Believe it or not, the semester begins.

1.22 △ Texts: Applied Multivariate Statistical Analysis R Johnson

Week1-2 △ Test: 2.\2b. Test 1

HUDM 6122 6.1 Test 2

Multivariate △ Final project: 2-3, suitable dataset for analysis,

☆ Dataset

Chapter 1 Introduction to Multivariate Statistics

Intro.

• Def: Variables (measurements) 1 2 3 ...

Cases

1
2
3
6
5
7

• Types: - Analysis of interdependence → no X or Y

- Principal component analysis
- Factor analysis
- Multidimensional Scaling
- Cluster analysis

- dependence

Multivariate

Structural equation models

Canonical Correlation

Multivariate analysis of variance MANOVA

→ Covariance MANCOVA

Discriminatory / Classification analysis

Example.

- Data layout.
  - n experimental, p variates / characteristics observed/measured
  - $X \rightarrow$ n rows, p columns
  - x_{jk} k^{th} variable on the j^{th} unit

HW.

- mean vector in $R. \rightarrow$ Descriptive Statistics
  - Variances: $s_{kk} = s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk}-\bar{x}_k)^2 = \frac{w_{kk}}{n}$, $k=1,\dots,p$
    - n, not n-1
    - ↓ biased estimator ↓ not biased
  - covariance: $s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k) = \frac{w_{ik}}{n}$, $i=1,\dots,p$, $k=1,\dots,p$
    - n, not n-1

- W: Wilcoxon
- S: Sum of squares

Week 2-2

Ex 1

$$b'd = [2, -1, 4, 0]\begin{bmatrix} -1 \\ 3 \\ -2 \\ 1 \end{bmatrix}$$

$$= 2(-1) + (-1)\cdot 3 + 4\cdot(-2) + 0$$

$$= -13$$

$$b'b = 2^2 + (-1)^2 + 4^2 + 0^2 = 21$$

$$d'd = (-1)^2 + 3^2 + (-2)^2 + 1^2 = 15$$

$$(-13)^2 \le 21 \times 15$$

$$|b'd| < 315.$$

Ex 2

a) Graph: $n=3$ $p=2$. 3×2 data matrix

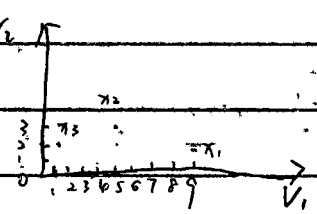
1.

$x_1 = [9, 1]$ $\quad y_1 = \begin{bmatrix} 9 \\ 5 \\ 1 \end{bmatrix}$ $\quad y_2 = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$ Step 1: 50 n. p.

$x_2 = [5, 3]$ Step 2: Ex raw (people)

$x_3 = [1, 2]$

Scatter plot.



2.

$\bar{x}$ ① vector $\quad \bar{x}_1 = \frac{9+5+1}{3} = 5$

② 2 element $\quad \bar{x}_2 = \frac{1+3+2}{3} = 2$

$\bar{x} = [5, 2]$

↓ somewhere in the middle. 在图中

b). deviation vectors

subtract average (by element) from each colum.

$$d_1 = \begin{bmatrix} 9 \\ 5 \\ 1 \end{bmatrix} - \begin{bmatrix} 5 \\ 5 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ -4 \end{bmatrix}$$ How far is each observation from its mean.

Vectors in 3 dimensions.

$$d_2 = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$$

Always add up to 0

Correlation, variance - covariance

length

a) $L_{d_1} = \sqrt{d_1' d_1} = \sqrt{4^2 + 0^2 + (-4)^2} = \sqrt{32}$

$L_{d_2} = \sqrt{d_2' d_2} = \sqrt{(-1)^2 + 1^2 + 0^2} = \sqrt{2}$

$d_1' d_2 = \begin{bmatrix} 4 & 0 & -4 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} = -4 + 0 + 0 = -4$

inner product $\begin{cases} + \\ - \\ 0 \end{cases}$ Cov $\begin{cases} + \\ - \\ 0 \end{cases}$

$$\cos\theta = \frac{d_1' d_2}{\sqrt{d_1' d_1} \cdot \sqrt{d_2' d_2}} = \frac{d_1' d_2}{L_{d_1} L_{d_2}} = \frac{-4}{\sqrt{32}\sqrt{2}} = -\frac{1}{2} \quad \theta = 120^\circ$$

$\theta = 120^\circ$. (θ between vector)

or between 1 , colum $(-\frac{1}{2})$

$\frac{1}{2} \times 1 \times \sin(\theta)$

Variance → S

$S_{11} = \frac{L_{d_1}^2}{n} = \frac{32}{3}$ Standard deviation $= \sqrt{S} = \sqrt{\text{Var}}$

$S_{22} = \frac{2}{3}$ $\frac{\text{length}}{n \text{ (or } n-1)}$ = variance = S

$r_{12} = -\frac{1}{2}$ as of angel = correlation

HT-1

CUM { Uni: / Multi:

Hypo { Σ known { Uni / Multi ; Σ unknown

Estimation and Hypothesis Testing

HuDu 6122

2.5

Week 3-2

- Univariate case

Suppose $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$

Then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$ is MLE of μ

$\hat{\sigma}^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ is MLE of σ^2

Then $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, where $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$

- Multivariate:

$x_1, \cdots x_n \sim N_p(\mu, \Sigma)$ Σ is given

Then $\bar{x} = \hat{\mu}$ is p×1 vector

$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x})$ is p×p matrix

$\hat{\mu}$ is MLE of μ

$\hat{\Sigma}$ is MLE of Σ

Σ covariance matrix

$\hat{\Sigma} = \frac{n-1}{n} S$ 取消 bias $S_n \neq S$ $\hat{\Sigma} = S_n$???

$\hat{\mu} \sim N_p(\mu, \frac{\Sigma}{n})$

- no chi χ^2 in multi for $\hat{\Sigma}$, instead W_p.

复习 cheat sheet

$(n-1)S \sim W_p(n-1, \Sigma)$

is Wishart distribution with n-1 df and dimension p×p, and scale Σ

$\bar{x} \to$

$S \to$ matrix

- Hypotheses when Σ is known → Z test

Recall the ~~univariate~~ case (Two-tail)

$x_1, \cdots x_n \sim N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$

Test statistic:

$$Z_{test} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Z-test ground of no reject Z-test & reject

RR: We reject H_0 if $|Z_{test}| > Z_{\frac{\alpha}{2}}$ (2-side)

When α is the given significant level (usually 0.05)

Slightly change the rule for multi

RR: rejection rule

Note: RR can be written as $Z^2_{test} > Z^2_{\frac{\alpha}{2}} = \chi^2_{1,\alpha}$

↳ Multivariate $Z'Z$

(inner product of it self)

P-value would be same

Multivariate:

Generic case

Σ not given

Multivariate case: Assume $X_1, \dots, X_n \sim N_p(\mu, \Sigma)$

p-variate

This means $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$

$\bar{X} \sim (\mu, \frac{\Sigma}{n})$

Parameters of Vectors

Then: $H_0: \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{10} \\ \vdots \\ \mu_{p0} \end{pmatrix} = \vec{\mu_0}$

$H_1: \mu \neq \mu_0$

Why not 分开算

Type I error

multiple comparison

Note: Why don't we do p univariate tests

$H_0: \mu_i = \mu_{i0}, i = 1, \dots, p$

A: The type I error will become inflated

Ex: test $p=3$

Each test $P(\text{Type I error}) = 0.05$

H_1: ① Avg

② χ^2 test

③ Observed value of

④ p-value

⑤ RR

2

P(At least one Type I error) = Overall type I error

= 1 - P(no error) = 1 - 0.95³ = 0.14

Bonferroni Correction ???

independent x_i

$\Sigma = \sigma^2$ [level]

Multivariate hypothesis:

$H_0: \mu = \mu_0$ vs $H_a: \mu \neq \mu_0$ (μ is $p \times 1$)

T.S.

$Z^2_{test} = n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0)$

scalar $1 \times p$ $p \times p$ $p \times 1$

= scalar (number)

follow χ^2 with p df.

$\Rightarrow Z^2_{test} \sim \chi^2_p$

x' transpose

$x' = t(x)$

Hypotheses

distance: $\Sigma \rightarrow$ direction

distance 제곱

Σ의 역행렬.

RR: Reject H_0 if $Z^2_{test} > \chi^2_{p(\alpha)}$

pchisq(α)

Ex: $p = 2$, $n = 20$

x_1 = Height

x_2 = Weight

Assume $\Sigma = \begin{pmatrix} 20 & 100 \\ 100 & 1000 \end{pmatrix}$ (Assume G)

Height → Correlated with cov

Variance of Height

Variance of weight

Covariance σ_{12}

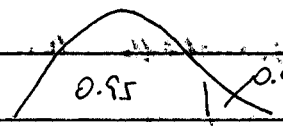
Variance is in ^{2} units

Test = $H_0: \mu = \begin{pmatrix} 70 \\ 170 \end{pmatrix}$ vs $H_1: \mu \neq \begin{pmatrix} 70 \\ 170 \end{pmatrix}$

R → correlation matrix.

Given: $\bar{x} = \begin{pmatrix} 71.45 \\ 164.7 \end{pmatrix}$

(invert Σ { for R)



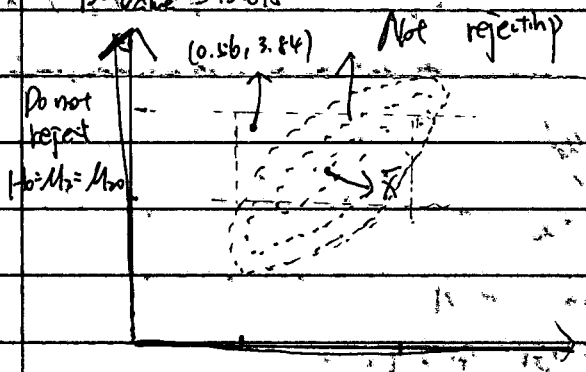
T.S. = B_{test}^2 = 54.

χ^2_{crit} = 9.99

Conclusion: 54 > 9.99

⇒ reject H_0

β - value = 0.015



Not rejecting Multivariate H_0

rectangular → do not reject. both univariate

rectangular ~~do~~ multi reject, uni - reject

But multi - reject, uni - reject

Do not reject univariate

$H_0: \mu_1 = \mu_{10}$

$Ax = c$ solve = "solve(A, c) gives x in R^n"

solve(A) → inverse of A.

What if σ^2 is unknown

A = use ttest

Univariate case:

$x_1, \ldots, x_n \sim N(\mu, \sigma^2)$

$H_0: \mu = \mu_0$ vs $H_1: \mu = \mu_0$ (default in R)

Sample Variance -

Covariance -

Variance - Covariance -

$$t_{test} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

RR: Reject H_0 if $|test| > t_{n-1}(\frac{\alpha}{2})$

$H_1: \mu < \mu_0$

- Multivariate → no one tail test.

HW-2

Σ known → χ^2
Σ unknown → small sample → T^2 ; $\hat{\Sigma} = \frac{n-1}{n} S$
large sample → ?

HW/M 6122
2.10
Week 4 I

Tests for μ when Σ is unknown

- Hotelling's T^2 test

Let $x_1, \dots, x_n \sim N_p(\mu, \Sigma)$

$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$

where μ is $p \times 1$

Σ is $p \times p$ (unknown)

T.S.

$$T^2_{test} = n \underset{1\times p}{(\bar{x} - \mu_0)'} \underset{p\times p}{S^{-1}} \underset{p\times 1}{(\bar{x} - \mu_0)}$$

Recall the univariate case:

$$t^2_{test} = \frac{(\bar{x} - \mu_0)^2}{S/n} = n(\bar{x} - \mu_0) S^{-1} (\bar{x} - \mu_0)$$

near 0 → OK

big → reject null

Under H_0, $T^2_{test} \sim T^2_{p, n-1}$ → similar F, rescale χ^2. Hotelling

R.R: reject $T^2_{test} > T^2_{p, n-1, \alpha}$ ← degree of freedom

Where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ is $p \times 1$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$$

is $p \times p$

tables in optional book

Properties:

1) We must have $n - 1 > p$, otherwise S is singular and S^{-1} doesn't exist.
more features than observations

2) In multivariate case, we don't consider one-tail alternatives like $H_1: \mu < \mu_0$

3) T^2 can be converted to F distribution

$$T^2_{p, n-1} = \frac{(n-1)p}{n-p} \cdot F_{p, n-p}$$

4) When $p = 1$, T^2 is the case of t^2_{n-1} distribution

5) When $n \to \infty$, $T^2_{p, n-1} \to \chi^2_p$ (univariate)

Ex: Table 5.1 on p.215

$p = 3$ vars, $n = 20$ ppl

x_1 = sweat rate

x_2 = sodium

x_3 = potassium

$H_0: \mu = \begin{pmatrix} 4 \\ 50 \\ 10 \end{pmatrix}$ vs $H_1: \mu \neq \begin{pmatrix} 4 \\ 50 \\ 10 \end{pmatrix}$

$\alpha = 0.1$ (n't. value = $\frac{(n-1)p}{n-p} F_{p, n-p}(\alpha)$

??? Why no $\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}$ < matrix ?

☆☆☆ F-table, page 1 α.

Quiz: ☆ Critical (F-table). draw

$T_{test} = 9.74$.

From the table:

Crit. value $= \frac{(n-1)p}{n-p} F(0.1) = \frac{19(3)}{17} F_{3,13}(0.1) \Rightarrow 2.1$

table $= 2.44$

☆ F related to t, cross-multiply ⇒ p value

p-value $= 0.065 < \alpha$

reject H_0

~~Likelihood ratio test~~ ☆☆☆ 不考. ??? 好奇

Idea:

T.S. $\lambda = \frac{\max_{H_0} L(\mu, \Sigma)}{\max L(\mu, \Sigma)}$ Algebra + Calculus

RR: If $\lambda <$ crit. value, then reject H_0

MANOVA [R] → Hotelling

☆☆☆ $\hat{\Sigma}_0$??? 好像 μ_0.

It can be shown that $\lambda = \left(\frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)} \right)^{n/2}$

where $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$

$\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)(x_i - \mu_0)'$

Fact: $\lambda^{\frac{2}{n}} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$

① 2 sample → Paired.

② Which variable? MANOVA.

+ Comparing two means from independent sample

Male vs Female: Same score or not?

~~In the univariate case:~~

Comparing 2 means

$X_1, \dots X_{1n_1} \sim N(\mu_1, \sigma_1^2)$

① H_0

② T.S. t-test Pooled

③ Crit. value df $= n_1 + n_2 - 2$

④ RR

[#] subgroup # obse

and $X_{21}, \dots X_{2n_2} \sim N(\mu_2, \sigma_2^2)$

Normal?

Normal?

Assumptions:

- The two samples are independent
- Equal variances: $\sigma_1^2 = \sigma_2^2 = \sigma^2$
- Normal Populations

Pooled: $S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$ Standard deviation pooled

Test: $H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

Test: $\frac{\bar{x}_1 - \bar{x}_2}{S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$

RR: If $|t_{test}| > t_{n_1+n_2-2}\left(\frac{\alpha}{2}\right)$, then reject H_0

In multivariate case

$x_{11}, \cdots, x_{1n_1} \sim N_p(\mu_1, \Sigma_1)$ ☆ $\mu_1 = \mu_2$ same dimensions

$x_{21}, \cdots, x_{2n_2} \sim N_p(\mu_2, \Sigma_2)$

Assumptions:

- The two samples are independent
- Equal cov matrices: $\Sigma_1 = \Sigma_2 = \Sigma$
- Normal populations

$H_0: \mu_1 = \mu_2$ vs $H_1: \mu_1 \neq \mu_2$

S_{pooled} — 个整体 相当 S.

μ_1 & μ_2 are $p \times 1$ vectors.

Define $S_{pooled} = \frac{1}{n_1+n_2-2}\left((n_1-1)S_1 + (n_2-1)S_2\right)$

T.S. $T^2_{test} = \frac{n_1 n_2}{n_1+n_2}(\bar{x}_1 - \bar{x}_2)' S^{-1}_{pooled}(\bar{x}_1 - \bar{x}_2)$ ☆☆☆

$\frac{1}{n_1} + \frac{1}{n_2} = \frac{n_1+n_2}{n_1 n_2} = \frac{1}{\frac{n_1 n_2}{n_1+n_2}}$

Under H_0, $T^2_{test} \sim T^2_{p,\, n_1+n_2-2}$

It can be shown that

$T^2_{test} \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} \cdot F_{p,\, n_1+n_2-p-1}$ ☆☆☆

RR: Reject H_0 when

$T^2_{test} >$ crit value

Hw:

$(A'A)' = A'(A')' = A'A$

$M' = M \Rightarrow$ symmetric

Rule: $(AB)' = B'A'$

$(A')' = A$.

$A_{n \times p}$.

2 group's diff { cov. / multi.

Assumptions

HT-3

HUDM 6122

2.12

Week 4-2

two

$W = \frac{x_1 + x_2 + x_3}{3}$

$= \frac{1}{3}x_1 + \frac{1}{3}x_2 + \frac{1}{3}x_3$

$= a'x$

$a = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

$E(a'X) =$

$Var(a'X) =$

CI { for 2 groups' difference { uni / multi

Paired 2 groups { uni / multi

☆☆☆

Group = μ_0

Hypotheses - Part 3 — Confidence Interval

- Simultaneous CI's — Bonferroni Adjustment

Given: $x_{11}, \dots x_{1n_1} \sim N_p(\mu_1, \Sigma)$

$x_{21}, \dots x_{2n_2} \sim N_p(\mu_2, \Sigma)$ — Multidimentional Confidence Region

Goal: CI for $\mu_1 - \mu_2$, which is $p \times 1$

Recall: Univariate case

Variance of Covariance

$H_0: \mu_1 - \mu_2 = \mu_0$ vs $H_1: \mu_1 - \mu_2 \neq \mu_0$

$t_{test} = \frac{\bar{x}_1 - \bar{x}_2 - \mu_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

☆☆☆

Table value

Why $\mu_1 \neq \mu_2$?

$\Rightarrow$ CI for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ — $\alpha \to t \Rightarrow$ CI

table

Multivariate case

$T^2_{test} = \frac{n_1 n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2 - \mu_0)' S_p^{-1} (\bar{x}_1 - \bar{x}_2 - \mu_0)$

$T^2_{test} \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$

Let $c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F^{(\alpha)}_{p, n_1 + n_2 - p - 1}$ — Linear Transformation

Scale · F.

Where α is given (0.05, 0.01 ...)

Th: For any $p \times 1$ vector a, — ① c_0 valid for any a

$100(1-\alpha)\%$ CI for $a'(\mu_1 - \mu_2)$ 在 [illegible] - plug in a.

is $a'(\bar{x}_1 - \bar{x}_2) \pm C\sqrt{a' S_p a (\frac{1}{n_1} + \frac{1}{n_2})}$

② pooled variance.

In particular we can choose

$a = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow i^{th}$ row $\rightarrow p \times 1$

then $a'(\mu_1 - \mu_2) = \mu_{1i} - \mu_{2i}$

and the CI for $\mu_{1i} - \mu_{2i}$ is

$\left[\bar{x}_{1i} - \bar{x}_{2i} \pm \right] \sqrt{c_{ii}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

x_{ji}: j ↓ group, i → variables.

Whether [variable] has a significant difference

$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_2 + \mu_3}{2}$

$a \rightarrow$ linear transformation

test of variable

$\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow$ 2nd variable $\rightarrow \bar{\mu}_{11} - \bar{\mu}_{21}$

$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \Rightarrow$ 1st variable $\rightarrow \bar{\mu}_{12} - \bar{\mu}_{22}$

Remark:

Q: When testing

$H_0: \mu_1 - \mu_2 = 0$ vs $H_a: \mu_1 - \mu_2 \neq 0$

and we reject H_0, how do we know which variable contributed the most to rejecting H_0.

A: Calculate

$\hat{a} = S_p^{-1}(\bar{x}_1 - \bar{x}_2)$

Then the largest in abs value element of $\hat{a}$ is corresponding to the var. contributing the most to rejecting the H_0.

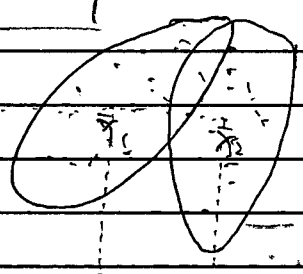
Covariance Matrix

$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$

$\frac{C_1(x_i - \bar{x})^2 + C_2(x_i - \bar{x})^2}{n_1 + n_2 - 2}$

Which variable contribute most to reject H_0.

$\bar{x}_{12} - \bar{x}_{22}$



$\bar{x}_{11}$ $\bar{x}_{21}$

$\star\star\star$: $\mu_1 - \mu_2 \neq 0$ reject

Difference

$\begin{cases} \text{Indep 应 用 } \mu_1-\mu_2 \\ \text{dep 应 } x_1-x_2 \text{ 再 } \bar{d} \end{cases}$

Independent
↓
~~Paired~~ observation test — paired obs

~~Univariate~~ case:

H te / reject Null / follow up

Suppose we have two sample: Sample size the same

x_{11} x_{21} $\to d_1 = x_{11} - x_{21}$

⋮ ⋮

x_{1n} x_{2n} $\to d_n = x_{1n} - x_{2n}$

Assume: there is a natural pairing between x_{1i} & x_{2i} $\forall i = 1, \dots, n$

Before & After pre- post test.

Dependence

That is the two samples are dependent

One Sample → Lab1, Lab2

Sth to link them

Still 1-n → independence,

but 1-2 group → dependent.

x_1 x_{21}

We reduce the two samples to one sample by calculating the difference: $d_i = x_{1i} - x_{2i}$ $i = 1, \dots, n$

Assume:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right)$$

T.S. $t_{test} = \frac{\bar{d}}{sd/\sqrt{n}}$, where $\bar{d}$ is average difference — one sample t-test

sd is st. dev of difference.

Under H_0, $t_{test} \sim t_{n-1}$

RR: Reject H_0 if $|t_{test}| > t_{n-1}(\frac{\alpha}{2})$

$H_0: \mu_1 - \mu_2 = 0$ vs $H_1: \mu_1 - \mu_2 \neq 0$

Check the normality distribution → post hoc? ☆☒☒!

~~Multivariate~~ case:

$$\begin{matrix} x_{11} - x_{21} = d_1 \\ \vdots \quad \vdots \to \vdots \\ x_{1n} - x_{2n} = d_n \end{matrix} \Big\} \ p \times 1 \text{ vector}$$

Assumption $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_{2p} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{pmatrix} \right)$

Hypotheses:

$H_0: \mu_1 = \mu_2$ vs $\mu_1 \neq \mu_2$ ($p \times 1$ vector)

Cheat Sheet

T.S.

$T^2_{test} = n \bar{d}' S_d^{-1} \bar{d}$

where $\bar{d}$ is mean diff ($p \times 1$)

S_d is cov. matrix of diff ($p \times p$)

Too large → reject

↳ How large → can't use → F-stats.

$T^2_{test} > \frac{(n-1)p}{n-p} F_{p, n-p}(\alpha)$

$n \rightarrow$ sample size always

~~AAA~~ 2 groups?

Separate?

↓

Both

d is normally distributed

HUDM 6122

2.19

Week 5-2

For χ^2

Part 2

Model: $X_{ij} = \mu + \tau_i + \epsilon_{ij}$

$i = 1, \ldots, g$

$j = 1, \ldots, n_i$ $\quad n = n_1 + \cdots + n_g$

$X_{ij}, \mu, \tau_i, \epsilon_{ij}$ are all $p \times 1$ vectors

$\epsilon_{ij} \overset{ind.}{\sim} N_p(0, \Sigma)$

with $\sum_{i=1}^{g} n_i \tau_i = 0$

$H_0: \tau_1 = \tau_2 = \cdots = \tau_g = 0$

Idea:

$$x_{ij} - \bar{x} = (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i)$$

$$\Rightarrow \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})(x_{ij} - \bar{x})' \quad (p \times p)$$

Everything Covariates (Why not inner product $A'A$)

Outer Product (AA')

$$= \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(x_i - \bar{x})' \quad \} \text{ Between}$$

$$+ \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \quad \} \text{ Within}$$

Define:

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \quad (p \times p)$$

$$= (n_1 - 1)S_1 + \cdots + (n_g - 1)S_g$$

df ★★★

B.t

Determinate ↑

W ↓

W ⇒ residual error - within

$n - g$

Generalized ★★★ ★★★ ?

$$B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \quad (p \times p)$$

$g - 1$

Total $= B + W$

$n - 1$

Generalized Covariance?

T.S. $\Lambda_{test} = \dfrac{\det(W)}{\det(B+W)}$ Wilk

Rej. Reject H_0 if Λ_{test} is "small"

If not in 6.3 → χ^2.

Otherwise, use Bartlett approximation

$-(n-1-\frac{(p+g)}{2}) \cdot \ln(\lambda_{test}) \approx \chi^2_{p(g-1)}$ Likelyhood Ratio Test

$\lambda_{test} = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$, where $\lambda_1, \dots, \lambda_s$ eigenvalues not full rank ⇒ S

of $W^{-1}B$ ... done know eigenvalue.

S ⇒ the number of the non zero eigenvalue

Ex (6.9 on p.304): $p=2$, $g=3$, $n_1=3$, $n_2=2$, $n_3=3$ $n=8$ $\alpha=0.01$

x_1 → Measurement from Group 1 x_2 x_3

9	3		0	4	3	8
6	2		2	0	1	9
9	7				2	7

Whenever is called vector, is a column

① $\bar{x}_1 = \begin{bmatrix} 8 \\ 4 \end{bmatrix}$ $\bar{x}_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ $\bar{x}_3 = \begin{bmatrix} 2 \\ 8 \end{bmatrix}$ ---

$\bar{x} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$

Covariance of Midterm

Sum of square treatment, Sum of Square residual Always symetrics

② $B = \begin{bmatrix} 78 & -12 \\ -12 & 48 \end{bmatrix}$ $W = \begin{bmatrix} 10 & 1 \\ 1 & 24 \end{bmatrix}$ ←

Covariance ← Covariance within groups (all three)

between groups $B+W = \begin{bmatrix} 88 & -11 \\ -11 & 32 \end{bmatrix}$ SST

③ $\lambda_{test} = \frac{|W|}{|B+W|} = \frac{\begin{vmatrix} 10 & 1 \\ 1 & 24 \end{vmatrix}}{\begin{vmatrix} 88 & -11 \\ -11 & 32 \end{vmatrix}} = \frac{10(24)-1(1)}{88(32)-(-11)(-11)} = \frac{239}{6215} = 0.0385$

Midterm. ☆ 2×2 determinant.

2×2 matrix ⑩ $p=2$, $g=3$ ⇒ table ($p=2$ $g=3$) $n=8$ ... Table

⇒ $\frac{n-g-1}{g-1} \cdot \frac{1-\sqrt{\lambda}}{\sqrt{\lambda}} \sim F_{2(g-1),\,2(n-g-1)}$

Strive to reject the null → variance ⇔ $\frac{8-3-1}{3-1} \cdot \frac{1-\sqrt{0.0385}}{\sqrt{0.0385}} = 8.19$

Crit. Value:

$F_{4,8}(0.01) = 7.01$

$qf^{1-\alpha}(0.99, 4, 8)$

Converted Statistic ⑤ Conclusion: Since $8.19 > 7.01$, we reject the @ $\alpha = 0.01$ level

reject the: the three group doesn't have the same average of 3 parameter

That is, the three groups do not have the same mean vectors.

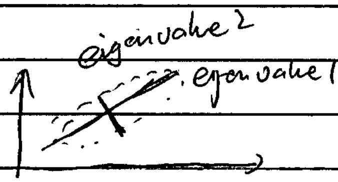
n → don't worry about it.

Third version — Use eigenvalue for Λ-test.

det is product of eigenvalues

~~☆☆☆~~ Wilk statistic

eigen = direction not change.



$p \times p$ & $p \times 1$

① $\bar{x}_i$, $\bar{x}$

② B W $B+W$

③ $\Lambda_{test} = \frac{|W|}{|B+W|}$ → Wilk's Λ

④ Third: p, g, n ⇒ T6.3 / χ^2

{ Wilk's Λ → Λtest
{ Crit. Value

Λtest > Crit. Value

(Wilk's lamda)

⑤ Conclu

g → Crit. values

$\lg = \log_{10} x$ — Common logarithm

~~$\ln$~~ $\ln = \log_e x$ — Natural logarithm

All tests. ~~☆☆☆~~

$\lambda_1, \dots, \lambda_s$ eigenvalues of $W^{-1}B$

Wilks: $\Lambda = \prod_{i=1}^{s} \left(\frac{1}{1+\lambda_i}\right)$ → Small

Roy: $\frac{\lambda_1}{1+\lambda_1}$, where λ_1 is the largest.

Pillai: $\sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i} = tr[(B+W)^{-1}B]$

Hotelling: $\sum_{i=1}^{s} \lambda_i = tr(W^{-1}B)$

} large value

F → Table 6.3

χ^2 → $p(g-1)$

χ^2
F
Looking χ^2, F table

{ Mon
{ Hote
{ Matrix

Nor

Hsiau 6122

[2.1]

Week 5-1

Analysis of Variance (MANOVA) t-test for more than 2 groups.

- univariate case (ANOVA)

Suppose we have "g" groups (populations)

From each we have a sample.

Group 1	...	Group g	
x_{11}		x_{g1}	n_1 $\bar{x}_{1\cdot}$ $\bar{x}$ n_g
x_{12}		x_{g2}	
⋮		⋮	
x_{1n_1}		x_{gn_g}	
$\bar{x}_1$	...	$\bar{x}_g$	$\bar{x}_{\cdot\cdot}$

Assume: For each $i = 1, \dots, g$

$x_{i1}, \dots, x_{ig} \sim N(\mu_i, \sigma^2)$ Normal

and the groups are independent. groups independent.

This set-up is known as one-way ANOVA 1 factor separate groups

It can be written as a probability model

$$x_{ij} = \mu + \tau_i + \varepsilon_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, \dots, g \quad j = 1, \dots, n_i$$

$\Rightarrow$ unobservable.

μ: Too specific effect of in group dummy

μ_i: group effect

τ_i: corres to group $1, \dots, g$

μ_i is the population mean of group i

ε_{ij} are random errors, normally distributed

We want to test:

$H_0: \mu_1 = \mu_2 = \dots = \mu_g$ vs.

H_1: At least two μ_i's are different μ & τ_i related.

Alternatively, we can test

$H_0: \tau_1 = \dots = \tau_g$

where $\sum_{i=1}^{g} n_i \tau_i = 0$

Idea

Between & Within Groups

$$x_{ij} = \bar{x} + (\bar{x}_i - \bar{x}) + (x_{ij} - \bar{x}_i) \quad i = 1, \dots, g \text{ (group)}, \quad j = 1, \dots, n_i \text{ (obs)}$$

$\Rightarrow$ observable

x_{ij}: observation

$\bar{x}$: overall average, no group

$\bar{x}_i$: ith group average, estimate of τ_i, Treatment effect

$x_{ij} - \bar{x}_i$: residual / error

$$\Leftrightarrow (x_{ij}-\bar{x})^2 = (\bar{x}_i-\bar{x})^2 + (x_{ij}-\bar{x}_i)^2 + 2(\bar{x}_i-\bar{x})(x_{ij}-\bar{x}_i)$$

$$\Leftrightarrow \sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = n_i(\bar{x}_i-\bar{x})^2 + \sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2 + 0$$

$$\Leftrightarrow \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{g} n_i(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

adding every single obs

SS_{Total} (Corrected) — $SS_{Treatment}$ (between groups sum of squares) — SS_{Resid} or SSE (within groups sum of squares)

R output

Group Different → SS_{Treat} ↑ B ... W

Code different → Mean.

ANOVA table:

Source	SS	df	MS	F
Treatment	SS_{Treat}	$g-1$	$SS_{Treat}/g-1$	MS_{Treat}
Residual	SS_{Resid}	$n-g$	$SS_{Resid}/n-g$	MS_{Resid}
Total	SS_{Total}	$n-1$		

bigger F ratio stats, ↑ reject null hyp

rej!!!

where $n = n_1 + n_2 + \dots n_g$

(not a matrix, different length)

T.S. $F_{test} = \dfrac{MS_{Treat}}{MS_{Resid}}$ ~ $F_{g-1,\,n-g}$

R.R. Reject H_0 if $F_{test} > F_{g-1,\,n-g}(\alpha)$.

(larger value for F statistics, reject)

$\bar{x}_{11}$ $\bar{x}_{12}$??? Noise.

x_{11}, x_{12}, x_{13} x_{21}, x_{22}, x_{23}

$\bar{x}_1$ $\bar{x}_2$

x_{12} ...

Note:

We reject H_0 if SS_{Treat} is "larger" than SS_{Resid}

$\Leftrightarrow \dfrac{SS_{Treat}}{SS_{Resid}}$ ↑ is "large"

SS Treat 커짐.

$\Leftrightarrow 1 + \frac{SS_{Treat}}{SS_{Resid}}$ is "large"

$\Leftrightarrow \frac{SS_{Total}}{SS_{Resid}}$ is "large"

$\Leftrightarrow \frac{SS_{Resid}}{SS_{Total}}$ is "small" ⇒ Analogy to Multivariate

① $p=1$, $g=$, $n=$
② $\bar{x}_i$, $\bar{x}$
③ SS_{To}, SS_{Tre}, SS_{Res}
④ Table
⑤ $F_{g-1, n-g}$
⑥ Con

Ex) 6.7 on p.298

Group 1	Group 2	Group 3
9	0	3
6	2	1
9		2
$n_1=3$	$n_2=2$	$n_3=3$

$g=3$ (In R: length cannot be different)

In R:

x	Group
9	g_1
6	g_1
9	g_1
0	g_2
2	g_2
3	g_3
1	g_3
2	g_3

$n = 3+2+3 = 8$

$\bar{x}_1 = \frac{9+6+9}{3} = 8$ Sum of Squares

$\bar{x}_2 = 1$

$\bar{x}_3 = 2$

$\bar{x} = \frac{9+6+\cdots+2}{8} = 4$

Corrected ?

- $SS_{Total} = SS - SS_{Mean}$ (Incorrected) $\sum (x_{ij} - \bar{x})^2$
  Sum of Squares $n \cdot \bar{x}^2$ $= \sum (x_{ij}^2 - 2\bar{x}x_{ij} + \bar{x}^2)$
  $= 9^2 + 6^2 + \cdots + 2^2 - n \cdot \bar{x}^2 = 216 - 128 = 88$ $= \sum x_{ij}^2 - 2\bar{x}\sum x_{ij} + n\bar{x}^2$ ($\sum x_{ij} \to n\bar{x}$)
- $SS_{Treat} = 3(8-4)^2 + 2(1-4)^2 + 3(2-4)^2$
  $= 78$
- $SS_{Resid} = 88 - 78 = 10$

$H_0: \mu_1 = \mu_2 = \mu_3$ $\alpha = 0.01$

Source	SS	df	MS	F	P-value
Treatment	78	2	39	$\frac{39}{2} = 19.5$	? in R
Resid	10	5	2		
Total	88	7			

F test = 19.5 df → critical value.

Crit. Value: $F_{2,5}(0.01) = 13.2$ ↓ page

Conclusion: Since $[9.5 > 13.2]$

we reject H_0

That is, at least two group [ ] means are different

In the multivariate case, the "between" and "within" SS are matrices!

$SS_{Treat} \to B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$

$\bar{x}_i, \bar{x} \to p \times 1$ vectors

$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x})'$ ← SS_{resid}

Main idea: W should be "smaller" compared to Total = B + W

T.S. $\frac{\det(W)}{\det(B+W)}$ → "small" → reject H_0

★★★ Statistic Test → only F table

H_0: hotellings table

- one M
- one H_0
- generals: Matrices + Normal Distribution

MANOVA – Part 3 Equal Covariance Assumption Testing

HUDM 6122

2.24

Week 6-1

– Testing for equal covariance matrices

Recall: MANOVA assumptions

$x_{ij} = \mu + \gamma_i + \varepsilon_{ij}$, $i = 1, \cdots, g$

$j = 1, \cdots, n_i$

① g, I_n

② within each g N

Assume: ε_{ij} is $N_p(0, \Sigma)$ ↘ same covariance for all groups

Groups are independent

$H_0: \gamma_1 = \cdots \gamma_g = 0$

③ Same

$H_0 = \Sigma_1 = \Sigma_2 = \cdots = \Sigma_g$

where Σ_i is the covariance of group i

Box's M test:

$$M = \left[\sum_{i=1}^{g} (n_i - 1)\right] \ln(\det(S_{pool})) - \sum_{i=1}^{g} (n_i - 1) \ln(\det(S_i))$$

Define:

$$C = \left[\sum_{i=1}^{g} \frac{1}{n_i} - \frac{1}{\sum_{i=1}^{g} (n_i - 1)}\right] \frac{2p^2 + 3p - 1}{6(p+1)(g-1)}$$

Newton Iteration

From Likelihood Ratio test.

p, g, n

↓ ↓

g=k=group

Under H_0,

$C = (1 - U) M \approx \chi^2_v$

where df. $v = \frac{p(p+1)(g-1)}{2}$

Under H_0,

Distribution.

RR:

Reject H_0 if

$C > \chi^2_v(\alpha)$

Remark: The test works well if $n_i > 20$, $i = 1, \cdots, g$

and $p, g \leq 5$

Remark: Very sensitive to violations of normality and as a result produces very small p-values

High sensitivity ⇒ low power

Reject Null