

多元统计分析

第5讲 多元正态分布 (3)

Johnson & Wichern Ch6.1-6.3, 6.6, Ch4.6-4.8

统计学研究中心 邓婉璐

wanludeng@tsinghua.edu.cn

2018-2019春季学期

Example – Scores for Statistical Courses

- Scores for 5 courses.
- Graduate students ($n=50$) vs Undergraduate students ($n=100$)
- Interested in
 - Whether different score for two groups (in terms of mean)?
 - What is our confidence about the difference?
 - What is our confidence about the GPA difference?
 - What if we have multiple ways for calculating GPA?

Outline

- Inference for two-sample normal population mean
- Assumption assessment
- Pipeline for real application
- (Advanced) Multiple testing – FDR vs FWER (e.g. Bonferroni correction)

Comparing Mean Vectors from Paired Samples

Sample and Sample Statistics

Sample:
n units in total,
for the j^{th} unit

X_{1j1} = variable 1 under treatment 1
 X_{1j2} = variable 2 under treatment 1
 \vdots
 X_{1jp} = variable p under treatment 1
 X_{2j1} = variable 1 under treatment 2
 X_{2j2} = variable 2 under treatment 2
 \vdots
 X_{2jp} = variable p under treatment 2

Denote:

$$E(D_j) = \delta = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_p \end{bmatrix}, \text{Cov}(D_j) = \Sigma_d$$

Sample difference: $D_{j1} = X_{1j1} - X_{2j1}$

$$D_{j2} = X_{1j2} - X_{2j2}$$

$$\vdots \quad \vdots$$

$$D_{jp} = X_{1jp} - X_{2jp}$$

$$D_j = X_{1j} - X_{2j} = [D_{j1}, \dots, D_{jp}]'$$

Sample statistics: $\bar{D} = \frac{1}{n} \sum_{j=1}^n D_j$

$$S_d = \frac{1}{n-1} \sum_{j=1}^n (D_j - \bar{D})(D_j - \bar{D})'$$

Confidence Region

Res 6.1

Let the differences D_1, \dots, D_n be a random sample from an $N_p(\delta, \Sigma_d)$ population. Then

$$T^2 = n(\bar{D} - \delta)' S_d^{-1} (\bar{D} - \delta)$$

is distributed as an $\frac{(n-1)p}{n-p} F_{p, n-p}$ random variable, whatever the true δ and Σ_d .

If n and $n-p$ are both large, then T^2 is approximately distributed as a χ^2 random variable, regardless of the form of the underlying population of differences.

Hypothesis testing ?

Simultaneous Confidence Interval ?

Example

$$H_0: \delta=0$$

Undergraduate:

```
> head(data1)
```

	Prob	Inference	Computing	MVA	LinearReg
1	83	75	88	86	79
2	85	74	90	85	79
3	84	81	90	84	75
4	83	77	91	91	82
5	87	70	89	85	83
6	84	79	91	86	73

Graduate:

```
> head(data2)
```

	Prob	Inference	Computing	MVA	LinearReg
1	83	81	85	82	76
2	88	71	82	81	77
3	84	75	82	80	78
4	83	74	85	81	69
5	81	71	85	81	69
6	85	74	82	80	75

Difference:

```
> head(D)
```

	Prob	Inference	Computing	MVA	LinearReg
1	0	-6	3	4	3
2	-3	3	8	4	2
3	0	6	8	4	-3
4	0	3	6	10	13
5	6	-1	4	4	14
6	-1	5	9	6	-2

Reject H_0

```
> T2 > cf
      [,1]
[1,] TRUE
> T2
      [,1]
[1,] 813.571
> cf
[1] 13.18691
```

Relevant to the order of the sample

e.g., > T2

```
      [,1]
[1,] 606.2999
```

> T2

```
      [,1]
[1,] 1070.666
```

So we require meaningful 'paired'!

Confidence Region

Res 6.1

Let the differences D_1, \dots, D_n be a random sample from an $N_p(\delta, \Sigma_d)$ population. Then

$$T^2 = n(\bar{D} - \delta)' S_d^{-1} (\bar{D} - \delta)$$

is distributed as an $\frac{(n-1)p}{n-p} F_{p, n-p}$ random variable, whatever the true δ and Σ_d .

If n and $n-p$ are both large, then T^2 is approximately distributed as a χ^2 random variable, regardless of the form of the underlying population of differences.

Equivalent
representation

$$X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \end{pmatrix}$$

$$\bar{X} = \frac{1}{n} \sum_{j=1}^n X_j$$

$$\bar{D} = C\bar{X}$$

$$S = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})'$$

$$S_d = CSC'$$

$$T^2 = n(C\bar{X})'(CSC')^{-1}(C\bar{X})$$

$$= n\bar{X}'S^{-1}\bar{X}$$

当C可逆的时候

Comparing Mean Vectors from Two Populations – Equal Variance, Normal Distribution

Sample and Sample Statistics

Sample

$$\vec{x}_{11}, \vec{x}_{12}, \dots, \vec{x}_{1n_1}$$

$$\vec{x}_{21}, \vec{x}_{22}, \dots, \vec{x}_{2n_2}$$

Sample mean

$$\bar{x}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_{1j}$$

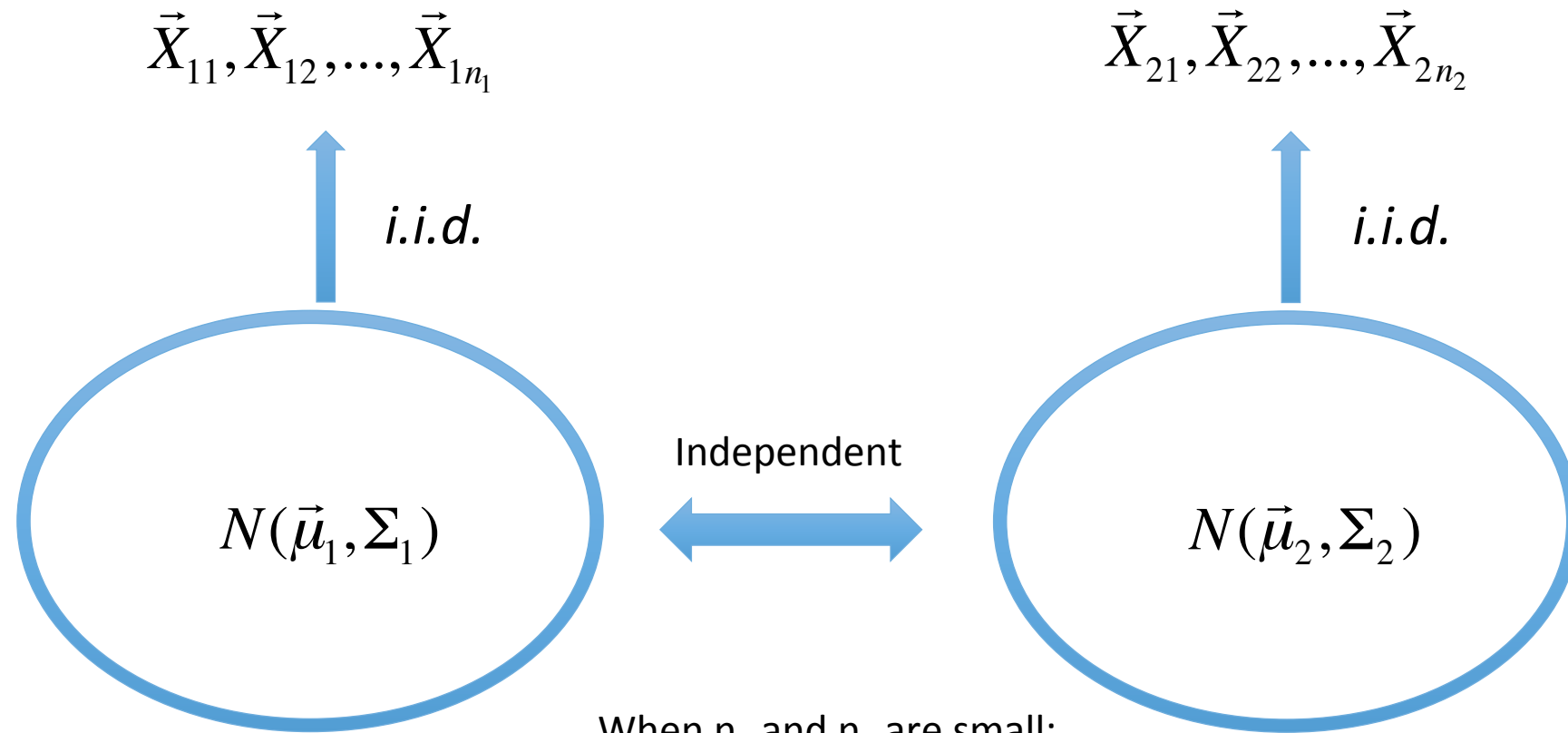
$$\bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} x_{2j}$$

Sample
covariance
matrix

$$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)'$$

$$S_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)'$$

Assumptions



When n_1 and n_2 are small:

$$\Sigma_1 = \Sigma_2 = \Sigma$$

Recall: Hotelling's T^2 for One Population

Consider the point null hypothesis,

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0,$$

For p-dim:

$$\begin{aligned} T^2 &= \sqrt{n}(\bar{X} - \mu_0)' \left(\frac{\sum_{j=1}^n (X_j - \bar{X})(X_j - \bar{X})'}{n-1} \right)^{-1} \sqrt{n}(\bar{X} - \mu_0) \\ &= \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right)' \left(\frac{\text{Wishart random matrix}}{d.f.} \right)^{-1} \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right) \end{aligned}$$

$$\text{Under } H_0: = N_p(0, \Sigma)' \left[\frac{1}{n-1} W_{p,n-1}(\Sigma) \right]^{-1} N_p(0, \Sigma)$$

Recall: Hotelling's T^2 for One Population

Under H_0 we have

$$\frac{n-p}{p} \frac{T^2}{n-1} \sim F_{p, n-p},$$

and we reject if

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

Hotelling's T^2 for Two Populations

Consider the point null hypothesis,

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq \delta_0$$

We need to find the corresponding sample mean and sample variance:

Since $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$

$$Cov(\bar{X}_1 - \bar{X}_2) = Cov(\bar{X}_1) + Cov(\bar{X}_2) = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

方差矩阵相等的时候

Thus, sample mean: $\bar{X}_1 - \bar{X}_2$

What about sample covariance matrix?


Hotelling's T^2 for Two Populations

Heuristically, consider:

Both S_1 and S_2 are unbiased estimator for Σ , with n_1-1 d.f. and n_2-1 d.f. respectively.

Combine the information together to get an estimator for Σ :

$$\begin{aligned} S_{pooled} &= \frac{\sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)' + \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)'}{n_1 + n_2 - 2} \\ &= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2 \sim \frac{W_p(n_1 + n_2 - 2, \Sigma)}{n_1 + n_2 - 2} \end{aligned}$$

 $\left\{ \begin{array}{l} (n_1 - 1)S_1 \sim W_p(n_1 - 1, \Sigma) \\ (n_2 - 1)S_2 \sim W_p(n_2 - 1, \Sigma) \end{array} \right.$

This expression can also be obtained through MLE.

Hotelling's T^2 for Two Populations

Consider the point null hypothesis,

$$H_0 : \mu_1 - \mu_2 = \delta_0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq \delta_0$$

Thus,

$$T^2 = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} (\bar{X}_1 - \bar{X}_2 - \delta_0)' S_{pooled}^{-1} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-\frac{1}{2}} (\bar{X}_1 - \bar{X}_2 - \delta_0)$$

$$= \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right)' \left(\frac{\text{Wishart random matrix}}{d.f.} \right)^{-1} \left(\begin{array}{c} \text{multivariate normal} \\ \text{random variable} \end{array} \right)$$

$$= N_p(0, \Sigma)' \left[\frac{W_p(n_1 + n_2 - 2, \Sigma)}{n_1 + n_2 - 2} \right]^{-1} N_p(0, \Sigma)$$

Hypothesis Testing for Two Populations

Under H_0 we have $\frac{(n_1 + n_2 - p - 1)}{p} \frac{T^2}{n_1 + n_2 - 2} \sim F_{p, n_1 + n_2 - p - 1}$

and we reject if

$$T^2 > T^2(\alpha) = \frac{p}{(n_1 + n_2 - p - 1)} (n_1 + n_2 - 2) F_{p, n_1 + n_2 - p - 1}(\alpha)$$

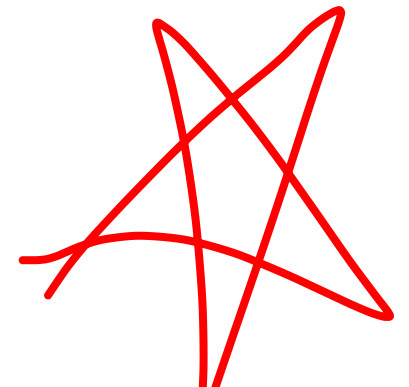
Confidence Region

Res 6.2
Confidence Region

Equivalently, the corresponding $100(1-\alpha)\%$ confidence region is

$$(\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))' \left(\left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right)^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)) \leq T^2(\alpha)$$

先有置信区域，然后考虑各个方向的投影：点乘某一个向量



Simultaneous Coverage Intervals

Recall 1-population:

Another type of interval with a simultaneous coverage property under the normal likelihood is the T^2 interval

$$I_a(c) = a'\bar{x} \pm c\sqrt{a'Sa}$$

Projection of Confidence Region!

with the property that if

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

then

$$\mathbb{P}[I_a(c) \text{ covers } a'\mu \text{ for all } a \neq 0] = 1 - \alpha.$$

$1-\alpha = P(T^2 \leq c^2) = P(t_a^2 \leq c^2, \forall a)$ 其中 t_a 为枢轴量, 定义见下

Simultaneous Coverage Intervals

On the other hand, can we find a lower bound for c , such that

$$I_a(c) = a'\bar{x} \pm c\sqrt{a'\frac{S}{n}a} \quad \text{covers } a'\mu \text{ for all } a \neq 0 \quad ?$$



By maximization lemma
(2-50) in Chapter 2

$$\begin{aligned} \max_a t^2 &= \max_a \frac{n(a'(\bar{x} - \mu))^2}{a'Sa} \\ &= n \max_a \frac{(a'(\bar{x} - \mu))^2}{a'Sa} = n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) = T^2 \end{aligned}$$

Simultaneous Coverage Intervals

Res 6.3
Simultaneous
Confidence
Intervals

$$\text{Let } c^2 = \frac{p}{(n_1 + n_2 - p - 1)} (n_1 + n_2 - 2) F_{p, n_1 + n_2 - p - 1}(\alpha)$$

Then with probability $1-\alpha$,

$$a'(\bar{X}_1 - \bar{X}_2) \pm c \sqrt{a' \left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} a}$$

will cover $a'(\mu_1 - \mu_2)$ for all a . In particular $\mu_{1i} - \mu_{2i}$ will be covered by

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii, pooled}}$$

Proof with projection is simpler.

Simultaneous Coverage Intervals

Res 6.3
Simultaneous
Confidence
Intervals

$$\text{Let } c^2 = \frac{p}{(n_1 + n_2 - p - 1)} (n_1 + n_2 - 2) F_{p, n_1 + n_2 - p - 1}(\alpha)$$

Then with probability $1-\alpha$,

$$a'(\bar{X}_1 - \bar{X}_2) \pm c \sqrt{a' \left(\frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} a}$$

will cover $a'(\mu_1 - \mu_2)$ for all a . In particular $\mu_{1i} - \mu_{2i}$ will be covered by

$$(\bar{X}_{1i} - \bar{X}_{2i}) \pm c \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii, pooled}}$$

Which direction quantifies the largest population difference?

$$a \propto S_{pooled}^{-1} (\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2))$$

Example

$$H_0: \mu_1 - \mu_2 = 0$$

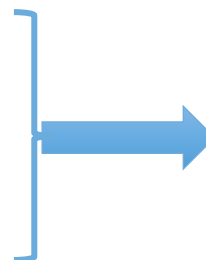
Undergraduate: > X1

	[,1]
Prob	84.98
Inference	74.78
Computing	89.16
MVA	84.97
LinearReg	78.60

Graduate:

> X2

	[,1]
Prob	82.78
Inference	70.54
Computing	82.22
MVA	81.08
LinearReg	72.76



Reject H_0

```
> T2 > cf
      [,1]
[1,] TRUE
> T2
      [,1]
[1,] 1027.917
> cf
[1] 11.70147
```

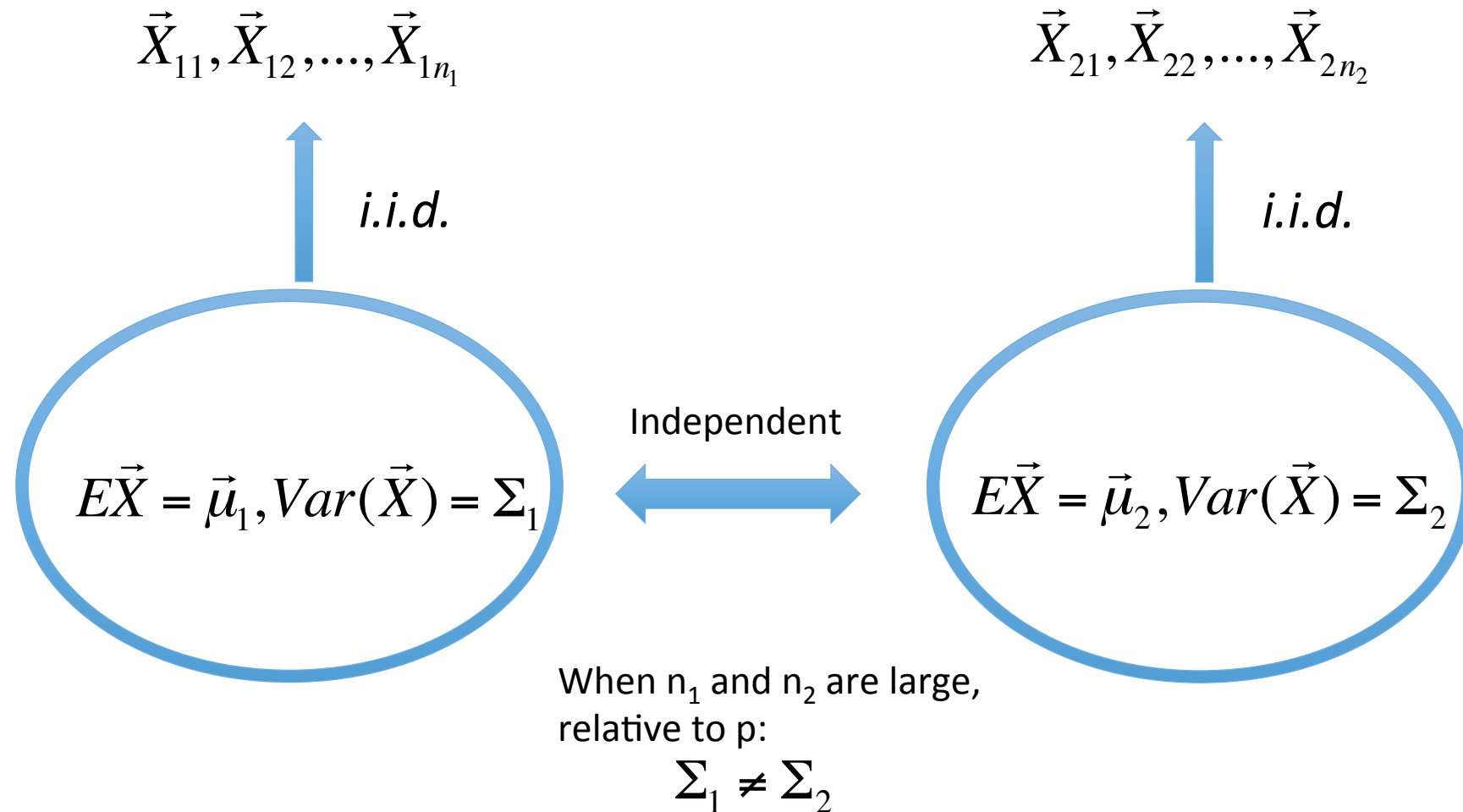
Direction for largest difference

```
> solve(Sp) %*% (X1-X2)
      [,1]
Prob      -0.4338307
Inference -0.5747851
Computing  3.3909387
MVA        0.3036343
LinearReg  1.6292411
```

Irrelevant to the order of the sample

Comparing Mean Vectors from Two Populations – Unequal Variance, Large Sample

Assumptions



Large Sample Inference

Recalling large sample theories.

Since $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$

$$\text{Cov}(\bar{X}_1 - \bar{X}_2) = \text{Cov}(\bar{X}_1) + \text{Cov}(\bar{X}_2) = \frac{1}{n_1} \Sigma_1 + \frac{1}{n_2} \Sigma_2$$

Thus, by CLT: $\bar{X}_1 - \bar{X}_2 \rightarrow N_p(\mu_1 - \mu_2, \frac{1}{n_1} \Sigma_1 + \frac{1}{n_2} \Sigma_2)$

by LLN:

$$\left. \begin{array}{l} S_1 \xrightarrow{p} \Sigma_1 \\ S_2 \xrightarrow{p} \Sigma_2 \end{array} \right\} \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \xrightarrow{p} \frac{1}{n_1} \Sigma_1 + \frac{1}{n_2} \Sigma_2$$

Res 4.13



Confidence Region

Res 6.4 Confidence Region

Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are large.
An approximate $100(1-\alpha)\%$ confidence ellipsoid for $(\mu_1 - \mu_2)$ is given by

$$(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)) \leq \chi_p^2(\alpha)$$

这是因为，样本量比较大的时候，
T₂近似服从 χ^2 分布

Example

$$H_0: \mu_1 - \mu_2 = 0$$

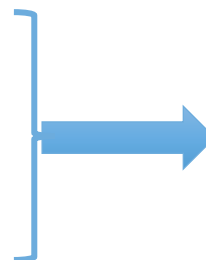
Undergraduate: > X1

	[,1]
Prob	84.98
Inference	74.78
Computing	89.16
MVA	84.97
LinearReg	78.60

Graduate:

> X2

	[,1]
Prob	82.78
Inference	70.54
Computing	82.22
MVA	81.08
LinearReg	72.76



Reject H_0

```
> T2 > c_chisq
      [,1]
[1,] TRUE
> T2
      [,1]
[1,] 1075.743
> c_chisq
[1] 11.0705
```

Irrelevant to the order of the sample!

Simultaneous Coverage Intervals

Res 6.4
Simultaneous
Confidence
Intervals

Let $c^2 = \chi_p^2(\alpha)$

Then with probability $1-\alpha$,

$$a'(\bar{X}_1 - \bar{X}_2) \pm c \sqrt{a'(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2)a}$$

will cover $a'(\mu_1 - \mu_2)$ for all a .

Proof with projection is simpler.

Sample Covariance Matrix

Assume $n_1=n_2$

Large sample size;
unequal variance

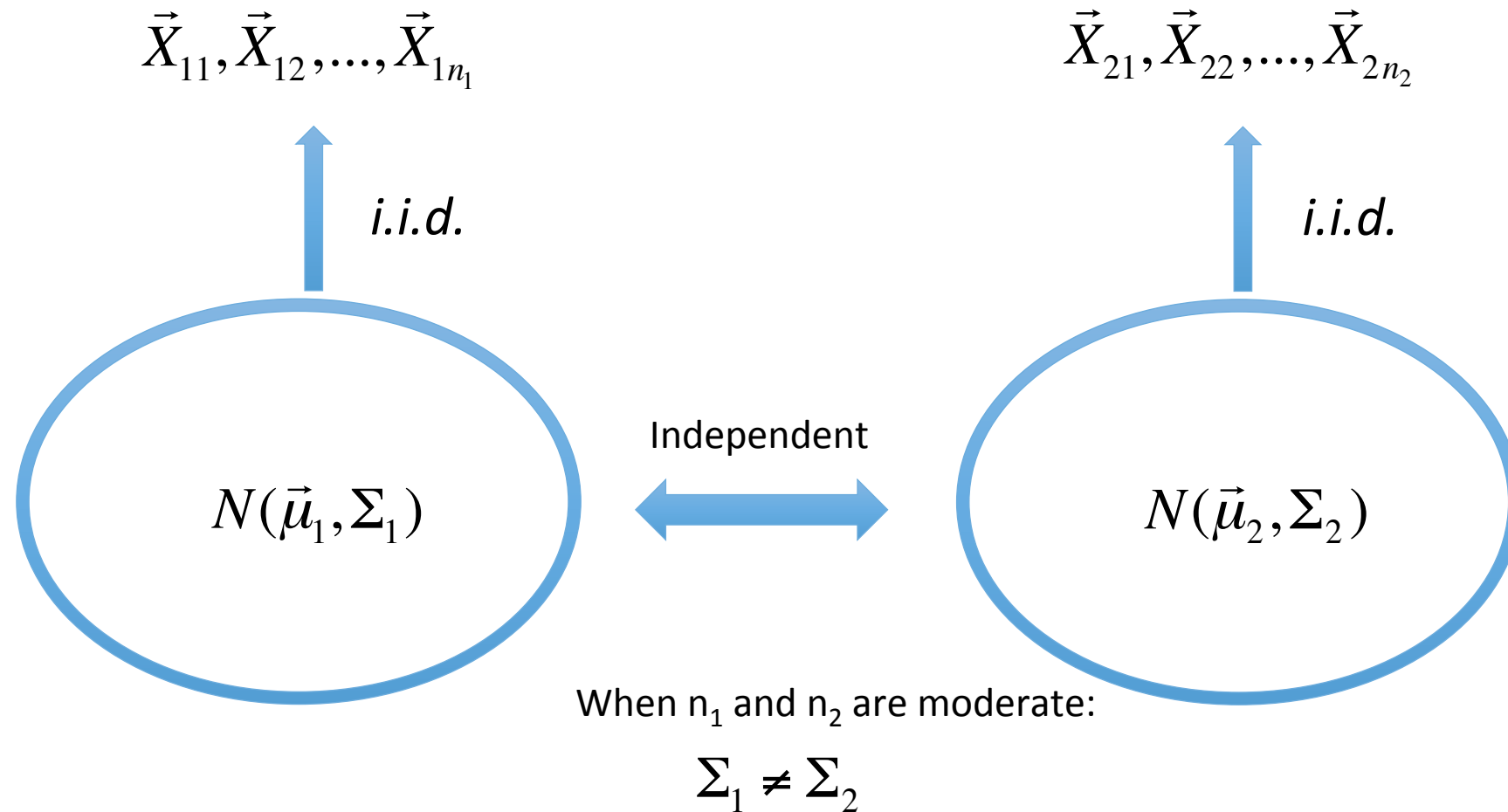
$$\frac{1}{n_1}S_1 + \frac{1}{n_2}S_2 = \frac{1}{n}(S_1 + S_2) = \frac{(n-1)S_1 + (n-1)S_2}{(n+n-2)}\left(\frac{1}{n} + \frac{1}{n}\right) = S_{pooled}\left(\frac{1}{n} + \frac{1}{n}\right)$$

Moderate sample size
normal distribution;
equal variance

Comparing Mean Vectors from Two Populations – Unequal Variance, Moderate Sample Size

(not required)

Assumptions



Confidence Region

Behrens-Fisher problem
Confidence Region

Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are positive.
An approximate $100(1-\alpha)\%$ confidence ellipsoid for $(\mu_1 - \mu_2)$ is given by

$$(\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2))' \left(\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} (\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)) \leq \frac{vp}{v - p + 1} F_{p, v-p+1}(\alpha)$$

与大样本时采用的统计量相同

Corresponding to Welch solution to the Behrens-Fisher problem in the univariate ($p=1$) case.

When can we apply T^2 Statistics?

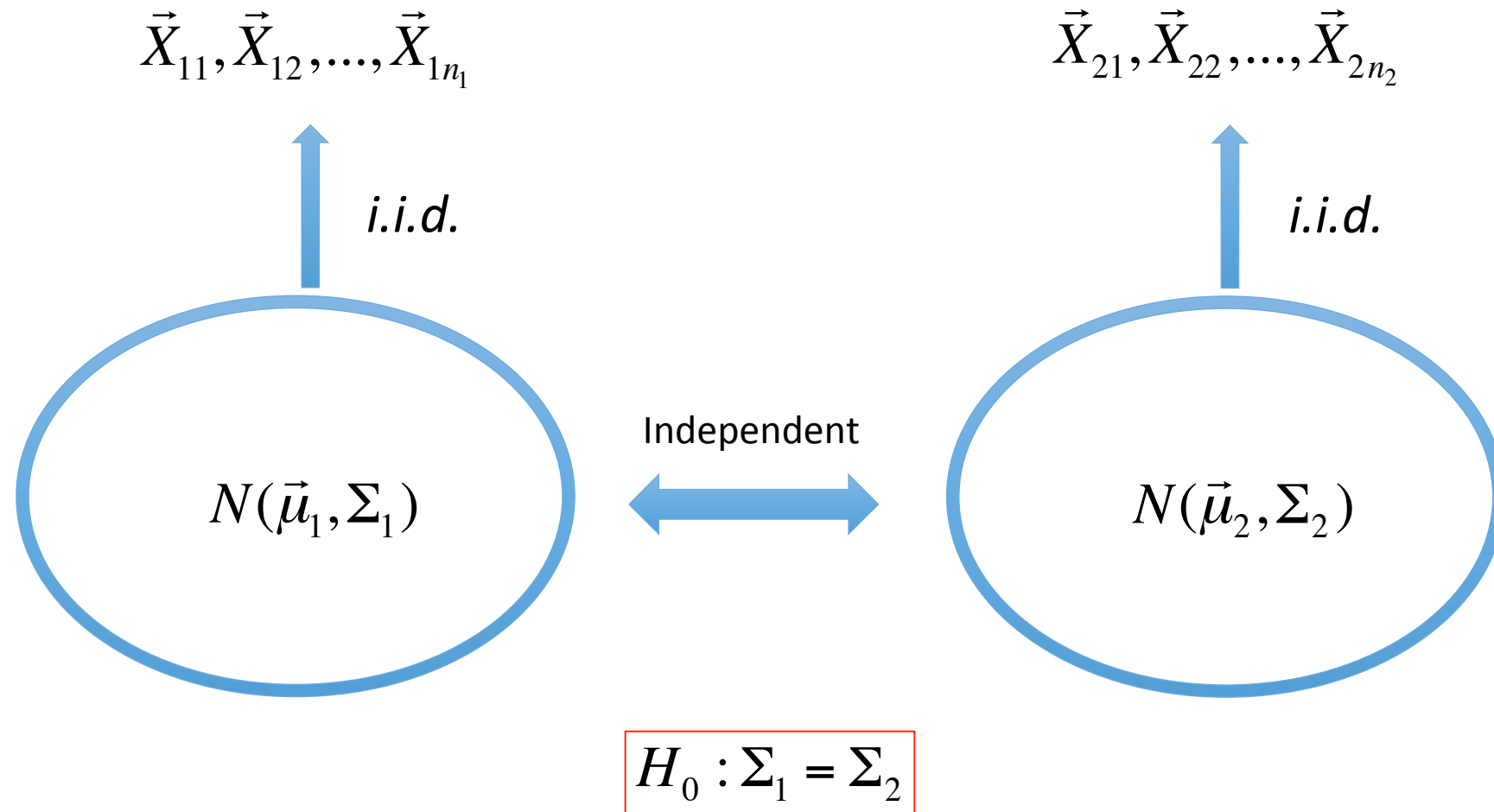
Assumptions underlying T^2 Statistics

Let's recall the assumptions underlying the Hotelling's T^2 test.

- **Equal variance.** The data from both populations have common variance-covariance matrix Σ .
- **Independence.** The subjects from both populations are independently sampled.
(Note that this does not mean that the variables are independent of one another.)
- **Normality.** Both populations are multivariate normally distributed.

Testing Equal Variance Matrix for Two Populations

Assumptions



LRT for Testing Equal Variance

$$H_0 : \Sigma_1 = \Sigma_2$$

Hint: Nested Model.

Recall:

Res 5.2
LRT

Denote the likelihood ratio test statistic $\Lambda = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)}$

and $v - v_0 = (\text{dimension of } \Theta) - (\text{dimension of } \Theta_0)$

Then under H_0 , asymptotically

$$-2 \ln \Lambda \sim \chi^2_{v-v_0}$$

一般的极大似然检验

LRT for Testing Equal Variance

Box's M Test
(a modified LRT)

For samples from two multivariate normal population, under H_0 , the following statistic follows a chi-square distribution approximately,

$$(1 - u)(-2 \ln \Lambda) \rightarrow \chi_v^2, \text{ where}$$

$$\Lambda = \prod_{i=1}^2 \left(\frac{|S_i|}{|S_{pooled}|} \right)^{(n_i-1)/2},$$

$$v = 2\left(\frac{1}{2}p(p+1)\right) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1),$$

$$u = \left[\sum_i \frac{1}{n_i - 1} - \sum_i \frac{1}{n_i} \right] \left(\frac{2p^2 + 3p - 1}{6(p+1)(2-1)} \right)$$

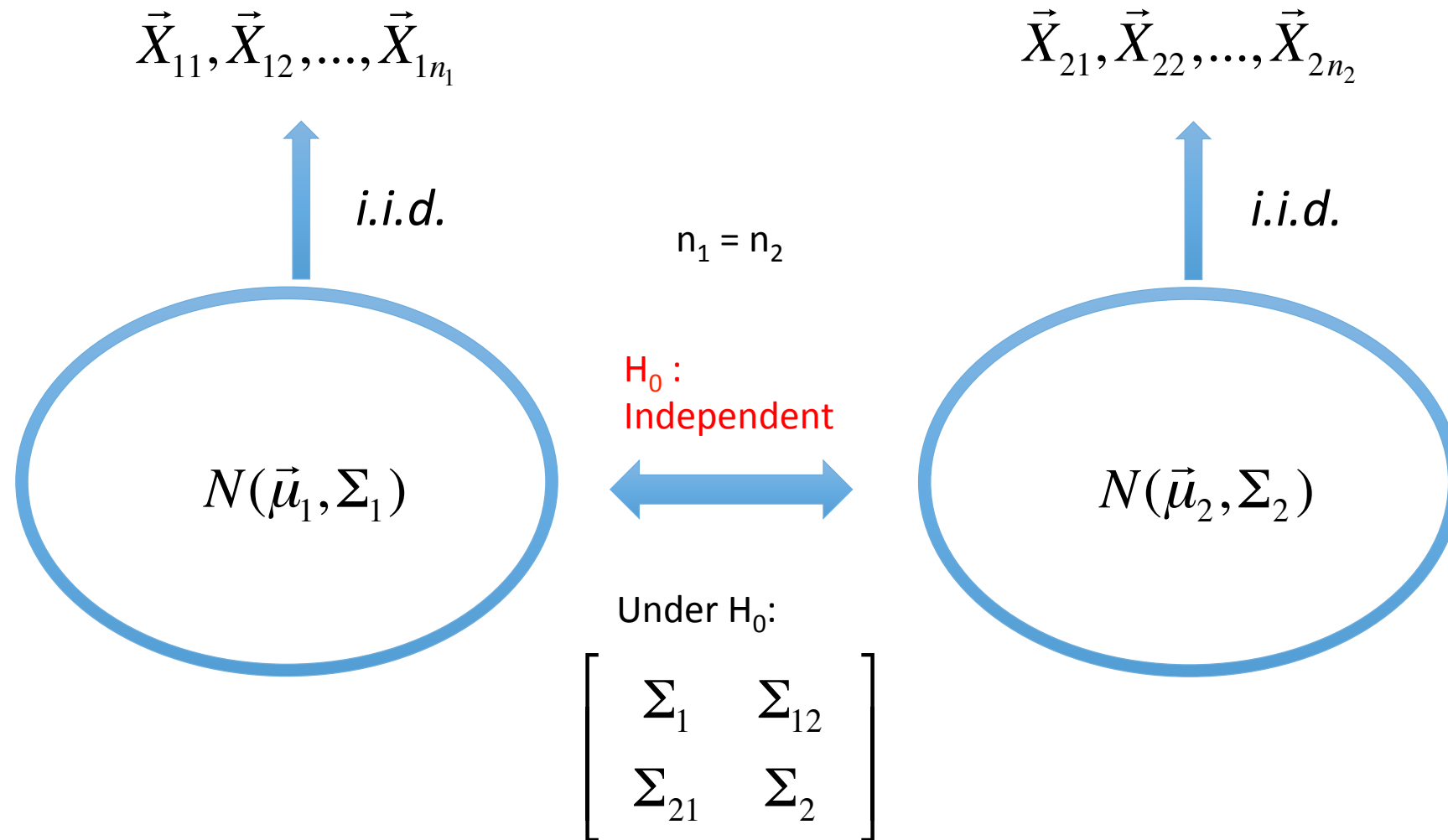
finite population
correction factor

Comments:

- Ref to Ch 6.6 for results for multiple groups.
- Understand the idea; details are not required.
- Alternative: Bartlett's Test
- sensitive to departures from normality

Testing Independence for Two Populations

Assumptions



LRT for Testing Independence

$$H_0 : \Sigma_{12} = 0$$

Hint: Nested Model.

For samples from two multivariate normal population, under H_0 , the LRT statistic follows a chi-square distribution asymptotically,

$$u(-2 \ln \Lambda) \rightarrow \chi_v^2, \text{ where}$$

$$\Lambda = \left(\frac{\prod_i |S_i|}{|S|} \right)^{-n/2},$$

finite population
correction factor

$$u = 1 - \frac{p}{n} - \frac{3}{2n},$$

$$v = p^2$$

Understand the idea;
details are not required.

Example

Scores for Courses in Statistics

Two group: Undergraduate vs Graduate

Assumption Assessment

$$H_0: \Sigma_1 = \Sigma_2$$

```
> C > c_chisq_2  
[1] FALSE  
> C  
[1] 3.176235  
> c_chisq_2  
[1] 24.99579
```

Can't reject H_0

H_0 : Two groups are independent.

```
> C > c_chisq_3  
[1] FALSE  
> C  
[1] 30.89084  
> c_chisq_3  
[1] 37.65248
```

Can't reject H_0

Assessing the Assumption of Normality - Transformation

Assessing the Assumption of Normality

➤ Univariate case

- Histogram
- Q-Q plot

➤ Multivariate case

Let \mathbf{X} be distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

(1) the marginal distribution X_i is normal

← Histogram or
normal probability plot

(2) the linear combination of X_i is normal

← $\hat{\mathbf{e}}_1' \mathbf{x}_j$, where $\mathbf{S}\hat{\mathbf{e}}_1 = \hat{\lambda}_1 \hat{\mathbf{e}}_1$

(3) contour : $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ is distributed as χ_p^2

Mahalanobis Distance

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$$

$$\{d_1^2, d_2^2, \dots, d_n^2\}$$

Q-Q Plot

➤ Univariate case

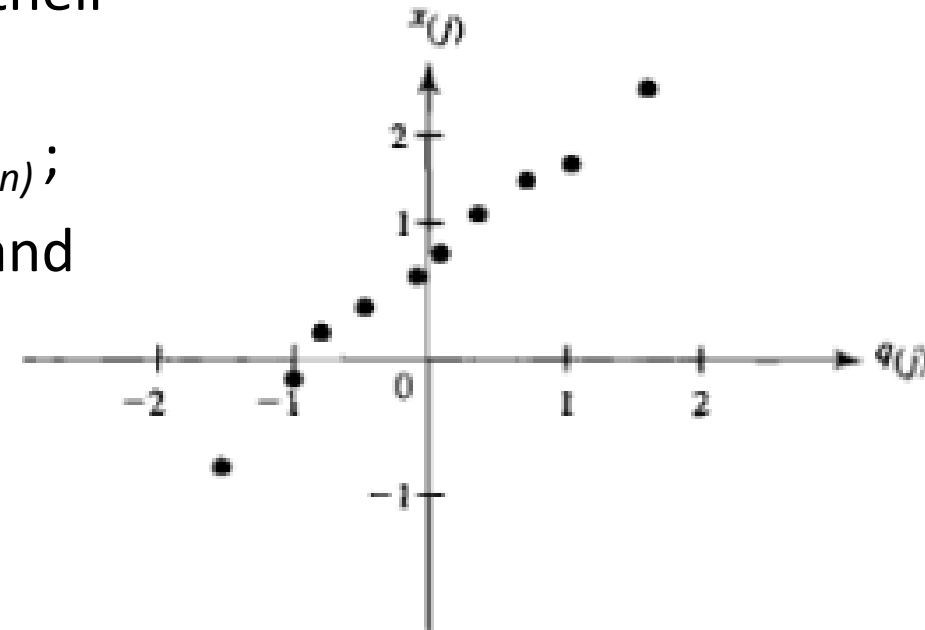
Let $x_{(1)}, \dots, x_{(n)}$ represents the ordered observations. $p_{(j)} = \frac{j - \frac{1}{2}}{n}; q_{(j)} = \Phi^{-1}(p_{(j)})$

Steps:

1. Order the original observations to get $x_{(1)}, \dots, x_{(n)}$ and their corresponding probability values $p_{(1)}, \dots, p_{(n)}$;
2. Calculate the standard normal quantiles get $q_{(1)}, \dots, q_{(n)}$;
3. Plot the pairs of observations $(q_{(1)}, x_{(1)}), \dots, (q_{(n)}, x_{(n)})$, and examine the “straightness” of the outcome,

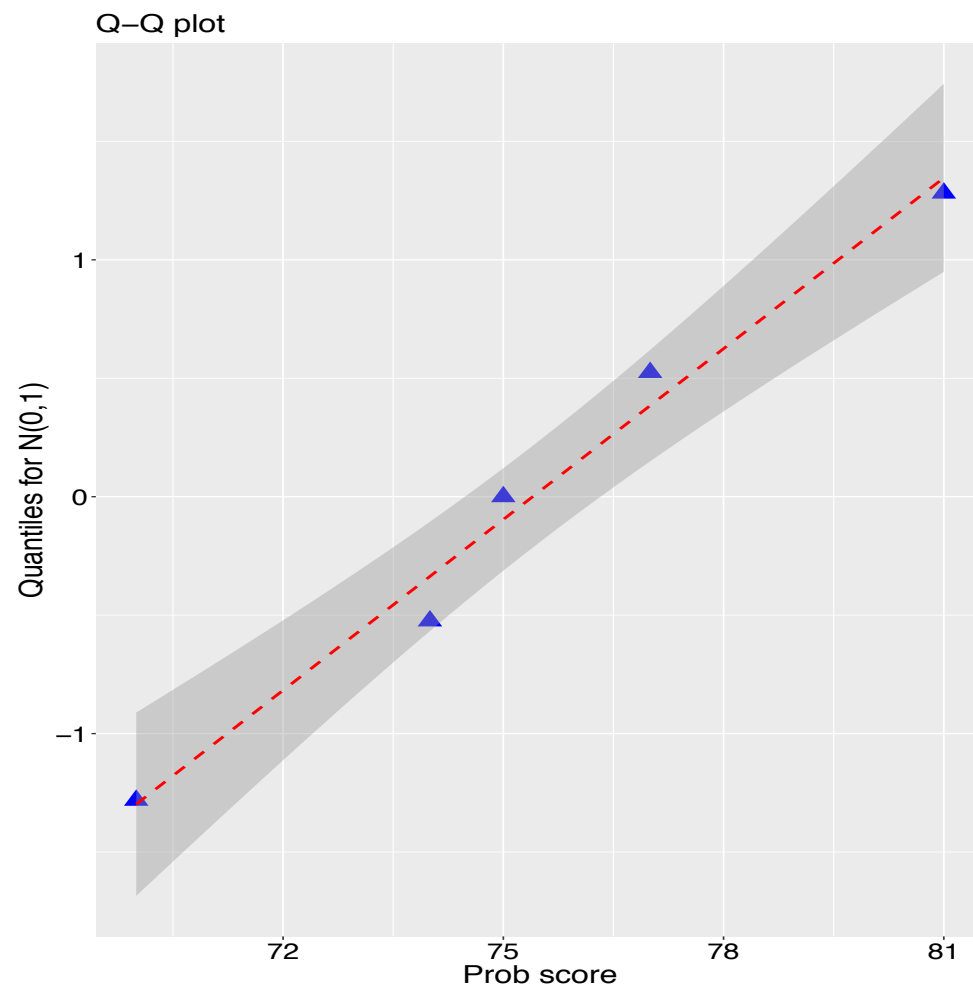
e.g. test by

$$r_Q = \frac{\sum_{j=1}^n (x_{(j)} - \bar{x})(q_{(j)} - \bar{q})}{\sqrt{\sum_{j=1}^n (x_{(j)} - \bar{x})^2} \sqrt{\sum_{j=1}^n (q_{(j)} - \bar{q})^2}}$$



Example

```
> x = data1[1:5,2]
> x
[1] 75 74 81 77 70
> x = sort(x)
> x
[1] 70 74 75 77 81
> n = length(x)
> p = (1:n-1/2)/n
> p
[1] 0.1 0.3 0.5 0.7 0.9
> q = qnorm(p)
> q
[1] -1.2815516 -0.5244005  0.0000000  0.5244005  1.2815516
```



Tests for multivariate normality

➤ Multivariate case

– Chi-square plot

Steps:

1. Order the squared distances from smallest to largest as $d^2_{(1)}, \dots, d^2_{(n)}$ and their corresponding probability values $p_{(1)}, \dots, p_{(n)}$;
2. Calculate quantiles for the chi-square distribution with p degrees of freedom: $q_{(1)}, \dots, q_{(n)}$, where $q_{(j)} = q_{c,p}(p_{(j)}) = \chi^2_p(1-p_{(j)})$, the upper percentiles of the chi-squared distribution;
3. Plot the pairs of observations $(q_{(1)}, d^2_{(1)}), \dots, (q_{(n)}, d^2_{(n)})$, and examine the “straightness” of the outcome (compared to the line $y=x$)

Tests for multivariate normality

Not required

- There are a number of tests for multivariate normality, some of which are implemented in the R package **MVN**.
- For example, Mardia's test (based on multivariate extensions of skewness and kurtosis measures).

Lizard Data:

```
mardiaTest(lizard,qqplot=F)

##      Mardia's Multivariate Normality Test
## -----
##      data : lizard
##
##      g1p           : 0.4536732
##      chi.skew       : 1.890305
##      p.value.skew   : 0.9971138
##
##      g2p           : 11.8382
##      z.kurtosis     : -1.443159
##      p.value.kurt   : 0.1489756
##
##      chi.small.skew : 2.246763
##      p.value.small  : 0.9940751
##
##      Result         : Data are multivariate normal.
## -----
```

When Your Data is NOT Normal

Count data, proportion data, correlation, ...

Transformation!

Original Scale

1. Counts, y

2. Proportions, p

3. Correlations, r

4. ...

Transformed Scale

$$\sqrt{y}$$

$$\text{logit}(p) = \frac{1}{2} \log\left(\frac{p}{1-p}\right)$$

$$z(r) = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right)$$

Fisher's z transformation

Power transformations



When the choice of a transformation to improve the approximation to normality is not obvious.
Let the data suggest a transformation.

When Your Data is NOT Normal

- Box-Cox transformations- the best known one among power transformations

$$z = \frac{x^\lambda - 1}{\lambda}, \quad \text{with } z = \log(x) \text{ for } \lambda = 0$$

The solution maximizing the following expression

$$l(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{j=1}^n (z_j - \bar{z})^2 \right] + (\lambda - 1) \sum_{j=1}^n \ln x_j$$

- Can be generalized to p -dim

$$l(\lambda_1, \dots, \lambda_p) = -\frac{n}{2} \ln |S(\lambda)| + (\lambda_1 - 1) \sum_{j=1}^n \ln x_{j1} + \dots + (\lambda_p - 1) \sum_{j=1}^n \ln x_{jp}$$

Find maximization by iteration.

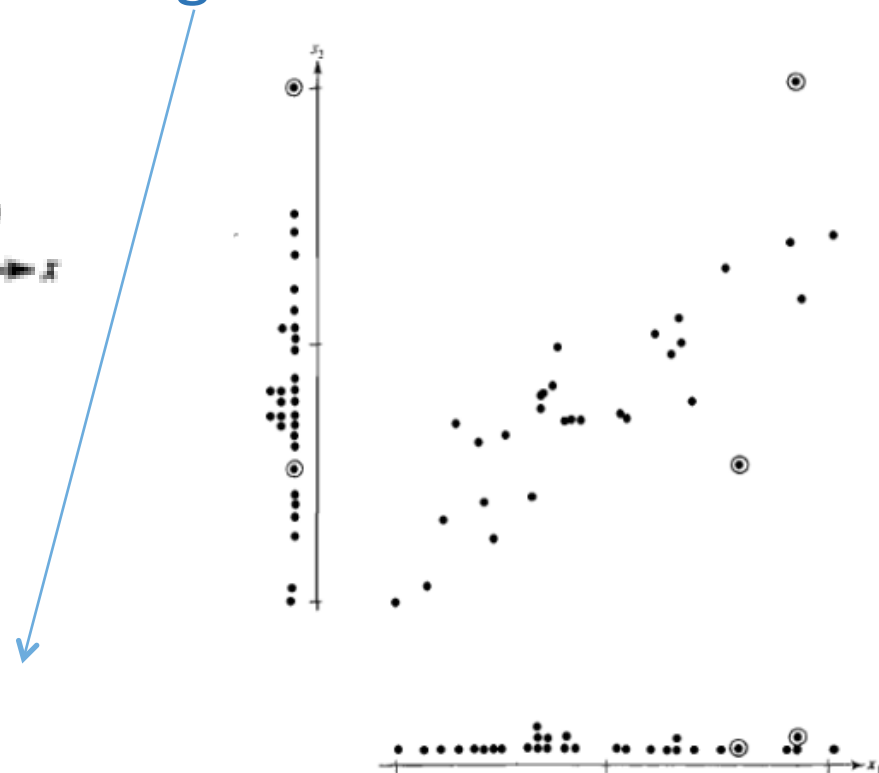
Outlier Detection

Outlier Detection

1. Make a dot plot for each variable
2. Make a scatter plot for each pair of variables
3. Calculate the standardized values, $z_{jk} = (x_{jk} - \bar{x}_k) / \sqrt{s_{kk}}$
examine these standardized values for large or small values



relative to the sample size and the
number of variables

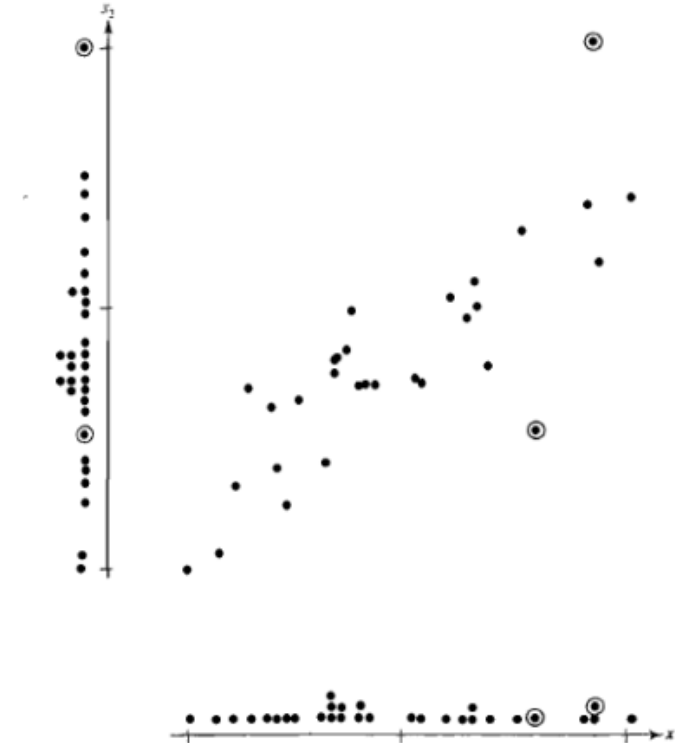
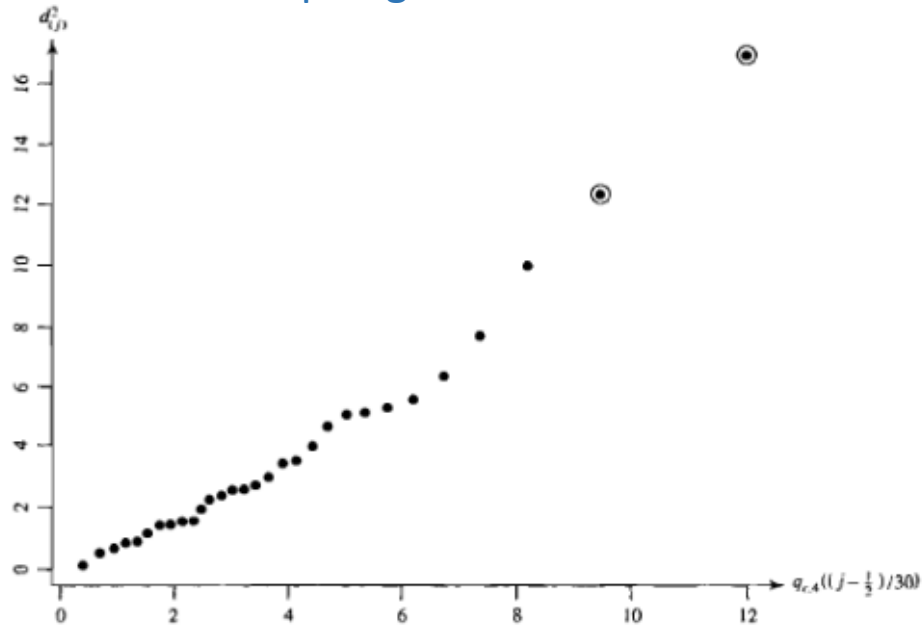


Outlier Detection

4. Calculate the Mahalanobis distance $(\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$

Examine the distances for **unusually large** values

measured by an appropriate percentile of the chi-square distribution with p degrees of freedom.



Outlier Detection

4. Calculate the Mahalanobis distance $(\mathbf{x}_j - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$

Examine the distances for unusually large values

What to do with outliers?

Depending upon the nature of the outlier and the objectives of the investigation,

Outliers may be deleted or appropriately weighted in subsequent analysis.

Pipeline for Real Data Analysis

For Real Data Analysis

In practice, the data analyses should proceed as follows:

- Step 1. For small samples, use histograms, scatterplots, and rotating scatterplots to assess the multivariate normality of the data. If the data do not appear to be normally distributed, apply appropriate normalizing transformations.
- Step 2. Assess the assumption that the population variance-covariance matrices are homogeneous.
- Step 3.
 - (A.1) Carry out the two-sample Hotelling's T-square test for equality of the population mean vectors.
 - (A.2) If the test in Step 2 is significant, use the modified two-sample Hotelling's T-square test.
 - (B.1) If the two-sample Hotelling's T-square test is not significant, conclude that there is no statistically significant evidence that the two populations have different mean vectors and stop.
 - (B.2) Otherwise, go to Step 4.
- Step 4. Compute either simultaneous confidence intervals. For the significant variables, draw conclusions regarding which population has the larger mean.

Multiple Testing

False Discovery Rate

vs

Familywise Error Rate

Hypothesis Test

$$H_0 : \mu_1 = \mu_2$$

$$H_{0j} : \mu_{1j} = \mu_{2j},$$

Different!

The classical tests we just covered won't work when $p > n$, but the methods we'll talk about now can be applied in that setting as well. Moreover, they aren't restricted to that setting, and in fact the problem of multiple testing is relevant whenever $p > 1$.

Bonferroni Correction

- If we test H_{0j} at level α , then by definition the probability of a Type I error is α . Accordingly, we can recast level α testing as reject $H_0 \tilde{p} < \alpha$.
- Since we are testing p such hypotheses, the probability of making at least one type I error – referred to as the familywise error rate (FWER) is larger than α .

$$\mathbb{P} \left[\bigcup_{j=1}^p \tilde{p}_j < \frac{\alpha}{p} \mid H_0 \right] \leq \sum_{j=1}^p \mathbb{P} \left[\tilde{p}_j < \frac{\alpha}{p} \mid H_0 \right] = p \frac{\alpha}{p} = \alpha,$$

- So the Familywise error rate is controlled at level α .

Bonferroni Correction

- Suppose that the hypothesis tests are all independent, and that every one of the null hypotheses H_{0j} is actually true. Then if we test each at level α , the number of Type I errors (false positives) V is distributed as

$$V \sim \text{Binomial}(p, \alpha)$$

- Then in particular, $P[V > p\alpha] = 0.5$, assuming $p\alpha$ is an integer. So if $\alpha = 0.05$ and $p = 1000$, we will make more than 50 “mistakes” (type I errors) half of the time. This is an obvious problem in the era of modern science, where often p is in the hundreds or thousands and n is similar to p or smaller.

False Discovery Rate

- On the other hand, controlling the FWER is pretty conservative. It might not be so bad to have a few type I errors.
- In particular, it might be ok so long as type I errors are a relatively small proportion of the total number of rejections of H_{0j} .
- In particular, we want to choose a level α to perform each test such that $V / (V + S)$ is small, where

$$V = \sum_j \mathbb{1}_{\{\tilde{p}_j < \alpha, H_{0j}\}}$$
$$S = \sum_j \mathbb{1}_{\{\tilde{p}_j < \alpha, H_{1j}\}},$$

False Discovery Rate

	Null hypothesis is true (H_0)	Alternative hypothesis is true (H_A)	Total
Test is declared significant	V	S	R
Test is declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

In this notation, the FWER is equal to $P[V \geq 1]$. The quantity

$$Q_e = \mathbb{E}[Q] = \mathbb{E} \left[\frac{V}{V + S} \right]$$

is called the **False discovery rate (FDR)**. Q_e is the expectation of the unobserved random variable Q .

False Discovery Rate

- Let p_0 be the number of null hypotheses that are true. The following two basic facts are important:
 1. If all the null hypotheses are true, then FDR is equivalent to FWER.
 2. When $p_0 < p$, the FDR is less than or equal to the FWER. Therefore, any procedure that controls FWER also controls FDR. However, a procedure that controls FDR only is less strict, so there is the potential for higher power. In particular, when $p - p_0$ is large, S tends to be large, resulting in a larger difference between $E[Q]$ and $P[V \geq 1]$.
- Thus, controlling FDR at level α is less conservative than controlling FWER at level α , since it allows us to make multiple type I errors on expectation, so long as they don't account for too high a proportion of the total number of “discoveries” (hypothesis tests for which we reject the null).

Motivating Example

- The following dataset has gene expression measurement for two samples:
 - n_1 from prostate cancer tissue; n_2 from healthy prostate tissue.
 - The expression levels of $p = 6033$ genes were measured.
 - The scientific question of interest is whether the genes are differentially expressed in cancer and healthy tissue. ([Reference: Efron's Large Scale Inference \(2014\)](#))

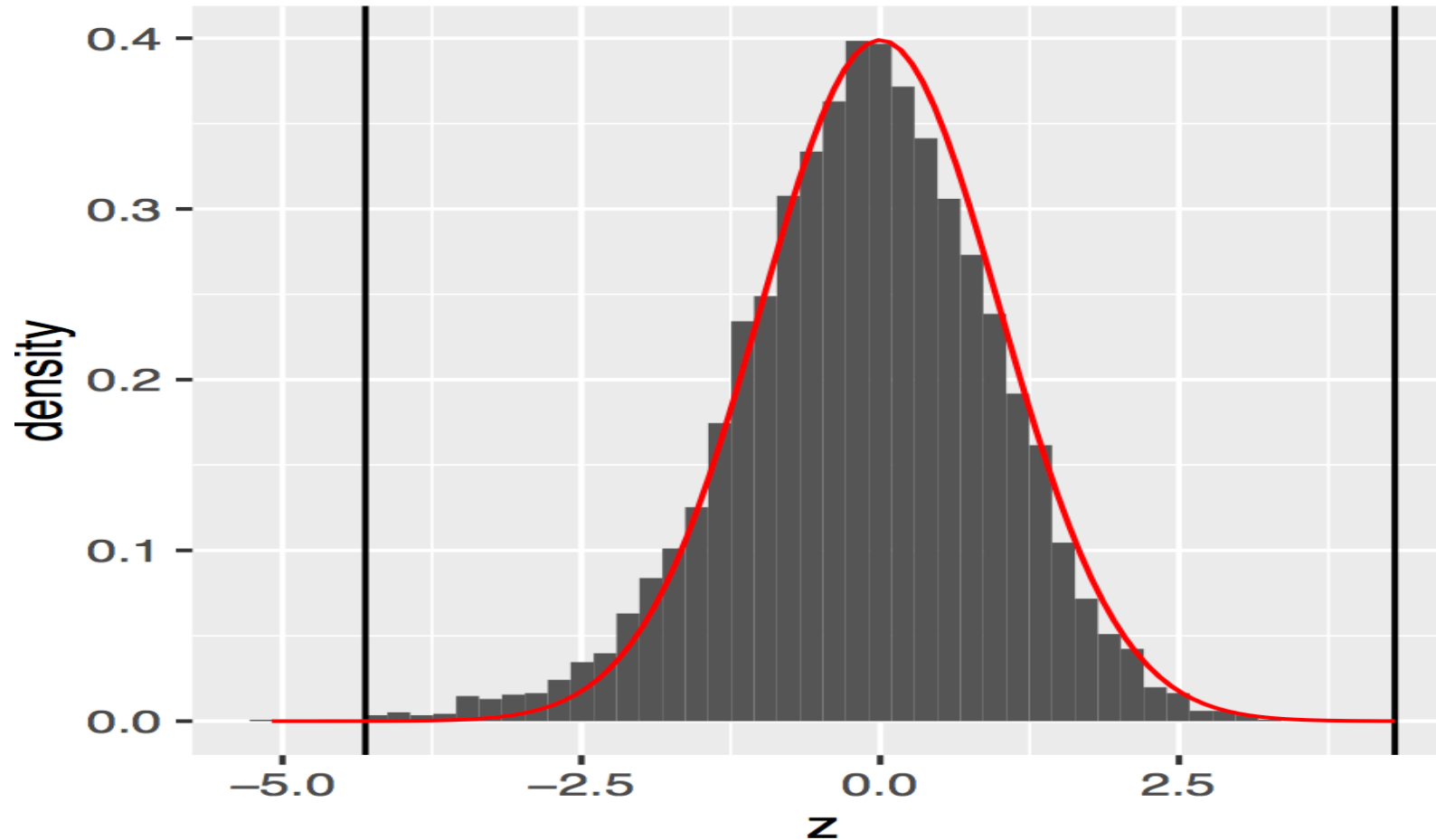
$$H_0 : \mu_{1j} = \mu_{2j}$$

$$t_j = \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\sqrt{s_j}}, \quad s_j = \frac{s_{1j} + s_{2j}}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

It will be convenient to transform to z-values from t-values, i.e.

$$z_j = \Phi^{-1}(2T_{n_1+n_2-2}(|t_j|)),$$

Motivating Example



z scores for the prostate data, with the Bonferroni threshold at level 0.05 shown as vertical line and the theoretical distribution of the z score under the null overlaid in red.

Motivating Example

As you can see, if we applied the Bonferroni bounds to control FWER, we would not have many “discoveries” (rejections of the null). In fact, there are only 2 out of the 6033 genes.

```
bonf.p <- p.adjust(pvals,method='bonferroni')  
sum(bonf.p<.05)  
  
## [1] 2
```

FDR control: the method of Benjamini and Hochberg

Perhaps the earliest, and simplest, method for control of FDR was proposed by [Benjamini and Hochberg \(1995\)](#). We give the procedure here. For a proof that the procedure controls FDR at the specified level, see the original paper. Let $\tilde{p}_1 \leq \tilde{p}_2 \leq \dots \tilde{p}_p$ be the ordered p-values. Let k be the largest j for which

$$\tilde{p}_j \leq \frac{j}{p} \alpha,$$

and reject all H_j , $j = 1, \dots, k$. If the test statistics are either independent or satisfy a positive correlation condition (see [Benjamini and Yekutieli \(2001\)](#)), this procedure controls FDR at level α .

Motivating Example

Let's use the Benjamini-Hochberg procedure on the prostate cancer data:

```
bh.p <- p.adjust(pvals,method='BH')  
sum(bh.p<.05)  
  
## [1] 21
```

As expected, we have considerably more “discoveries” (rejections of the null) – 21 instead of 2 at the 0.05 level.

Summary

Summary

- Inference for two-sample normal population mean
 - Different situations with different assumption, but no new idea
 - Paired samples
 - (Unequal) moderate sample size, equal covariance matrix, normal distribution
 - Large sample size, (unequal) covariance matrix
 - Moderate / large sample; precise / asymptotic distribution under null
- Assumption assessment
 - Normality (Q-Q plot); Transformations to near normality; Outlier detection
 - Equal variance: LRT
 - Independence: LRT
- Pipeline for real application
- (Not required) Alternative to Multiple testing – FDR