Multivariate Statistical Analysis

Lecture 10

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

Sample Correlation Coefficient

Given the sample $\mathbf{x}_1, \dots, \mathbf{x}_N$ from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the maximum likelihood estimator of the correlation between the i-th and the j-th components is

$$r_{ij} = \frac{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N}(x_{i\alpha} - \bar{x}_i)^2}\sqrt{\sum_{\alpha=1}^{N}(x_{j\alpha} - \bar{x}_j)^2}},$$

where $x_{i\alpha}$ is the i-th component of $\mathbf{x}_\alpha$ and

$$\bar{x}_i = \frac{1}{N}\sum_{\alpha=1}^{N} x_{i\alpha}.$$

We shall find the distribution of r_{ij}.

If the population correlation

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

is zero, then the density of sample correlation r_{ij} is

$$k_N(r) = \frac{\Gamma\left(\frac{N-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{N-2}{2}\right)}(1 - r_{ij}^2)^{\frac{N-4}{2}}.$$

Sample Correlation Coefficient

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be observation from $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

We denote

$$\mathbf{x}_\alpha = \begin{bmatrix} x_{1\alpha} \\ x_{2\alpha} \end{bmatrix}, \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N} \mathbf{x}_\alpha \quad \text{and} \quad \mathbf{A} = \sum_{\alpha=1}^{N} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^\top.$$

We have shown that $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \sum_{\alpha=1}^{n} \mathbf{z}_\alpha \mathbf{z}_\alpha^\top,$$

where $n = N - 1$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent distributed to $\mathcal{N}_2\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$

Sample Correlation Coefficient

We denote

$$a_{11.2} = a_{11} - \frac{a_{12}^2}{a_{22}} \qquad \text{and} \qquad \sigma_{11.2} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}.$$

Lemma

Based on above notations, we have

(a) $\dfrac{a_{11}}{\sigma_{11}} \sim \chi_n^2$ *and* $\dfrac{a_{22}}{\sigma_{22}} \sim \chi_n^2$;

(b) $a_{12} \mid a_{22} \sim \mathcal{N}\left(\sigma_{22}^{-1}\sigma_{12}a_{22}, \sigma_{11.2}a_{22}\right)$;

(c) $\dfrac{a_{11.2}}{\sigma_{11.2}} \sim \chi_{n-1}^2$ *is independent on* a_{12} *and* a_{22}.

We can show that

$$z = \frac{x}{\sqrt{y/(n-1)}} \sim t_{m-1},$$

where

$$x = \frac{a_{12} - \sigma_{22}^{-1}\sigma_{12}a_{22}}{\sqrt{\sigma_{11.2}a_{22}}} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad y = \frac{a_{11.2}}{\sigma_{11.2}} \sim \chi^2_{n-1}$$

are independent.

Sample Correlation Coefficient

If population correlation

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

is non-zero, the density of sample correlation r_{ij} is

$$\frac{2^{n-2}(1-\rho^2)^{\frac{n}{2}}(1-r_{ij}^2)^{\frac{n-3}{2}}}{(n-2)!\pi} \sum_{\alpha=0}^{\infty} \frac{(2\rho r_{ij})^{\alpha}}{\alpha!} \left(\Gamma\left(\frac{n+\alpha}{2}\right)\right)^2.$$

Sample Correlation Coefficient

If population correlation

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

is non-zero, the density of sample correlation r_{ij} is

$$\frac{2^{n-2}(1-\rho^2)^{\frac{n}{2}}(1-r_{ij}^2)^{\frac{n-3}{2}}}{(n-2)!\pi} \sum_{\alpha=0}^{\infty} \frac{(2\rho r_{ij})^{\alpha}}{\alpha!} \left(\Gamma\left(\frac{n+\alpha}{2}\right)\right)^2.$$

Outline

Tests for the Hypothesis of Lack of Correlation

Consider the hypothesis $H : \rho_{ij} = 0$ for some particular pair (i, j).

1. For testing H against alternatives $\rho_{ij} > 0$, we reject H if $r_{ij} > r_0$ for some positive r_0. The probability of rejecting H when H is true is

$$\int_{r_0}^{1} k_N(r)\,\mathrm{d}r.$$

2. For testing H against alternatives $\rho_{ij} < 0$, we reject H if $r_{ij} < -r_0$.
3. For testing H against alternatives $\rho_{ij} \neq 0$, we reject H if $r_{ij} > r_1$ or $r_{ij} < -r_1$ for some positive r_1. The probability of rejection when H is true is

$$\int_{-1}^{-r_1} k_N(r)\,\mathrm{d}r + \int_{r_1}^{1} k_N(r)\,\mathrm{d}r.$$

Tests for Lack of Correlation

We have shown that

$$\sqrt{N-2} \cdot \frac{r_{ij}}{\sqrt{1-r_{ij}^2}}$$

has the t-distribution with $N-2$ degrees of freedom.

We can also use t-tables. For $\rho_{ij} \neq 0$, reject H if

$$\sqrt{N-2} \cdot \frac{|r_{ij}|}{\sqrt{1-r_{ij}^2}} > t_{N-2}(\alpha),$$

where $t_{N-2}(\alpha)$ is the two-tailed significance point of the t-statistic with $N-2$ degrees of freedom for significance level α.

Outline

The Likelihood Ratio Criterion

The likelihood ratio criterion:

1. Let $L(\mathbf{x}, \boldsymbol{\theta})$ be the likelihood function of the observation $\mathbf{x}$ and the parameter vector $\boldsymbol{\theta} \in \Omega$.
2. Let a null hypothesis be defined by a proper subset ω of Ω. The likelihood ratio criterion is

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\theta} \in \omega} L(\mathbf{x}, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Omega} L(\mathbf{x}, \boldsymbol{\theta})}.$$

3. The likelihood ratio test is the procedure of rejecting the null hypothesis when $\lambda(\mathbf{x})$ is less than a predetermined constant.

Test $\rho = \rho_0$ by the Likelihood Ratio Criterion

The likelihood ratio criterion is

$$\frac{\sup_\omega L(\mathbf{x}, \boldsymbol{\theta})}{\sup_\Omega L(\mathbf{x}, \boldsymbol{\theta})} = \left(\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2}\right)^{\frac{N}{2}}.$$

The likelihood ratio test is

$$\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2} \leq c$$

where c is chosen by the prescribed significance level.

Test $\rho = \rho_0$ by the Likelihood Ratio Criterion

The critical region can be written equivalently as

$$(\rho_0^2 c - \rho_0^2 + 1)r^2 - 2\rho_0 cr + c - 1 + \rho_0^2 \geq 0,$$

that is,

$$r > \frac{\rho_0 c + (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c - \rho_0^2 + 1} \quad \text{and} \quad r < \frac{\rho_0 c - (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c - \rho_0^2 + 1}.$$

Thus the likelihood ratio test of $H : \rho = \rho_0$ against alternatives $\rho \neq \rho_0$ has a rejection region of the form $r > r_1$ and $r < r_2$.