Multivariate Statistical Analysis

Lecture 10

Fudan University

luoluo@fudan.edu.cn

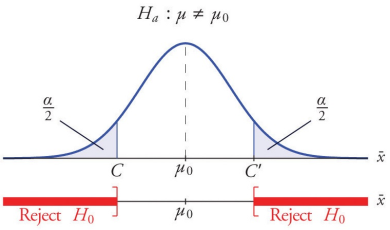
Outline

Outline

Hypothesis Testing for the Mean (Covariance is Known)

In the univariate case, the difference between the sample mean and the population mean is normally distributed. We consider

$$z = \frac{\sqrt{N}}{\sigma}(\bar{x} - \mu_0).$$



What about multivariate case?

Hypothesis Testing for the Mean (Covariance is Known)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ constitute a sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

What about multivariate case to test $\boldsymbol{\mu} = \boldsymbol{\mu}_0$?

$$\frac{\sqrt{N}}{\sigma}(\bar{x} - \mu_0) \implies \frac{N}{\sigma^2}(\bar{x} - \mu_0)^2 \implies N(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0).$$

Rejection Region

Let $\chi_p^2(\alpha)$ be the number such that

$$\Pr\left\{N(\bar{\mathbf{x}}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) > \chi_p^2(\alpha)\right\} = \alpha.$$

To test the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is a specified vector, we use as our rejection region (critical region)

$$N(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0) > \chi_p^2(\alpha).$$

If above inequality is satisfied, we reject the null hypothesis.

Confidence Region

Consider the statement made on the basis of a sample with mean $\bar{\mathbf{x}}$:
"The mean of the distribution satisfies

$$N(\bar{\mathbf{x}} - \boldsymbol{\mu}^*)^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}^*) \leq \chi_p^2(\alpha).$$

as an inequality on $\boldsymbol{\mu}^*$." This statement is true with probability $1 - \alpha$.

Thus, the set of $\boldsymbol{\mu}^*$ satisfying above inequality is a confidence region for $\boldsymbol{\mu}$ with confidence $1 - \alpha$.

Two-Sample Problems

Suppose there are two samples:

1. $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{N_1}^{(1)}$ from $\mathcal{N}(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma})$;
2. $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{N_2}^{(2)}$ from $\mathcal{N}(\boldsymbol{\mu}^{(2)}, \boldsymbol{\Sigma})$;

where $\boldsymbol{\Sigma}$ is known.

How to test the hypothesis $\boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)}$?

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}.

Sample Correlation Coefficient

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

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

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 $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$

Sample Correlation Coefficient

We denote

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

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

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

Sample Correlation Coefficient

We can show that

$$
\begin{aligned}
z =& \frac{x}{\sqrt{y/(n-1)}} \\
=& \frac{\sqrt{n-1}\left(r - \sigma_{12}\sigma_{22}^{-1}\sqrt{a_{22}/a_{11}}\right)}{\sqrt{1-r^2}}
\end{aligned}
$$

where

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

are independent.

If $\sigma_{12} = 0$, then $z = \dfrac{x}{\sqrt{y/(n-1)}} \sim t_{n-1}$.

Sample Correlation Coefficient

If population correlation

$$\rho = \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$$

is non-zero ($\sigma_{12} \neq 0$), the density of sample correlation r is

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