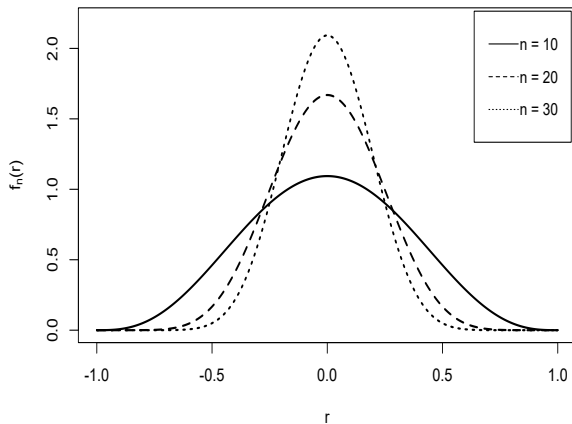
$t = \sqrt{(n-2)r^2/(1-r^2)}$ , 由随机变量函数的密度函数公式, 可得  $r$  的密度函数为:

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

### 定理5.4.1

假设  $\mathbf{x}_1, \dots, \mathbf{x}_n$  是来自  $p$  元正态分布  $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  的独立同分布样本, 如果  $\rho_{ij} = 0$ , 则样本相关系数  $r_{ij}$  的密度函数由式(1)所定义。

## $\rho = 0$ 时, $r$ 的密度函数曲线



# 样本相关系数的精确分布

- 密度函数关于原点对称
- 当 $n > 4$ 时，它在 $r = 0$ 处有众数
- 密度函数是偶函数
- 样本相关系数 $r$ 的奇数阶矩等于0
- $r$ 的偶数阶矩为

$$E(r^{2m}) = \frac{\Gamma\left(\frac{n-1}{2}\right) \Gamma\left(m + \frac{1}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n-1}{2} + m\right)}.$$

## 定理5.4.1的应用：假设检验

对于给定的数对 $(i, j)$ ，考虑如下的假设检验：

$$H_0 : \rho_{ij} = 0, \quad H_1 : \rho_{ij} > 0. \quad (2)$$

- 当 $r_{ij} > r_0$ 时，将拒绝原假设 $H_0$
- 问题：如何确定阈值 $r_0$ ？

## 定理5.4.1的应用：假设检验

对于给定的数对 $(i, j)$ ，考虑如下的假设检验：

$$H_0 : \rho_{ij} = 0, \quad H_1 : \rho_{ij} > 0. \quad (2)$$

- 当 $r_{ij} > r_0$ 时，将拒绝原假设 $H_0$
- 问题：如何确定阈值 $r_0$ ？
- 对给定的显著性水平 $\alpha$ ，当原假设 $H_0$ 为真时，拒绝原假设 $H_0$ 的概率为

$$\int_{r_0}^1 f_n(r) dr = \alpha.$$

- 若感兴趣的是原假设 $H_0$ 对备择假设 $H_1 : \rho_{ij} < 0$ 时，当 $r_{ij} < -r_0$ 时，将拒绝原假设 $H_0$ 。



## 定理5.4.1的应用：假设检验

考虑下面的假设检验：

$$H_0 : \rho_{ij} = 0, \quad H_1 : \rho_{ij} \neq 0. \quad (3)$$

- 当  $r_{ij} > \tilde{r}_0$  或  $r_{ij} < -\tilde{r}_0$  时，将拒绝原假设  $H_0$
- $\tilde{r}_0$  满足：当原假设  $H_0$  为真时，拒绝原假设  $H_0$  的概率为

$$\int_{\tilde{r}_0}^1 f_n(r) dr = \alpha/2.$$

## 定理5.4.1的应用：假设检验

考虑下面的假设检验：

$$H_0 : \rho_{ij} = 0, \quad H_1 : \rho_{ij} \neq 0. \quad (3)$$

- 当  $r_{ij} > \tilde{r}_0$  或  $r_{ij} < -\tilde{r}_0$  时，将拒绝原假设  $H_0$
- $\tilde{r}_0$  满足：当原假设  $H_0$  为真时，拒绝原假设  $H_0$  的概率为

$$\int_{\tilde{r}_0}^1 f_n(r) dr = \alpha/2.$$

- 对于阈值  $r_0$  和  $\tilde{r}_0$ ，可以通过上面的积分确定，也可以通过查表确定。

## 定理5.4.1的应用：假设检验

- 当 $\rho = 0$ 时,  $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$
- 可以通过自由度是 $n-2$ 的 $t$ 分布来确定阈值 $r_0$ 和 $\tilde{r}_0$
- 令 $t_{n-2}(\alpha)$ 为自由度为 $n-2$ 的 $t$ 分布的上 $\alpha$ 分位点
- 对假设检验问题(2), 若

$$\sqrt{n-2} \frac{r_{ij}}{\sqrt{1-r_{ij}^2}} > t_{n-2}(\alpha)$$

时, 则拒绝原假设 $H_0$ 。

## 定理5.4.1的应用：假设检验

- 对于双边假设检验问题(3)，令 $t_{n-2}(\alpha/2)$ 为自由度为 $n-2$ 的 $t$ 分布的上 $\alpha/2$ 分位点，若

$$\sqrt{n-2} \frac{|r_{ij}|}{\sqrt{1-r_{ij}^2}} > t_{n-2}(\alpha/2)$$

时，则拒绝原假设 $H_0$ 。

## 应用：回归系数的检验

- 当  $\rho = 0$  时,  $t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}} \sim t_{n-2}$ , 可用于回归系数是否为0的检验问题
- 假设  $x_{i2} = a + bx_{i1} + \varepsilon_i$ ,  $i = 1, \dots, n$ , 可得斜率  $b$  的极大似然估计为:

$$\hat{b} = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}.$$

## 应用：回归系数的检验

- 计算并整理，有

$$\begin{aligned} & \frac{\sqrt{n-2} \frac{r}{\sqrt{1-r^2}}}{\frac{\hat{b} \sqrt{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}}{\sqrt{\sum_{i=1}^n [x_{i2} - \bar{x}_2 - \hat{b}(x_{i1} - \bar{x}_1)]^2 / (n-2)}}} \\ & \sim t_{n-2}. \end{aligned}$$

- 上面结果可以对回归系数 $b$ 进行单边或双边的显著性检验。

## $\rho \neq 0$ 时样本相关系数的分布

- 分子分母同除 $\sqrt{1 - \rho^2}$ ，简单计算有：

$$\begin{aligned} r &= \frac{\rho \|\mathbf{u}\| + \sqrt{1 - \rho^2} w_1}{\sqrt{(\rho \|\mathbf{u}\| + \sqrt{1 - \rho^2} w_1)^2 + (1 - \rho^2) \sum_{k=2}^{n-1} w_k^2}} \\ &= \frac{\delta + w_1}{\sqrt{(\delta + w_1)^2 + \sum_{k=2}^{n-1} w_k^2}}, \end{aligned}$$

其中 $\delta = \phi \|\mathbf{u}\|$ ,  $\phi = \rho / \sqrt{1 - \rho^2}$ 。

## $\rho \neq 0$ 时样本相关系数的分布

- 由前可知:  $w_1 \sim N(0, 1)$ ,  $\sum_{k=2}^{n-1} w_k^2 \sim \chi_{n-2}^2$
- 因此, 给定 $\mathbf{u}$ 时,  $r$ 的条件分布与 $t/\sqrt{t^2 + n - 2}$ 的分布相同, 即

$$r = \frac{\delta + w_1}{\sqrt{(\delta + w_1)^2 + \sum_{k=2}^{n-1} w_k^2}} \stackrel{d}{=} \frac{t}{\sqrt{t^2 + n - 2}},$$

- 其中 $t = \frac{\delta + w_1}{\sqrt{\sum_{k=2}^{n-1} w_k^2 / (n - 2)}}$ 服从自由度为 $n - 2$ 的非中心化 $t$ 分布,  
非中心化参数为 $\delta = \phi \|\mathbf{u}\|$ ,  $\phi = \rho / \sqrt{1 - \rho^2}$ 。



## $\rho \neq 0$ 时样本相关系数的分布

※ **目标：**当 $\rho \neq 0$ 时，计算样本相关系数 $r$ 的密度函数

■ **解决方案：**为了找到样本相关系数 $r$ 的密度函数，可分下面三步完成

- 1 给定 $\mathbf{u}$ ，计算 $t$ 的条件密度函数 $f_n(t|\mathbf{u})$
- 2 给定 $\mathbf{u}$ ，计算 $r$ 的条件密度函数 $f_n(r|\mathbf{u})$
- 3 计算样本相关系数 $r$ 的密度函数 $f_n(r)$

## $\rho \neq 0$ 时样本相关系数的分布

### ■ 第一步：计算 $t$ 的条件密度函数 $f_n(t|\mathbf{u})$

- 给定 $\|\mathbf{u}\|$ 时，由于 $w_1$ ， $\sum_{k=2}^{n-1} w_k^2$  和 $\|\mathbf{u}\|$ 相互独立，且

$$w_1 \sim N(0, 1) \quad \sum_{k=2}^{n-1} w_k^2 \sim \chi_{n-2}^2$$

- 注意到： $t = \frac{\delta + w_1}{\sqrt{\sum_{k=2}^{n-1} w_k^2 / (n-2)}}$ ，可用卷积公式计算如下。

## $\rho \neq 0$ 时样本相关系数的分布

$$\begin{aligned} f_n(t|\delta, \mathbf{u}) &= \int_0^\infty \sqrt{\frac{w}{n-2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[t\sqrt{w/(n-2)}-\delta]^2} \frac{w^{n/2-2}}{2^{n/2-1}\Gamma(\frac{1}{2}n-1)} e^{-w/2} dw \\ &= \sum_{k=0}^{\infty} \frac{t^k \delta^k e^{-\delta^2/2}}{\sqrt{2\pi} k! (n-2)^{\frac{1}{2}(k+1)} 2^{n/2-1} \Gamma(\frac{1}{2}n-1)} \int_0^\infty w^{\frac{1}{2}(k+n-1)-1} e^{-\frac{1}{2}w(1+\frac{t^2}{n-2})} dw \\ &= \sum_{k=0}^{\infty} \frac{t^k \delta^k e^{-\delta^2/2} 2^{\frac{1}{2}(k+n-1)} \Gamma(\frac{1}{2}(k+n-1))}{\sqrt{2\pi} k! (n-2)^{\frac{1}{2}(k+1)} 2^{n/2-1} \Gamma(\frac{1}{2}n-1)} \left(1 + \frac{t^2}{n-2}\right)^{-\frac{1}{2}(k+n-1)} \\ &= \sum_{k=0}^{\infty} \frac{t^k \delta^k e^{-\delta^2/2} 2^{k/2} \Gamma(\frac{1}{2}(k+n-1))}{\sqrt{\pi} k! (n-2)^{\frac{1}{2}(k+1)} \Gamma(\frac{1}{2}n-1)} \left(1 + \frac{t^2}{n-2}\right)^{-\frac{1}{2}(k+n-1)}. \end{aligned}$$

### ■ 第二步：计算 $r$ 的条件密度函数 $f_n(r|\mathbf{u})$

- 由 $t = \sqrt{n-2}r/\sqrt{1-r^2}$ 知,  $1 + t^2/(n-2) = 1/(1-r^2)$ , 且 $dt/dr = \sqrt{n-2}/\sqrt{(1-r^2)^3}$ 。
- 给定 $\mathbf{u}$ 时,  $r$ 的条件分布密度函数为:

$$\begin{aligned} f_n(r|\mathbf{u}) &= \sum_{k=0}^{\infty} \frac{r^k \delta^k e^{-\delta^2/2} 2^{k/2} \Gamma\left(\frac{1}{2}(k+n-1)\right)}{\sqrt{\pi} k! \Gamma\left(\frac{1}{2}n-1\right)} (1-r^2)^{\frac{n-4}{2}} \\ &= \sum_{k=0}^{\infty} \frac{r^k \phi^k \|\mathbf{u}\|^k e^{-\frac{1}{2}\phi^2 \|\mathbf{u}\|^2} 2^{k/2} \Gamma\left(\frac{1}{2}(k+n-1)\right)}{\sqrt{\pi} k! \Gamma\left(\frac{1}{2}n-1\right)} (1-r^2)^{\frac{n-4}{2}}. \end{aligned}$$

### ■ 第三步：计算 $r$ 的密度函数 $f_n(r)$

- 由于 $\|\mathbf{u}\|^2 \sim \chi_{n-1}^2$ ，并假设 $\|\mathbf{u}\|^2$ 的密度函数为 $f(\|\mathbf{u}\|^2)$ ，则 $r$ 的密度函数为：

$$\begin{aligned} f_n(r|\rho) &= \int_0^\infty f_n(r|\mathbf{u})f(\|\mathbf{u}\|^2)d\|\mathbf{u}\|^2 \\ &= \sum_{k=0}^\infty \frac{r^k \phi^k 2^{k/2} \Gamma(\frac{1}{2}(k+n-1))}{\sqrt{\pi} k! \Gamma(\frac{1}{2}n-1)} (1-r^2)^{\frac{1}{2}(n-4)} E[\|\mathbf{u}\|^k e^{-\frac{1}{2}\phi^2 \|\mathbf{u}\|^2}], \end{aligned}$$

$$\text{其中 } E[\|\mathbf{u}\|^k e^{-\frac{1}{2}\phi^2 \|\mathbf{u}\|^2}] = \frac{2^{k/2} \Gamma(\frac{1}{2}(k+n-1))}{\Gamma(\frac{1}{2}(n-1))} (1+\phi^2)^{-\frac{k+n-1}{2}}.$$

## $\rho \neq 0$ 时样本相关系数的分布

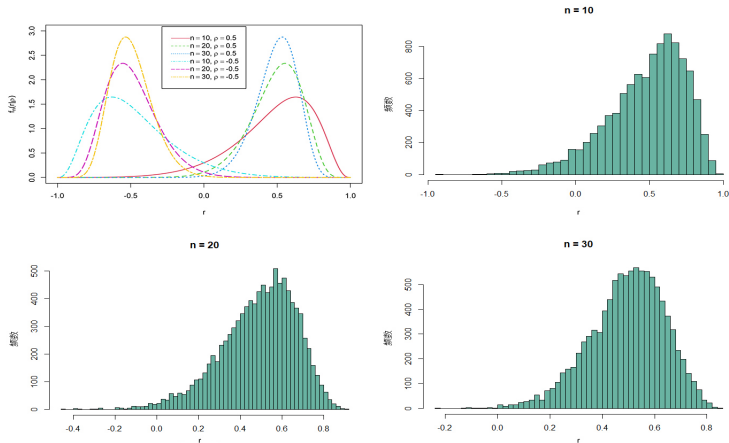
● 因此,  $r$  的密度函数为

$$\begin{aligned} f_n(r|\rho) &= \sum_{k=0}^{\infty} \frac{r^k \phi^k 2^k \Gamma^2\left(\frac{1}{2}(k+n-1)\right)}{\sqrt{\pi} k! \Gamma\left(\frac{1}{2}(n-1)\right) \Gamma\left(\frac{1}{2}n-1\right)} \\ &\quad \times (1-r^2)^{\frac{1}{2}(n-4)} (1+\phi^2)^{-\frac{1}{2}(k+n-1)} \\ &= \frac{(1-r^2)^{\frac{1}{2}(n-4)} (1-\rho^2)^{\frac{1}{2}(n-1)}}{\sqrt{\pi} \Gamma\left(\frac{1}{2}(n-1)\right) \Gamma\left(\frac{1}{2}n-1\right)} \sum_{k=0}^{\infty} \frac{r^k \rho^k 2^k \Gamma^2\left(\frac{1}{2}(k+n-1)\right)}{k!}. \end{aligned} \quad (4)$$

### 定理5.4.2

来自总体相关系数为  $\rho$  的二元正态分布的  $n$  个观测值的样本相关系数  $r$  的密度函数由式(4)所定义。

## $\rho \neq 0$ 时, $r$ 的密度函数曲线



**Figure:** 密度函数曲线和基于10000次重复试验的直方图；当 $n = 10$ 时,  $\bar{r} = 0.4814$ ；当 $n = 20$ 时,  $\bar{r} = 0.4894$ ；当 $n = 30$ 时,  $\bar{r} = 0.5072$ 。

## $\rho \neq 0$ 时样本相关系数的分布

- 样本相关系数 $r$ 的精确分布最早由Fisher (1915)得到, 他给出了密度函数的另一个形式:

$$\frac{(1 - \rho^2)^{\frac{1}{2}(n-1)}(1 - r^2)^{\frac{1}{2}(n-4)}}{(n-3)!\pi} \left| \frac{d^{n-2}}{dx^{n-2}} \left\{ \frac{\arccos(-x)}{\sqrt{1-x^2}} \right\} \right|_{x=r\rho}.$$

- Hotelling (1953)对 $r$ 的密度函数做了一个彻底研究, 并推荐使用如下的密度函数:

$$\frac{(n-2)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n-\frac{1}{2})} (1-\rho^2)^{\frac{1}{2}(n-1)} (1-r^2)^{\frac{1}{2}(n-4)} (1-\rho r)^{n-\frac{3}{2}} F\left(\frac{1}{2}, \frac{1}{2}; n-\frac{3}{2}; \frac{1+\rho r}{2}\right),$$

其中 $F(a, b; c; x) = \sum_{j=0}^{\infty} \frac{\Gamma(a+j)\Gamma(b+j)\Gamma(c)}{\Gamma(a)\Gamma(b)\Gamma(c+j)} \frac{x^j}{j!}$  是一个超几何函数。



## $\rho \neq 0$ 时样本相关系数的分布

- Olkin 和Pratt (1958)证明, 样本相关系数 $r$ 是总体相关系数 $\rho$ 的一个有偏估计, 并给出了 $\rho$ 的唯一一致最小方差无偏估计为

$$G(r) = r \cdot F\left(\frac{1}{2}, \frac{1}{2}; \frac{n-2}{2}; 1-r^2\right).$$

- David (1938) 计算了样本相关系数 $r$ 的分布函数

$$F(r^*|n, \rho) = \Pr(r \leq r^*).$$

- 由密度函数的对偶性知:

$$F(r^*|n, \rho) = 1 - F(-r^*|n, -\rho).$$

- 为了方便实际的应用, David (1938)把相应的数值制成表格。

- 考虑下面的假设检验问题：

$$H_0 : \rho = \rho_0, \quad H_{11} : \rho > \rho_0; \quad (5)$$

$$H_0 : \rho = \rho_0, \quad H_{21} : \rho < \rho_0; \quad (6)$$

$$H_0 : \rho = \rho_0, \quad H_{31} : \rho \neq \rho_0. \quad (7)$$

- 令 $r_0$ 是分布函数 $F(r^*|n, \rho_0)$ 的上 $\alpha$ 分位点，如果 $r > r_0$ ，则拒绝 $H_0$ ，并接受 $H_{11}$ ，即 $\rho > \rho_0$ 成立；
- 令 $r'_0$ 是分布函数 $F(r^*|n, \rho_0)$ 的下 $\alpha$ 分位点，如果 $r < r'_0$ ，则拒绝 $H_0$ ，并接受 $H_{21}$ ，即 $\rho < \rho_0$ 成立；

## 精确分布下 $\rho$ 的假设检验

- 当 $r > r_1$ 和 $r < r'_1$ , 其中 $r_1$ 和 $r'_1$ 满足 $1 - F(r_1|n, \rho_0) + F(r'_1|n, \rho_0) = \alpha$ , 则拒绝 $H_0$ , 并接受 $H_{31}$ , 即 $\rho \neq \rho_0$ 。
- David (1937)建议 $r_1$ 和 $r'_1$ 满足:

$$1 - F(r_1|n, \rho_0) = F(r'_1|n, \rho_0) = \alpha/2.$$

### 例子

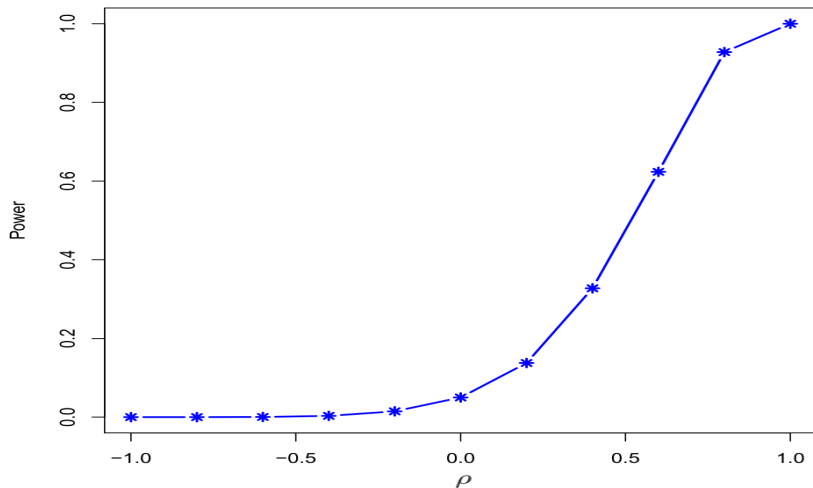
假设计划在5%的显著性水平下, 使用15个观测值来检验 $H_0 : \rho = 0.5 \leftrightarrow H_{11} : \rho \neq 0.5$ 。在David (1938)的表格中, 可以发现 $F(0.027|15, 0.5) = 0.025$ 和 $F(0.805|15, 0.5) = 0.975$ , 因而当 $r < 0.027$ 或者 $r > 0.805$ 时, 就拒绝原假设 $H_0$ 。

- David (1938)的表格可以计算功效函数，若拒绝域是 $r > r_1$ 和 $r < r'_1$ ，检验的功效就是真实总体相关系数的函数，即 $1 - F(r_1|n, \rho_0) + F(r'_1|n, \rho_0)$ ，反映的是当总体相关系数是 $\rho$ 时拒绝原假设 $H_0$ 的概率。
- 考虑检验 $\rho = 0$ 的单边功效函数。在5%的显著性水平下，功效函数见表格和图。

## 精确分布下 $\rho$ 的假设检验

$\rho$	拒绝概率	$\rho$	拒绝概率
-1.0	0.0000	0.2	0.1376
-0.8	0.0000	0.4	0.3275
-0.6	0.0004	0.6	0.6235
-0.4	0.0032	0.8	0.9279
-0.2	0.0147	1.0	1.0000
0.0	0.0500		

# 精确分布下 $\rho$ 的假设检验



- David (1938)的表格可用于计算 $\rho$ 的置信区间, 对给定 $n$ 和显著性水平 $\alpha$ ,  $r'_1 = f_1(\rho)$ 和 $r_1 = f_2(\rho)$  是 $\rho$ 的两个函数, 使得它们满足

$$\Pr\{f_1(\rho) < r < f_2(\rho) | \rho\} = 1 - \alpha. \quad (8)$$

- 如果 $r_1$ 和 $r'_1$ 满足 $1 - F(r_1 | n, \rho) = F(r'_1 | n, \rho) = \frac{\alpha}{2}$ , 则 $f_1(\rho)$  和 $f_2(\rho)$  是 $\rho$ 的单调递增函数。

- 若 $\rho = f_k^{-1}(r)$ 是 $r = f_k(\rho)$ 的反函数, 其中 $k = 1, 2$ , 那么不等式 $f_1(\rho) < r$ 等价于 $\rho < f_1^{-1}(r)$ 且 $r < f_2(\rho)$ 等价于 $f_2^{-1}(r) < \rho$ .
- 则式(8)可以重新写成

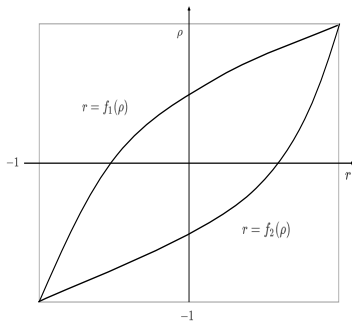
$$\Pr\{f_2^{-1}(r) < \rho < f_1^{-1}(r) | \rho\} = 1 - \alpha. \quad (9)$$

- 对于给定 $n$ 和显著性水平 $\alpha$ , 下图给出了曲线 $r = f_1(\rho)$ 和 $r = f_2(\rho)$ 。



## 精确分布下 $\rho$ 的区间估计

- 在检验 $\rho = \rho_0$ 时, 直线 $\rho = \rho_0$ 和这两条曲线的交点给出了临界点 $r_1$ 和 $r'_1$ 。
- 对于给定的样本相关系数 $\tilde{r}$ ,  $\rho$ 的置信区间就是直线 $r = \tilde{r}$ 夹在这两条曲线之间的部分 $(f_2^{-1}(\tilde{r}), f_1^{-1}(\tilde{r}))$ 。



# 样本相关系数的渐近正态分布

- 定义下面的样本相关系数：

$$r(n) = \frac{v_{ij}(n)}{\sqrt{v_{ii}(n)v_{jj}(n)}}, \quad i \neq j.$$

- 既然样本相关系数具有位置和尺度的不变性，则 $r(n)$ 也可以改写成：

$$r(n) = \frac{c_{ij}(n)}{\sqrt{c_{ii}(n)c_{jj}(n)}}, \quad i \neq j,$$

$$\text{其中 } c_{ij}(n) = \frac{v_{ij}(n)}{\sqrt{\sigma_{ii}(n)\sigma_{jj}(n)}}.$$

## 样本相关系数的渐近正态分布

- 令  $c_{ij}(n), c_{ii}(n), c_{jj}(n)$  的分布与下面矩阵元素的分布相同:

$$\begin{aligned} & \sum_{k=1}^n \begin{pmatrix} Z_{ki}^* \\ Z_{kj}^* \end{pmatrix} \begin{pmatrix} Z_{ki}^* & Z_{kj}^* \end{pmatrix} \\ &= \sum_{k=1}^n \begin{pmatrix} Z_{ki}/\sqrt{\sigma_{ii}} \\ Z_{kj}/\sqrt{\sigma_{jj}} \end{pmatrix} \begin{pmatrix} Z_{ki}/\sqrt{\sigma_{ii}} & Z_{kj}/\sqrt{\sigma_{jj}} \end{pmatrix} \end{aligned}$$

- 其中  $(Z_{ki}^*, Z_{kj}^*)'$  独立且服从二元正态分布, 即

$$N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad \text{其中 } \rho = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}.$$

### 定理5.2.4

假设 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 是来自 $p$ 元正态分布 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的i.i.d.随机样本, 记 $\mathbf{V} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ , 则

$$\mathbf{B}_n = \frac{1}{\sqrt{n}}(\mathbf{V} - n\boldsymbol{\Sigma})$$

收敛于一个正态随机矩阵 $\mathbf{B} = (b_{ij})$ , 其元素均值为0, 协方差为

$$E(b_{ij}b_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}.$$

● 令

$$\mathbf{U}(n) = \frac{1}{n} \begin{pmatrix} c_{ii}(n) \\ c_{jj}(n) \\ c_{ij}(n) \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \\ \rho \end{pmatrix}. \quad (10)$$

● 利用定理5.2.4, 可证明向量 $\mathbf{U}(n)$ 的渐近正态分布为

$$\sqrt{n}(\mathbf{U}(n) - \mathbf{b}) \xrightarrow{d} N_3(\mathbf{0}, \Sigma), \quad n \rightarrow \infty,$$

其中

$$\Sigma = \begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1 + \rho^2 \end{pmatrix}. \quad (11)$$

## 样本相关系数的渐近正态分布

总结，给出下面一般性定理：

### 定理5.4.3

令 $\{U(n)\}$ 是一列 $m$ 维的随机向量， $\mathbf{b}$ 是一个 $m$ 维的固定向量，当 $n \rightarrow \infty$ 时，使得 $\sqrt{n}(U(n) - \mathbf{b}) \xrightarrow{d} N_m(\mathbf{0}, \Sigma)$ ，这儿 $\mathbf{0}$ 是 $m$ 维的零向量， $\Sigma$ 为 $m \times m$ 的协方差矩阵。假设 $f(\mathbf{u})$ 为一个向量值函数，其每个分量 $f_j(\mathbf{u})$ 在 $\mathbf{u} = \mathbf{b}$ 处有非零导数，定义矩阵 $\Phi_b$ ，其 $(i, j)$ 元素为 $\left. \frac{\partial f_j(\mathbf{u})}{\partial u_i} \right|_{\mathbf{u}=\mathbf{b}}$ 。当 $n \rightarrow \infty$ 时，则

$$\sqrt{n}[f(U(n)) - f(\mathbf{b})] \xrightarrow{d} N_m(\mathbf{0}, \Phi_b' \Sigma \Phi_b).$$

## 样本相关系数的渐近正态分布

- 可以验证：分别由式(10)和(11)所定义的 $U(n)$ ， $\mathbf{b}$  和  $\Sigma$  满足定理5.4.3 的条件，这时也定义满足定理5.4.3条件的函数 $f$ ，即

$$r = \frac{u_3}{\sqrt{u_1 u_2}}.$$

- 下面计算矩阵 $\Phi_{\mathbf{b}}$ 的元素，即

$$\left. \frac{\partial r}{\partial u_1} \right|_{\mathbf{u}=\mathbf{b}} = -\frac{1}{2} u_3 u_1^{-3/2} u_2^{-1/2} \Big|_{\mathbf{u}=\mathbf{b}} = -\frac{1}{2} \rho,$$

$$\left. \frac{\partial r}{\partial u_2} \right|_{\mathbf{u}=\mathbf{b}} = -\frac{1}{2} u_3 u_2^{-3/2} u_1^{-1/2} \Big|_{\mathbf{u}=\mathbf{b}} = -\frac{1}{2} \rho,$$

$$\left. \frac{\partial r}{\partial u_3} \right|_{\mathbf{u}=\mathbf{b}} = u_1^{-1/2} u_2^{-1/2} \Big|_{\mathbf{u}=\mathbf{b}} = 1,$$

且 $f(\mathbf{b}) = \rho$ 。

# 样本相关系数的渐近正态分布

- 利用定理5.4.3, 则可证得

$$\sqrt{n}[r(n) - \rho] \xrightarrow{d} N(0, \sigma^2), \quad n \rightarrow \infty, \quad (12)$$

其中

$$\begin{aligned} \sigma^2 &= \left(-\frac{1}{2}\rho, -\frac{1}{2}\rho, 1\right) \begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1 + \rho^2 \end{pmatrix} \begin{pmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{pmatrix} \\ &= (1 - \rho^2)^2. \end{aligned}$$

- 有了样本相关系数 $r(n)$ 的渐近正态分布, 可以构造 $\rho$ 的区间估计。

**问题:** 渐近方差中含有未知参数 $\rho$ , 如何解决这个问题?



## 方法1：插入(plug-in)方法

- 渐近方差  $\sigma^2 = (1 - \rho^2)^2$  包含未知的参数  $\rho$ ，可用它的极大似然估计  $r(n)$  来替换。
- 当  $n \rightarrow \infty$  时，可以证明  $r(n) \xrightarrow{P} \rho$ 。
- 利用Slutsky定理，可得

$$\frac{\sqrt{n}[r(n) - \rho]}{1 - r^2(n)} = \frac{\sqrt{n}[r(n) - \rho]}{1 - \rho^2} \frac{1 - \rho^2}{1 - r^2(n)} \xrightarrow{d} N(0, 1).$$

- $\rho$  的置信水平为  $1 - \alpha$  的置信区间为

$$\left[ r(n) - \frac{1 - r^2(n)}{\sqrt{n}} z_{1-\alpha/2}, \quad r(n) + \frac{1 - r^2(n)}{\sqrt{n}} z_{1-\alpha/2} \right],$$

其中  $z_{1-\alpha/2}$  是标准正态分布的上  $\alpha/2$  分位点。

## 方法2: Fisher Z变换方法

- **解决办法:** 求函数 $f$ , 使得 $f(r(n))$ 的渐近方差为1, 即

$$\sqrt{n}[f(r(n)) - f(\rho)] \xrightarrow{d} N(0, 1).$$

- 根据定理5.4.3, 有

$$\sqrt{n}[f(r(n)) - f(\rho)] \xrightarrow{d} N(0, (f'(\rho))^2(1 - \rho^2)^2).$$

- 解函数 $f$ , 使得 $(f'(\rho))^2(1 - \rho^2)^2 = 1$ 。解得函数为:  $f(x) = \frac{1}{2} \ln \frac{1+x}{1-x}$ 。

- 故有

$$\sqrt{n} \left[ \frac{1}{2} \ln \frac{1+r(n)}{1-r(n)} - \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \right] \xrightarrow{d} N(0, 1).$$

## 方法2: Fisher Z变换方法

- 由上面结果, 可构造  $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$  的置信水平为  $1-\alpha$  的置信区间为:

$$\left[ \frac{1}{2} \ln \frac{1+r(n)}{1-r(n)} - \frac{1}{\sqrt{n}} z_{1-\alpha/2}, \quad \frac{1}{2} \ln \frac{1+r(n)}{1-r(n)} + \frac{1}{\sqrt{n}} z_{1-\alpha/2} \right].$$

- 为构造  $\rho$  的置信区间, 对上面的置信区间进行变换, 可得到  $\rho$  的置信水平为  $1-\alpha$  的置信区间为:

$$\left[ \frac{\frac{1+r(n)}{1-r(n)} \exp\left(-\frac{2}{\sqrt{n}} z_{1-\alpha/2}\right) - 1}{\frac{1+r(n)}{1-r(n)} \exp\left(-\frac{2}{\sqrt{n}} z_{1-\alpha/2}\right) + 1}, \quad \frac{\frac{1+r(n)}{1-r(n)} \exp\left(\frac{2}{\sqrt{n}} z_{1-\alpha/2}\right) - 1}{\frac{1+r(n)}{1-r(n)} \exp\left(\frac{2}{\sqrt{n}} z_{1-\alpha/2}\right) + 1} \right].$$

■ 假设随机向量  $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $p > 2$ , 将  $\mathbf{X}$ ,  $\boldsymbol{\mu}$  和  $\boldsymbol{\Sigma}$  剖分为

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

- ▶  $\mathbf{X}^{(1)}$  和  $\boldsymbol{\mu}^{(1)}$  为  $q \times 1$  的向量
- ▶  $\boldsymbol{\Sigma}_{11}$  为  $q \times q$  矩阵
- ▶  $\mathbf{X}^{(2)}$  和  $\boldsymbol{\mu}^{(2)}$  为  $(p - q) \times 1$  的向量
- ▶  $\boldsymbol{\Sigma}_{22}$  为  $(p - q) \times (p - q)$  矩阵
- ▶  $\boldsymbol{\Sigma}_{12}$  为  $q \times (p - q)$  矩阵

## 样本偏相关系数

- 给定  $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$  时  $\mathbf{X}^{(1)}$  的条件分布服从  $q$  元正态分布, 即  $(\mathbf{X}^{(1)} | \mathbf{X}^{(2)} = \mathbf{x}^{(2)}) \sim N_q(\boldsymbol{\mu}^{(1)} + \boldsymbol{\beta}(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}), \boldsymbol{\Sigma}_{11.2})$ .
- $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$  为  $\mathbf{X}^{(1)}$  对  $\mathbf{X}^{(2)}$  的**回归系数**。
- $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$  为**条件协方差矩阵**, 它的元素用  $\sigma_{ij \cdot q+1, \dots, p}$  表示,  $i, j = 1, \dots, q$ 。
- 给定  $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$  条件下,  $X_i$  与  $X_j$  的**偏相关系数**  $(\text{corr}(\mathbf{X}^{(1)} | \mathbf{X}^{(2)}))$  定义为:

$$\rho_{ij \cdot q+1, \dots, p} = \frac{\sigma_{ij \cdot q+1, \dots, p}}{(\sigma_{ii \cdot q+1, \dots, p} \sigma_{jj \cdot q+1, \dots, p})^{1/2}}, \quad i, j = 1, \dots, q.$$

- 类似地，对样本均值 $\bar{\mathbf{x}}$ ，离差矩阵 $\mathbf{V}$ 和样本协方差矩阵 $\mathbf{S}$ 进行如下剖分：

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{\mathbf{x}}^{(1)} \\ \bar{\mathbf{x}}^{(2)} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix},$$
$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix}.$$

- 由定理5.1.2，在给定 $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ 时，则条件期望 $\mu_{1.2}$ 的MLE为：

$$\begin{aligned} \hat{\mu}_{1.2} &= \bar{\mathbf{x}}^{(1)} + (\mathbf{V}_{12}/n)(\mathbf{V}_{22}/n)^{-1}(\mathbf{x}^{(2)} - \bar{\mathbf{x}}^{(2)}) \\ &= \bar{\mathbf{x}}^{(1)} + \mathbf{V}_{12}\mathbf{V}_{22}^{-1}(\mathbf{x}^{(2)} - \bar{\mathbf{x}}^{(2)}). \end{aligned}$$

## 样本偏相关系数

■ 由  $\mathbf{V}_{12}\mathbf{V}_{22}^{-1} = \mathbf{S}_{12}\mathbf{S}_{22}^{-1}$ , 所以条件期望  $\mu_{1.2}$  的MLE也可表示为:

$$\hat{\mu}_{1.2} = \bar{\mathbf{x}}^{(1)} + \mathbf{S}_{12}\mathbf{S}_{22}^{-1}(\mathbf{x}^{(2)} - \bar{\mathbf{x}}^{(2)}).$$

■ 条件协方差阵  $\Sigma_{11.2}$  的MLE为:

$$\hat{\Sigma}_{11.2} = \frac{\mathbf{V}_{11}}{n} - \frac{\mathbf{V}_{12}}{n} \left( \frac{\mathbf{V}_{22}}{n} \right)^{-1} \frac{\mathbf{V}_{21}}{n} = \frac{\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}}{n} = \frac{\mathbf{V}_{11.2}}{n}.$$

■ 可得偏相关系数的极大似然估计为:

$$\hat{\rho}_{ij \cdot q+1, \dots, p} = \frac{\hat{\sigma}_{ij \cdot q+1, \dots, p}}{(\hat{\sigma}_{ii \cdot q+1, \dots, p} \hat{\sigma}_{jj \cdot q+1, \dots, p})^{1/2}}, \quad i, j = 1, \dots, q,$$

其中  $\hat{\sigma}_{ij \cdot q+1, \dots, p}$  是  $\hat{\Sigma}_{11.2}$  的第  $(i, j)$  元素。

## 定理5.4.4

设 $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , 并且 $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'(i = 1, \dots, n)$ 为来自 $p$ 元正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一组随机样本,  $n > p$ ,  $\bar{\mathbf{x}}$ 为样本均值向量,  $\mathbf{V}$ 为样本离差阵,  $\mathbf{S}$ 为样本协方差阵, 在给定 $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ 时, 则

- ① 条件期望 $\boldsymbol{\mu}_{1.2}$ 和条件协方差阵 $\boldsymbol{\Sigma}_{11.2}$ 的MLE分别为 $\hat{\boldsymbol{\mu}}_{1.2} = \bar{\mathbf{x}}^{(1)} + \mathbf{S}_{12}\mathbf{S}_{22}^{-1}(\mathbf{x}^{(2)} - \bar{\mathbf{x}}^{(2)})$ 和 $\hat{\boldsymbol{\Sigma}}_{11.2} = \mathbf{V}_{11.2}/n$ , 同时 $\boldsymbol{\beta}$ 的MLE为 $\hat{\boldsymbol{\beta}} = \mathbf{V}_{12}\mathbf{V}_{22}^{-1}$ ;
- ② 进一步, 条件期望 $\boldsymbol{\mu}_{1.2}$ 和条件协方差阵 $\boldsymbol{\Sigma}_{11.2}$ 的无偏估计分别为 $\bar{\mathbf{x}}^{(1)} + \mathbf{S}_{12}\mathbf{S}_{22}^{-1}(\mathbf{x}^{(2)} - \bar{\mathbf{x}}^{(2)})$ 和 $\mathbf{S}_{11.2}$ , 其中

$$\mathbf{S}_{11.2} = \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21} = \mathbf{V}_{11.2}/(n-1).$$



## 定理5.4.5

设 $\mathbf{x}_1, \dots, \mathbf{x}_n$ 为来自 $p$ 元正态总体 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的一组随机样本,  $n > p$ , 则给定后 $p - q$ 个分量条件下, 前 $q$ 个分量的偏相关系数 $\rho_{ij \cdot q+1, \dots, p}$ 的MLE为

$$\hat{\rho}_{ij \cdot q+1, \dots, p} = \frac{v_{ij \cdot q+1, \dots, p}}{(v_{ii \cdot q+1, \dots, p} v_{jj \cdot q+1, \dots, p})^{1/2}}, \quad i, j = 1, \dots, q,$$

其中 $(v_{ij \cdot q+1, \dots, p}) = \mathbf{V}_{11} - \mathbf{V}_{12} \mathbf{V}_{22}^{-1} \mathbf{V}_{21} = \mathbf{V}_{11.2}$ 。

■ 为了简单, 把极大似然估计  $\hat{\rho}_{ij \cdot q+1, \dots, p}$  记为  $r_{ij \cdot q+1, \dots, p}$ , 称为**样本偏相关系数**。

■ 由估计的回归系数  $\hat{\beta}$ , 矩阵  $\mathbf{V}_{11.2}$  也可以表示为:

$$\begin{aligned}\mathbf{V}_{11.2} &= \sum_{k=1}^n \left[ \mathbf{x}_k^{(1)} - \bar{\mathbf{x}}^{(1)} - \hat{\beta}(\mathbf{x}_k^{(2)} - \bar{\mathbf{x}}^{(2)}) \right] \left[ \mathbf{x}_k^{(1)} - \bar{\mathbf{x}}^{(1)} - \hat{\beta}(\mathbf{x}_k^{(2)} - \bar{\mathbf{x}}^{(2)}) \right]' \\ &= \mathbf{V}_{11} - \hat{\beta} \mathbf{V}_{22} \hat{\beta}'\end{aligned}$$

● 其中向量  $\hat{\epsilon}_k^{(1.2)} = \mathbf{x}_k^{(1)} - \bar{\mathbf{x}}^{(1)} - \hat{\beta}(\mathbf{x}_k^{(2)} - \bar{\mathbf{x}}^{(2)})$  称为  $\mathbf{x}_k^{(1)}$  对  $\mathbf{x}_k^{(2)}$  的回归残差, 样本偏相关系数就是这些回归残差之间简单的相关系数。



谢谢，请多提宝贵意见！