Multivariate Statistical Analysis

Lecture 09

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

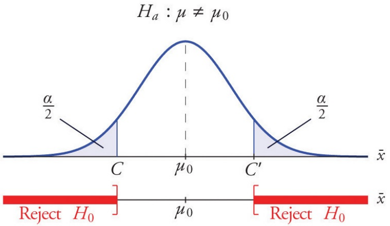
1. Noncentral Chi-Squared Distribution
2. Hypothesis Testing for the Mean (Covariance is Known)
3. Sample Correlation Coefficient

Hypothesis Testing for the Mean

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

We consider

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



1. For significance level $\alpha = 0.05$ and $p = 1$, we have $1 - \alpha = 0.95$.
2. What about multivariate case?

Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent, standard normal random variables, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the (central) chi-squared distribution (χ^2-distribution) with n degrees of freedom. One may write $y \sim \chi_n^2$.

We have $\mathbb{E}[y] = n$ and $\mathrm{Var}[y] = 2n$.

The probability density function of the (central) chi-squared distribution is

$$f(y;\, n) = \begin{cases} \dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} \exp\left(-\dfrac{y}{2}\right), & y > 0; \\ 0, & \text{otherwise}, \end{cases}$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t)\,\mathrm{d}t.$$

Chi-Squared Distribution

The derivation for the density of Chi-square distribution:

1. Show that $\Gamma(1/2) = \sqrt{\pi}$.
2. For $y_1 = x^2$ with $x \sim \mathcal{N}(0, 1)$, the density function of y_1 is
$$\frac{1}{\sqrt{2\pi y_1}} \exp\left(-\frac{1}{2} y_1\right).$$
3. For beta function $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$, we have
$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$
4. Show the density of $y_n = \sum_{i=1}^{n} x_i^2$ by induction.

Noncentral Chi-Squared Distribution

If $x_1, \ldots, x_n$ are independent and each x_i are normally distributed random variables with means μ_i and unit variances, then the sum of their squares,

$$y = \sum_{i=1}^{n} x_i^2,$$

is distributed according to the noncentral Chi-squared distribution with n degrees of freedom and noncentrality parameter

$$\lambda = \sum_{i=1}^{n} \mu_i^2.$$

One may write $y \sim \chi^2_{n,\lambda}$.

We have $\mathbb{E}[y] = n + \lambda$ and $\mathrm{Var}[y] = 2n + 4\lambda$.

Noncentral Chi-Squared Distribution

Theorem

If $y_1, \ldots, y_k$ *are independent and each* y_i *is distributed according to the noncentral* χ^2*-distribution with* n_i *degrees of freedom and noncentrality parameter* λ_i*, then*

$$\sum_{i=1}^{k} y_i \sim \chi^2_{n,\lambda},$$

where

$$n = \sum_{i=1}^{k} n_i \qquad \text{and} \qquad \lambda = \sum_{i=1}^{k} \lambda_i.$$

Noncentral Chi-Squared Distribution

Theorem

If the n-component random vector $\mathbf{y}$ *is distributed according to* $\mathcal{N}_n(\boldsymbol{\nu}, \mathbf{T})$ *with* $\mathbf{T} \succ \mathbf{0}$*, then*

$$\mathbf{y}^\top \mathbf{T}^{-1} \mathbf{y} \sim \chi^2_{n,\lambda},$$

where

$$\lambda = \boldsymbol{\nu}^\top \mathbf{T}^{-1} \boldsymbol{\nu}.$$

If $\boldsymbol{\nu} = \mathbf{0}$*, the distribution is the central* χ^2*-distribution.*

Noncentral Chi-Squared Distribution

Let $\mathbf{y} \sim \mathcal{N}_p(\boldsymbol{\lambda}, \mathbf{I})$, then

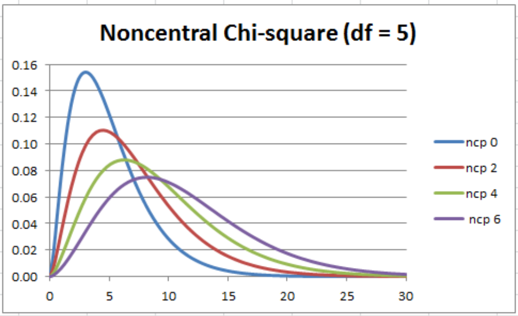
$$v = \mathbf{y}^\top \mathbf{y}$$

is distributed according to the noncentral χ^2-distribution with p degrees of freedom and noncentral parameter $\lambda = \boldsymbol{\lambda}^\top \boldsymbol{\lambda}$.

The probability density function is

$$\begin{aligned} & f(v; p, \lambda) \\ = & \begin{cases} \displaystyle\sum_{\beta=0}^{\infty} \frac{(\lambda/2)^\beta \exp\left(-(\lambda/2)\right)}{\beta!} \cdot \frac{1}{2^{\frac{p+2\beta}{2}} \Gamma\left(\frac{p}{2} + \beta\right)} y^{\frac{p}{2}+\beta-1} \exp\left(-\frac{v}{2}\right) & v > 0, \\ 0, & v \leq 0. \end{cases} \end{aligned}$$

Noncentral Chi-Squared Distribution

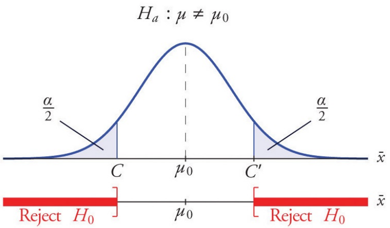


Outline

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

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