Tests of Hypotheses of Means

Instructor: Cheng-Ying Chou

Tests of Hypotheses of Means

- Introduction
- Review of Univariate One-sample t Test
- Multivariate One-sample T^2 Test
- Review of Univariate Two-sample t Test
- Multivariate Two-sample T^2 Test
- Profile Analysis
- Summary

Introduction

- Wechsler Adult Intelligence Scale scores of 101 elderly men and women (60-64, years old, Morrison, 2005)

Item	Mean	Variance	Covariance
Verbal	55.24	210.54	126.99
Performance	34.97	119.68	

Introduction

- The mean scores for the population between 18 and 59 are 60 and 50 for verbal and performance items, respectively.
- Univariate question:
  - Is the mean score of verbal item for elderly population different from that of the population between 18 and 59?
- Multivariate question:
  - Are the Wechsler Adult Intelligence Scale scores for elderly population different from those of the population between 18 and 59?

Introduction

Survivors and non-survivors for Bumpus's female sparrows

	Survivors		Nonsurvivors	
Variables	mean	variance	Mean	Variance
Total length	157.38	11.05	158.43	15.07
Alar extent	241.00	17.50	241.57	32.55
Length of beak and head	31.43	0.53	31.48	0.73
Length of humerus	18.50	0.18	18.45	0.43
Length of keel of sternum	20.8	0.58	20.84	1.32

Introduction

- Univariate questions:
  - Does difference exist in each of the five morphological measurements between the survivors and non-survivors?
- Multivariate questions:
  - Is there any difference in morphology between the survivors and non-survivors?

One sample t-test

- Let $x_1, x_2, \ldots, x_n$ denote a sample of n from the normal distribution with mean μ and variance σ^2.
- Hypothesis of means
  - H_0: $\mu=\mu_0$, H_a: $\mu\neq\mu_0$
  - Reject the null hypothesis at the α significance level if

$$|t| = \left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right| > t_{\alpha/2,n-1}$$

  - At 100(1- α) confidence interval

$$\bar{X} - t_{\alpha/2,n-1}\sqrt{s^2/n} \le \mu_0 \le \bar{X} + t_{\alpha/2,n-1}\sqrt{s^2/n}$$

Univariate one-sample t test

- Wechsler Adult Intelligence Scale (WAIS) verbal scores
- H_0: $\mu_v=60$, H_a: $\mu_v \neq 60$

$$\bar{X}_v = 55.24 \qquad \sigma_v^2 = 210.54$$

$$|t| = \left| \frac{55.24 - 60}{\sqrt{210.54/101}} \right| = 3.2969 > t_{0.005,100} = 2.60$$

- Reject the null hypothesis at 0.01 significance level. 99% confidence interval for μ_v is (51.4861, 58.9939)

Hotelling t squared test

- Univariate t^2 follow an F distribution with 1 and n-1 d.f.
- t and t^2 are unitless
  - Example of verbal scores
  - $t^2 = (-4.76)^2/(210.54/101)$

    $= (-4.76)(210.54/101)^{-1}(-4.76)$

    $=10.8693 > F_{0.01,1,100}=6.895$

Multivariate One-sample T2 Test

- Multivariate Hypothesis of One-sample Means

$$H_0 : \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{10} \\ \mu_{20} \\ . \\ . \\ \mu_{p0} \end{pmatrix} = \boldsymbol{\mu_0}$$

vs.

$$H_a : \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_p \end{pmatrix} \neq \begin{pmatrix} \mu_{10} \\ \mu_{20} \\ . \\ . \\ \mu_{p0} \end{pmatrix} = \boldsymbol{\mu_0}$$

The multivariate Test

- Let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ denote a sample of *n* from the *p*-variate normal distribution with mean vector $\vec{\mu}$ and covariance matrix $\mathbf{X}$.
- Suppose we want to test

$$H_0 : \vec{\mu} = \vec{\mu}_0 \quad \text{vs}$$
$$H_A : \vec{\mu} \neq \vec{\mu}_0$$

Covariance

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}(X_i - \mu_i)(X_j - \mu_j),$$

- Multivariate covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top,$$

Body measure of men and women

chest	waist	hips	gender	chest	waist	hips	gender
34	30	32	male	36	24	35	female
37	32	37	male	36	25	37	female
38	30	36	male	34	24	37	female
36	33	39	male	33	22	34	female
38	29	33	male	36	26	38	female
43	32	38	male	37	26	37	female
40	33	42	male	34	25	38	female
38	30	40	male	36	26	37	female
40	30	37	male	38	28	40	female
41	32	39	male	35	23	35	female

Covariance matrix by R

```
cov(measure[, c("chest", "waist", "hips")])
```

	chest	waist	hips
chest	6.631579	6.368421	3.000000
waist	6.368421	12.526316	3.578947
hips	3.000000	3.578947	5.944737

```
cov(subset(measure, gender == "female")[,
c("chest", "waist", "hips")])
```

	chest	waist	hips
chest	2.277778	2.166667	1.555556
waist	2.166667	2.988889	2.755556
hips	1.555556	2.755556	3.066667

```
cov(subset(measure, gender == "male")[,
c("chest", "waist", "hips")])
```

	chest	waist	hips
chest	6.7222222	0.9444444	3.944444
waist	0.9444444	2.1000000	3.077778
hips	3.9444444	3.0777778	9.344444

Multivariate One-sample T2 Test

Let $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$ denote a sample of *n* from the *p*-variate normal distribution with mean vector $\vec{\mu}$ and covariance matrix Σ.

Suppose we want to test

- Test statistic
- Hotelling T2 Statistic: a generalization of univariate *t* statistic by union-intersection principle

$$\text{or} \quad t^2 = n\left(\bar{\vec{x}} - \vec{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\vec{x}} - \vec{\mu}_0\right) > t^2_{\alpha/2}$$

Transformation of the Hotelling T^2 statistic

$$F = \frac{(n-p)\, n\left(\bar{\vec{x}} - \vec{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\vec{x}} - \vec{\mu}_0\right)}{p(n-1)}$$

$$F = \frac{n-p}{p(n-1)} T^2 = \frac{(n-p)\, n}{p(n-1)} \left(\bar{\vec{x}} - \vec{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\vec{x}} - \vec{\mu}_0\right) > F_{\alpha, p, n-p}$$

It turns out that if H_0 is true than has an F distribution with $\nu_1 = p$ and $\nu_2 = n - p$.

Multivariate One-sample T2 Test

- Wechsler Adult Intelligence Scale (WAIS) scores

The null hypothesis is given as

$$H_0 : \mu = \begin{pmatrix} \mu_v \\ \mu_p \end{pmatrix} = \begin{pmatrix} \mu_{10} = 60 \\ \mu_{20} = 50 \end{pmatrix}$$

The sample mean vector is

$$\bar{\mathbf{X}} = \begin{pmatrix} 55.24 \\ 34.97 \end{pmatrix}, \text{ and}$$

The sample covariance matrix is

$$\mathbf{S} = \begin{pmatrix} 210.54 & 129.99 \\ 129.99 & 119.68 \end{pmatrix}$$

Multivariate One-sample T2 Test

- Wechsler Adult Intelligence Scale verbal scores
  - The inverse of the sample covariance matrix is given as

$$\mathbf{S}^{-1} = \begin{pmatrix} 0.01319 & -0.01400 \\ -0.01400 & 0.02321 \end{pmatrix}$$

It follows that

$$T^2 = 101\begin{pmatrix} 54.24\text{-}60 & 34.97\text{-}50 \end{pmatrix}\begin{pmatrix} 0.01319 & -0.01400 \\ -0.01400 & 0.02321 \end{pmatrix}\begin{pmatrix} 54.24\text{-}60 \\ 34.97\text{-}50 \end{pmatrix}$$

$$= 357.43$$

$$F = \frac{101\text{-}2}{2(101\text{-}1)}357.43 = 176.93$$

$$> F_{0.01,2,99} = 4.98$$

The verbal and performance means of this elderly population is different (lower) from those of the general population.

Multivariate One-sample T^2 Test

The 99% region is given as

$$101\begin{pmatrix} 55.24 - \mu_v \\ 34.97 - \mu_p \end{pmatrix}\begin{pmatrix} -0.01319 & -0.01400 \\ -0.01400 & 0.02321 \end{pmatrix}\begin{pmatrix} 55.24 - \mu_v \\ 34.97 - \mu_p \end{pmatrix} \leq \frac{(101-1)2}{(101-2)} 4.83$$

$$\Rightarrow$$

$$1.33(\mu_v - 55.24)^2 - 2.83(\mu_v - 55.24)(\mu_p - 55.24) + (\mu_v - 34.97)^2 \leq 9.76$$

Multivariate One-sample T^2 Test

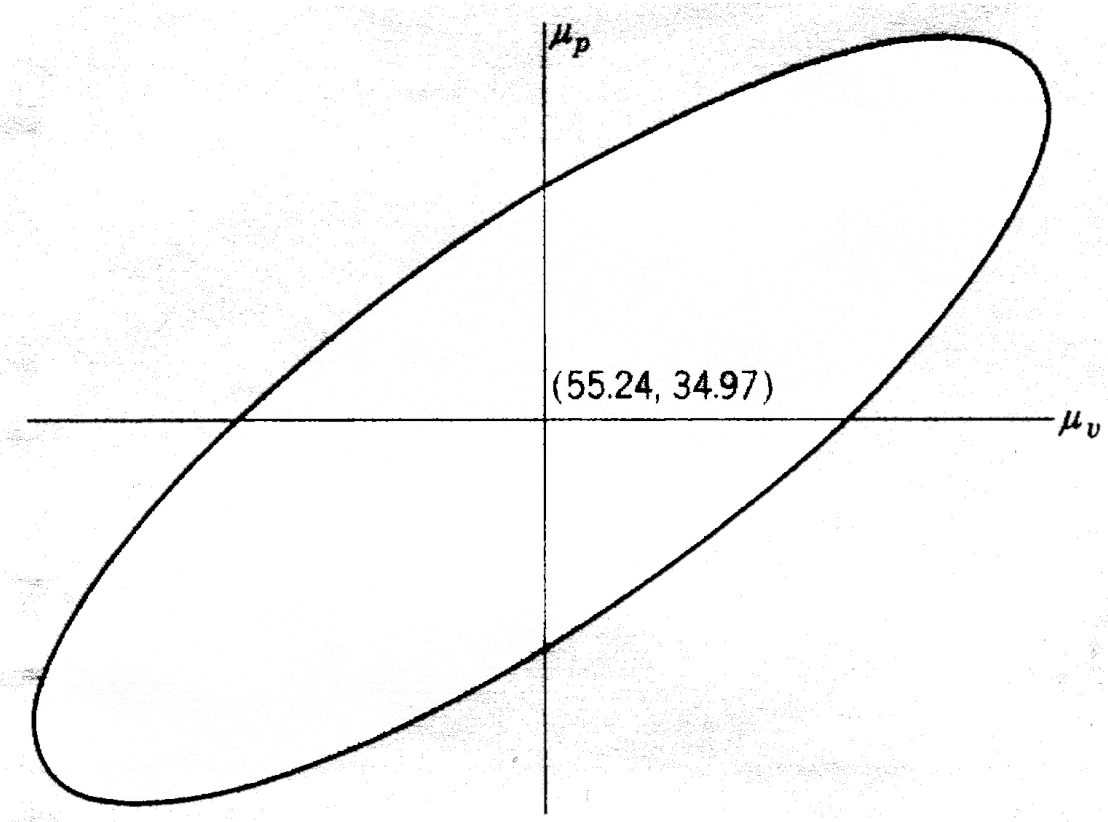


FIGURE 2.1

Multivariate One-sample T^2 Test

A 99% simultaneous CI for μ_v

$\mathbf{a}' = (1\ 0)$ and

$$\sqrt{\frac{1}{101}(1\ 0)\begin{pmatrix} 210.54 & 126.99 \\ 126.99 & 119.68 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix}} = 1.444$$

$$\sqrt{\frac{(101\text{-}1)2}{101\text{-}2}4.83} = 3.12$$

$$50.73 \ \leq \ \mu_{\mathrm{v}} \ \leq \ 59.75$$

The 99% CI for μ_{v} does not contain 60.

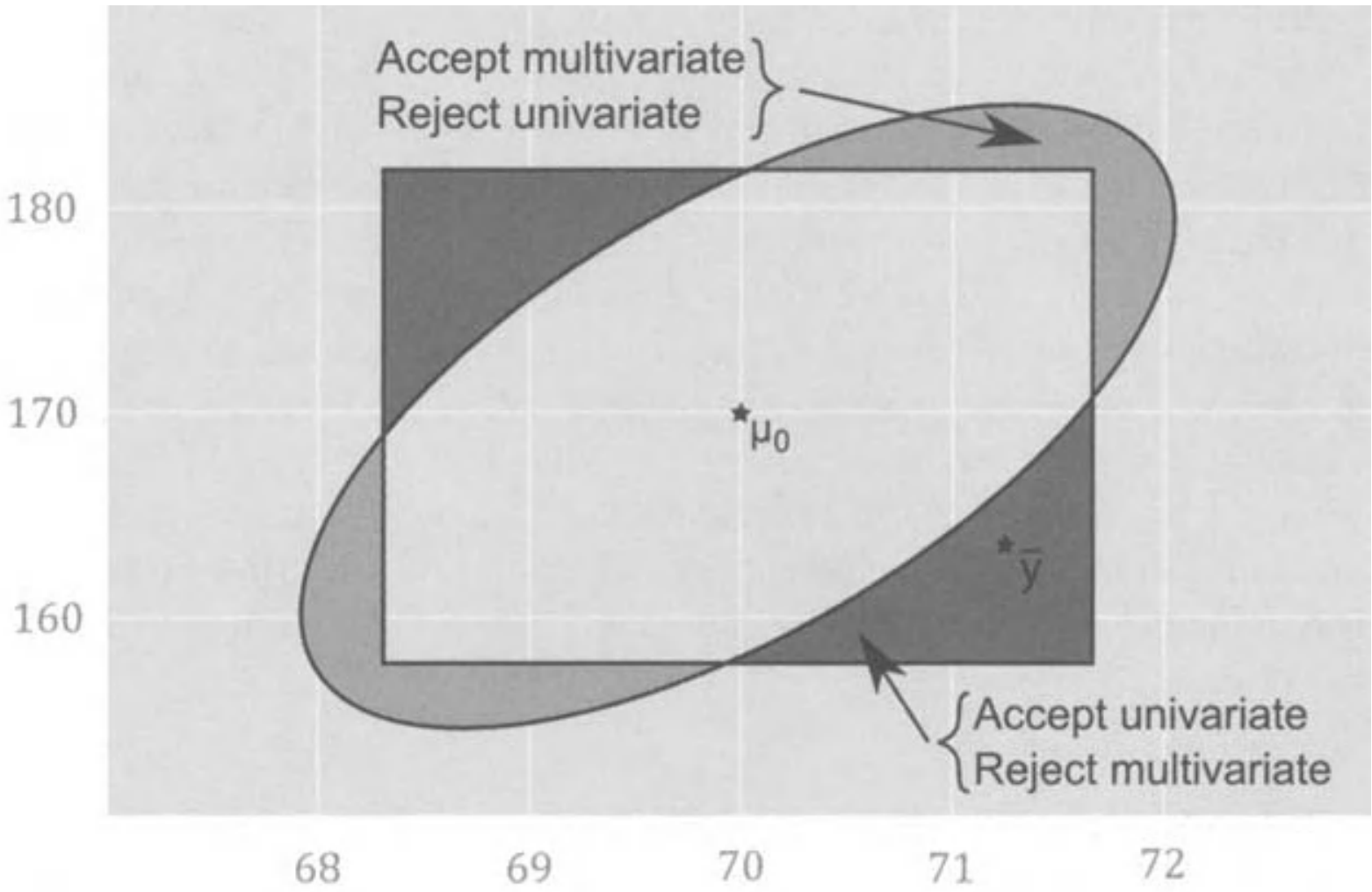
p univariate tests

- *P* univariate tests inflate the Type I error α.
- Ex: We do *p* = 10 separate univariate tests at the .05 level, the probability of at least one false rejection is greater than .05.
- If the variables were independent (they rarely are), we would have (under *Ho).*
-

$$P(\text{at least one rejection}) = 1 - P(\text{all 10 tests accept } \mathrm{H}_0)$$
$$= 1 - (0.95)^{10} = 0.40$$

- For correlated variables, the overall α~0.4-0.5.

Ex: a sample from bivariate normal distribution.



Testing the difference between two means

- t test for two means

t test for comparing two means (independent samples, variances not equal):

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(d.f. = the smaller of $n_1 - 1$ or $n_2 - 1$)

Equal variance case

Pooled variance: $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}$

d.f.=n_1+n_2-2

Formula for the confidence interval for difference of two means (small independent samples, variance unequal):

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(d.f. = smaller of $n_1 - 1$ and $n_2 - 1$)

Testing the difference between two means

- t test for comparing two means for dependent samples

t test for comparing two means for dependent samples:

$$t = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \qquad \text{where} \qquad \bar{D} = \frac{\Sigma D}{n} \qquad \text{and}$$

$$s_D = \sqrt{\frac{n\Sigma D^2 - (\Sigma D)^2}{n(n-1)}} \qquad (\text{d.f.} = n - 1)$$

Formula for confidence interval for the mean of the difference for dependent samples:

$$\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}$$

$(\text{d.f.} = n - 1)$

Example: Weights of Newborn Infants

- A researcher wishes to see if the average weights of newborn male infants are different from the average weights of newborn female infants. She selects a random sample of 10 male infants and finds the mean weight is 7 pounds 11 ounces and the standard deviation of the sample is 8 ounces. She selects a random sample of 8 female infants and finds that the mean weight is 7 pounds 4 ounces and the standard deviation of the sample is 5 ounces. Can it be concluded at $\alpha = 0.05$ that the mean weight of the males is different from the mean weight of the females? Assume that the variables are normally distributed and variances of both populations are equal.

- Step 1: State the hypotheses and identify the claim.
  - $H_0: \mu_1 = \mu_2$ and $H_1: \mu_1 \neq \mu_2$ (claim)

Example: Weights of Newborn Infants

- **Step 2: Find the critical values.**
  - Since the test is two-tailed and α = 0.05, the degrees of freedom are the smaller of n1 – 1 or n2 – 1.
  - In this case, n1 – 1 = 10 – 1 = 9 and n2 – 1 = 8 – 1 = 7.
  - From Table, the critical values are +2.365 and -2.365.
- **Step 3: Find the Test Value.**
  - Change the means to ounces. (1 lb = 16 oz)

$$7 \text{ lb } 11 \text{ oz} = 7 \times 16 + 11 = 123 \text{ oz}$$

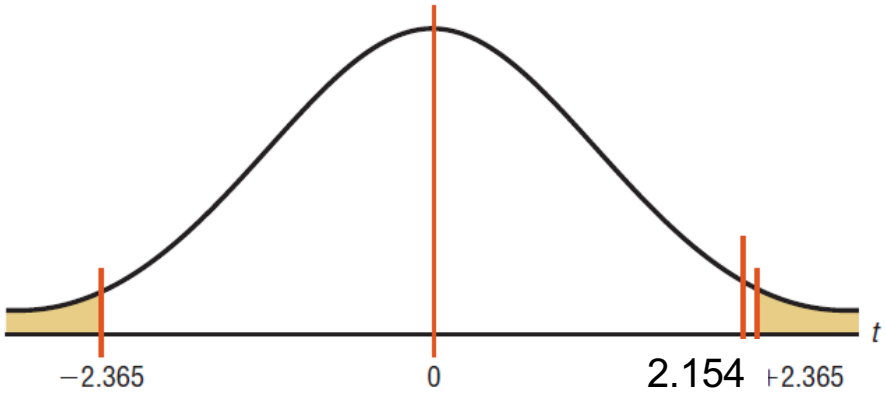
$$7 \text{ lb } 4 \text{ oz} = 7 \times 16 + 4 = 116 \text{ oz}$$

$$s_{sp}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} = \frac{(10-1)8^2 + (8-1)5^2}{(10+8-2)} = 46.938$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_{sp}^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{(123-116)-0}{\sqrt{46.938(\frac{1}{10} + \frac{1}{8})}} = \frac{7}{3.250} = 2.154$$

Example: Weights of Newborn Infants

- Step 4: Make the decision.
  - Do not reject the null hypothesis.



- Step 5: Summarize the results.
  - There is not enough evidence to support the claim that the mean of the weights of the male infants is different from the mean of the weights of the female infants.

Multivariate Two-sample T2 Test

- Multivariate Hypothesis of Two-sample Means

$$H_0 : \boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ . \\ . \\ \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ . \\ . \\ \mu_{2p} \end{pmatrix} = \boldsymbol{\mu}_2$$

vs.

$$H_a : \boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ . \\ . \\ \mu_{1p} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} \\ \mu_{22} \\ . \\ . \\ \mu_{2p} \end{pmatrix} = \boldsymbol{\mu}_2$$

Multivariate Two-sample T2 Test

- Data Structure for Two Independent Samples

Statistics	Sample 1	Sample 2
Sample size	n1	n2
Mean	$\bar{x}_1$	$\bar{x}_2$
Covariance matrix	S_1	S_2

- Two-sample Hotelling T2

$$t^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

where **S** is pooled covariance matrix

- Reject the null hypothesis at the α significance level if

Transformation of the Hotelling T^2 statistic

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

Multivariate Two-sample T^2 Test

- Define the difference $\delta = \mu_1 - \mu_2$

  A 100(1-α)% ellipsoidal confidence region for δ

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \delta) \le \frac{(n_1+n_2)(n_1+n_2-2)p}{(n_1 n_2)(n_1+n_2-p-1)} F_{\alpha,p,n_1+n_2-p-1}$$

Multivariate Two-sample T^2 Test

- The 100(1-α)% simultanous confidence intervals for all linear combinations $\mathbf{a}'\delta$ of mean differences

$$\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm \sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}\,\frac{(n_1+n_2)(n_1+n_2-2)p}{(n_1 n_2)(n_1+n_2-p-1)}\,F_{\alpha,p,n_1+n_2-p-1}}$$

Univariate Two-sample t Test

- Example: The total length of female sparrows between survivors and non-survivors

Statistics	Survivors	Non-surviviors
Sample size	21	28
Mean	157.38	158.43
Variance	11.05	15.07

Univariate Two-sample t Test

- Example: The total length of female sparrows between survivors and non-survivors

$$\text{Pooled variance: } s^2 = \frac{(21-1)\times 11.05 + (28-1)\times 15.07}{(21+28-2)} = 13.6$$

$$|t| = \left| \frac{157.38 - 158.43}{\sqrt{13.6(1/21 + 1/28)}} \right| = 0.99 < t_{0.025,47} = 2.015$$

Fail to reject the null hypothesis

at the 5% significance level.

Multivariate Two-sample T2 Test

- The mean vector and covariance matrix of survivors (n1 = 21)

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} 157.381 \\ 241.000 \\ 31.433 \\ 18.500 \\ 20.810 \end{pmatrix} \quad \mathbf{S}_1 = \begin{pmatrix} 11.048 & 9.100 & 1.557 & 0.870 & 1.286 \\ 9.100 & 17.500 & 1.910 & 1.310 & 0.880 \\ 1.557 & 1.910 & 0.531 & 0.189 & 0.240 \\ 0.870 & 1.310 & 0.189 & 0.176 & 0.133 \\ 1.286 & 0.880 & 0.240 & 0.133 & 0.575 \end{pmatrix}$$

- The mean vector and covariance matrix of non-survivors (n2 = 28)

$$\bar{\mathbf{x}}_2 = \begin{pmatrix} 158.429 \\ 241.571 \\ 31.479 \\ 18.446 \\ 20.839 \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} 15.069 & 17.190 & 2.243 & 1.746 & 2.931 \\ 17.190 & 32.550 & 3.398 & 2.950 & 4.066 \\ 2.243 & 3.398 & 0.728 & 0.470 & 0.559 \\ 1.746 & 2.950 & 0.470 & 0.434 & 0.506 \\ 2.931 & 4.066 & 0.559 & 0.506 & 1.321 \end{pmatrix}$$

Multivariate Two-sample T2 Test

- The pooled variance df = 21+28-2=47

$$\mathbf{S} = \frac{(20\mathbf{S}_1 + 27\mathbf{S}_2)}{47}$$

The Hotelling T2 is 2.824

n1=21, n2=28, p=5, df1=p=5,

df2=(n1+n2-p-1)=(21+28-5-1)=43

$$F = \frac{[(21+28-5-1)\times 2.824]}{[(21+28-2)5]} = 0.517 < F_{0.05,5,43} = 2.43$$

Fail to reject the null hypothesis at the 5% significance level. There is no evidence that the mean morphological characteristics are different between the survivors and non-survivors

Are male and female body shapes significantly different?

```
n1<-10; n2<-10; p<-3;
chest<-
c(34,37,38,36,38,43,40,38,40,41,36,36,34,33,36,37,34,36,38,35);

waist<-
c(30,32,30,33,29,32,33,30,30,32,24,25,24,22,26,26,25,26,28,23);

hips<-
c(32,37,36,39,33,38,42,40,37,39,35,37,37,34,38,37,38,37,40,35);
gender<-rep(c("male","female"),each=10)
measure<-data.frame(chest,waist,hips,gender)
cov_all<-cov(measure[, c("chest", "waist", "hips")])
covF<-cov(subset(measure, gender == "female")[,  c("chest",
"waist", "hips")])
covM<-cov(subset(measure, gender == "male")[,  c("chest",
"waist", "hips")])
#where the subset() returns all observations corresponding to
females (first statement) # or males (second statement).# pooled
covariance matrix for T2 calculation
```

Are male and female body shapes significantly different?

```
cov_pooled<- (covF*(n1-1)+covM*(n2-1))/(n1+n2-2);
invCp<-solve(cov_pooled);
invC<-solve(cov_all); # This will be the F value of 5.04, which is
not correct.

meanF<-colMeans(subset(measure, gender == "female")[,1:3]);
meanM<-colMeans(subset(measure, gender == "male")[,1:3]);
T2<-(n1*n2)/(n1+n2)*(meanF-meanM)%*%invCp%*%t(t(meanF-meanM));
F_calc<-(n1+n2-p-1)*T2/(n1+n2-2)/p; F_calc
F_crit<-qf(0.95,p,n1+n2-p-1); F_crit
```

```
     [,1]
[1,] 45.93111
```

```
3.238872
```

```
#install.packages("ICSNP")
library("ICSNP")
m1 <- with(measure, HotellingsT2(cbind(chest,waist,hips) ~
gender));m1
```

```
Hotelling's two sample T2-test

data:  cbind(chest, waist, hips) by gender
T.2 = 45.931, df1 = 3, df2 = 16, p-value = 4.364e-08
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

Multivariate paired two-sample T2 test

- The Paired T^2 Test
  - Take the difference in the corresponding vectors of measurements within each object between the two conditions
  - Obtain the sample mean difference vector and sample covariance matrix
  - Apply one-sample T2 test

Multivariate paired two-sample T2 test

- The Paired T^2 Test Example
  - Dataset: Measurement of uraic acid (X1) and total cholesterol level (X2) before and after the treatment for six patients

	Before		After		Difference	
Pat. No.	X11	x21	X12	X22	d1	d2
1	12.5	220	6.5	190	-6.0	-30
2	14.2	260	7.8	250	-6.4	-10
3	10.8	180	6.0	190	-4.8	10
4	13.4	200	7.2	220	-6.2	20
5	11.9	280	6.8	240	-5.1	-40
6	12.0	170	5.9	180	-6.1	10

Multivariate paired two-sample T2 test

- The sample mean difference vector is

$$\bar{d} = \begin{pmatrix} -5.7767 \\ -6.6667 \end{pmatrix}$$

$$\boxed{F = \frac{n-p}{p(n-1)} T^2 = \frac{(n-p)n}{p(n-1)} (\vec{\bar{x}} - \vec{\mu}_0)' \mathbf{S}^{-1} (\vec{\bar{x}} - \vec{\mu}_0) > F_{\alpha,p,n-p}}$$

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

- The sample covariance matrix based on the differences

$$\mathbf{S}_d = \begin{pmatrix} 0.4267 & -3.1333 \\ -3.1333 & 586.6667 \end{pmatrix} \qquad \mathbf{S}_d^{-1} = \begin{pmatrix} 2.4394 & 0.01303 \\ 0.01303 & 0.00174 \end{pmatrix}$$

$$T^2 = 6(-5.7667 - 6.6667) \begin{pmatrix} 2.4394 & 0.01303 \\ 0.01303 & 0.001774 \end{pmatrix} \begin{pmatrix} -5.7767 \\ -6.6667 \end{pmatrix} = 493.2151$$

$$F = \frac{6-2}{2} \frac{493.2151}{6-1} = 197.2860 > F_{0.05,2,4} = 6.944$$

Reject null hypothesis

Example

- Consider an example in which two classrooms of students are asked to rate their teacher in terms of their satisfaction and the level of teacher knowledge.

Teacher	Satisfaction	Knowledge
1	1	3
1	3	7
1	2	2
2	4	6
2	6	8
2	6	8
2	5	10
2	5	10
2	4	6

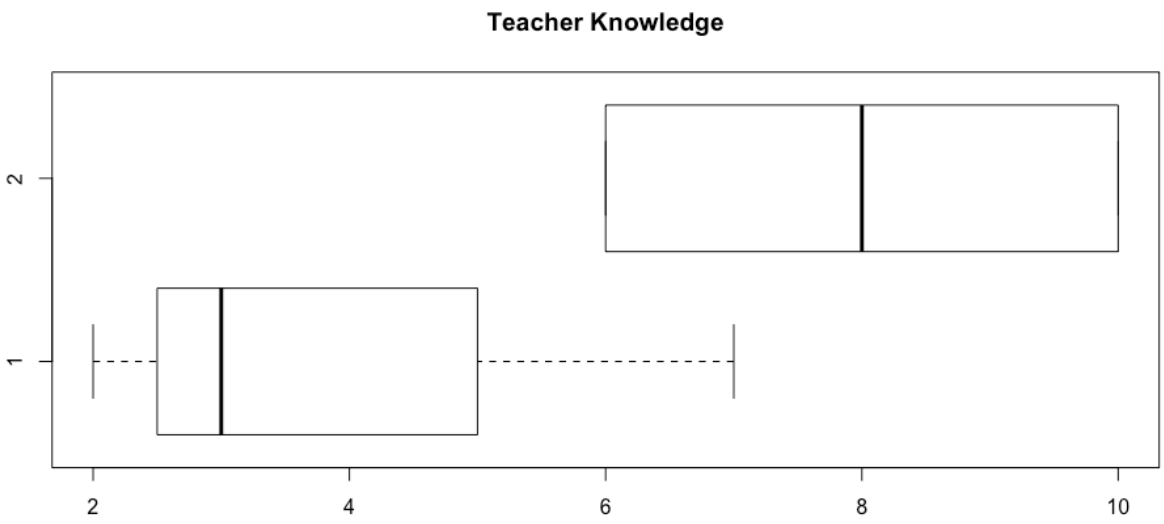
Running Hotelling T2 in R

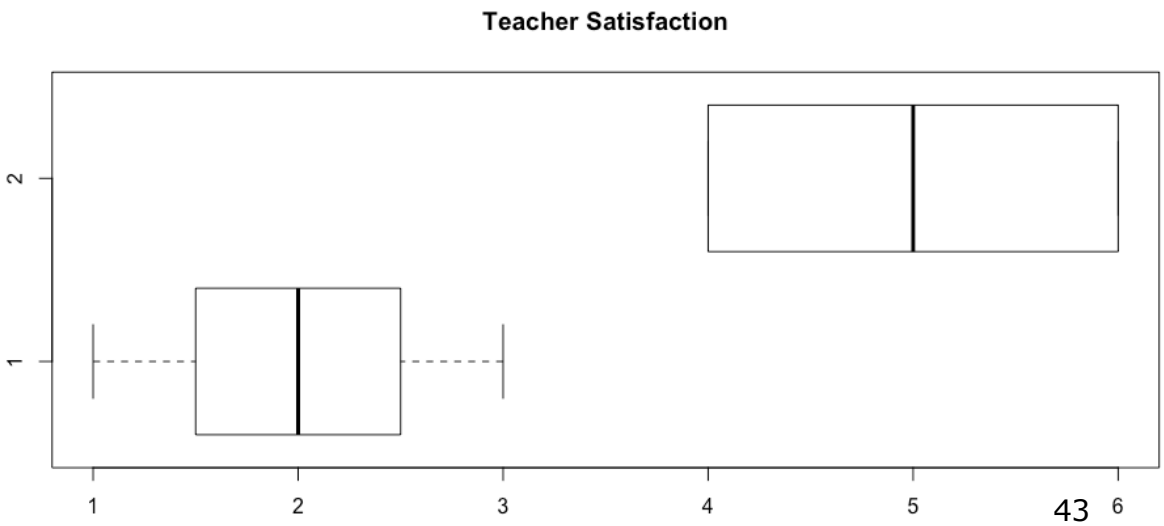
```
# Multivariate 2 sample T2 test
install.packages("ICSNP")
library("ICSNP")
math.teach <- data.frame(teacher = factor(rep(1:2,c(3,
6))), satis = c(1, 3, 2, 4, 6, 6, 5,5, 4),know = c(3,
7, 2, 6, 8, 8, 10, 10, 6));
with(math.teach, tapply(know, teacher, mean));
with(math.teach, tapply(satis, teacher, mean));
par(mfrow = c(2, 1));
boxplot(know ~ teacher, math.teach, main = "Teacher
Knowledge", horizontal = T);
boxplot(satis ~ teacher, math.teach, main = "Teacher
Satisfaction",horizontal = T);

m1 <- with(math.teach, HotellingsT2(cbind(satis, know)
~ teacher));
```

Graphical Display of Data

Hotelling's two sample T2-test

data: cbind(satis, know) by teacher
T.2 = 9, df1 = 2, df2 = 6, p-value = 0.01562
alternative hypothesis: true location difference is not equal to c(0,0)





Multivariate T2 test

- The vector ${}_pd_1$ is Gaussian multivariate-distributed with zero mean and unit covariance matrix $N({}_p0_1, {}_pI_p)$ and ${}_pM_p$ is a *pxp* matrix with unit scale matrix and *m* degrees of freedom.
- The Quadratic form $X=({}_1d^T{}_p\ M^{-1}{}_pd_1)$ has a Hotelling T2(p,m) distribution $X \sim T^2_{p,m}$

Transformation of the Hotelling T^2 statistic

$$\frac{m-p+1}{pm} X \sim F_{p,m-p+1}$$

$$t^2 = n\left(\vec{\bar{x}} - \vec{\mu}_0\right)' \mathbf{S}^{-1} \left(\vec{\bar{x}} - \vec{\mu}_0\right) > t^2_{\alpha/2}$$

Summary of multivariate T2 test

$$X \sim T^2_{p,m} \qquad \frac{m-p+1}{pm} X \sim F_{p,m-p+1}$$

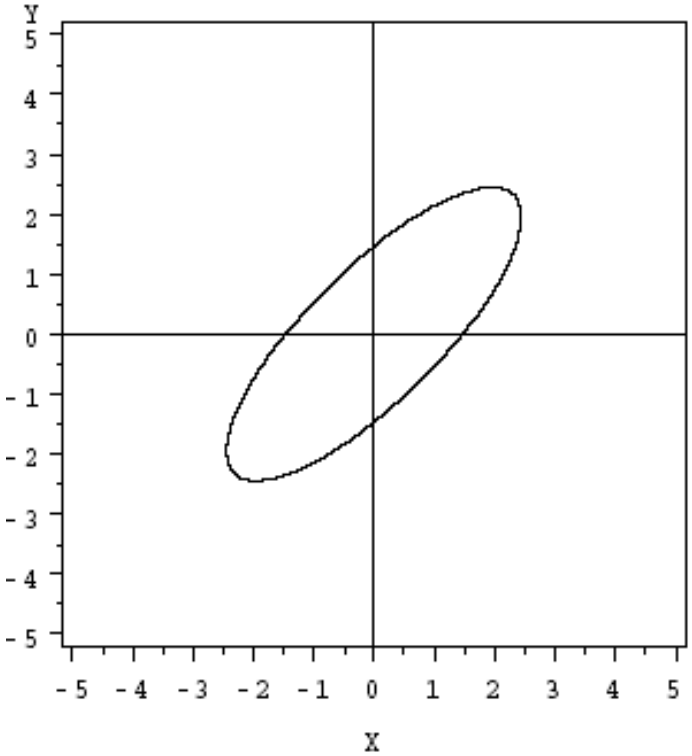
Type	df	F statistic
1-sample	n-1	$F = \frac{n-p}{p(n-1)} T^2 > F_{\alpha,p,n-p}$
2-sample	n1+n2-2	$F = \frac{(n_1+n_2-p-1)}{p(n_1+n_2-2)} T^2 > F_{\alpha,p,n_1+n_2-p-1}$
Paired 2 sample	n-1	$F = \frac{n-p}{p(n-1)} T^2 > F_{\alpha,p,n-p}$

Multivariate normal distribution

- The Mahalanobis distance $d^2=(x-\mu)'\Sigma^{-1}(x-\mu)$ has a chi-square distribution with *p* degrees of freedom, and for "large" samples the observed Mahalanobis distances have an approximate chi-square distribution.

- A Q-Q plot against chi-square quantiles can be used to examine.

- The probability of a random **X** falling inside the hyper-ellipse of Mahalanobis distance is
  - $\Pr\{(x-\mu)'\Sigma^{-1}(x-\mu)\leq\chi^2_{p,\alpha}\}=1-\alpha$



Determining the Quantiles

- The ith estimated quantile is determined as the chi-square value (with d.f.) for which the cumulative probability is (i-0.5) / n.

- To determine the full set of estimated chi-square quantiles, this is done for value of i from 1 to n.

Chi square distribution

- The Chi Square distribution is the distribution of the sum of squared standard normal deviates.
- Standard normal distribution

$$Z = \frac{X - \mu}{\sigma} \qquad Z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

Standardized variables

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} = (\Sigma^{1/2})^{-1}(\mathbf{x} - \mu)$$

- $\sum_{i=1}^{p} Z_i^2 = \mathbf{Z}'\mathbf{Z}$ has the χ^2 distribution with *p* degrees of freedom.

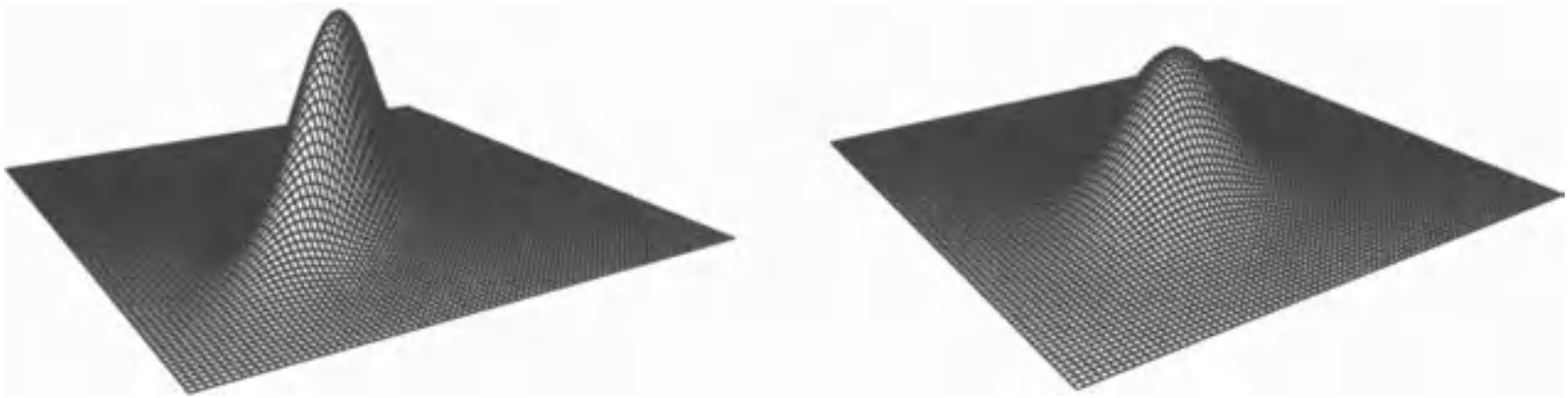


Figure 4.2 Bivariate normal densities.

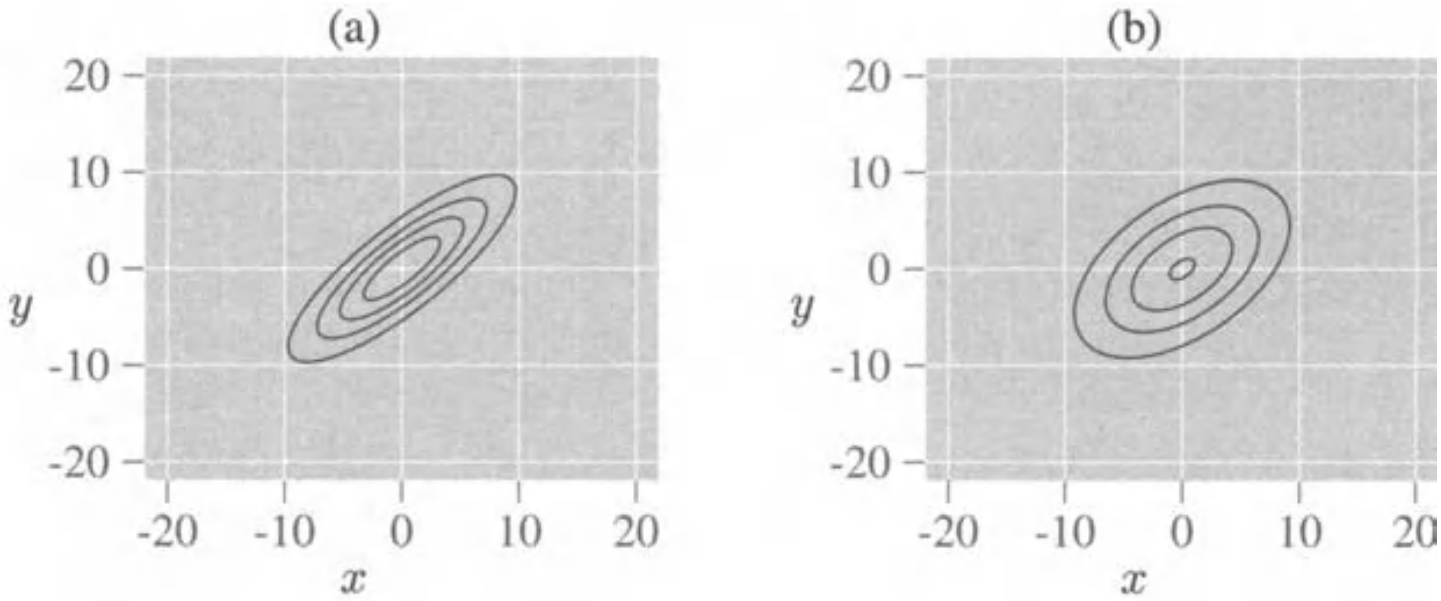


Figure 4.3 Contour plots for the distributions in Figure 4.2.

- Univariate Normal Density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty \qquad y \sim N(\mu, \sigma^2)$$

- Multivariate Normal Density

$$g(x) = \frac{1}{(\sqrt{2\pi})^p \, |\Sigma|^{1/2}} e^{-(x-\mu)'\Sigma^{-1}(x-\mu)/2}, \quad -\infty < x < \infty \qquad \mathbf{y} \sim N_p(\mu, \Sigma)$$

$$\mathbf{Z}'\mathbf{Z} = (\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) \sim \chi_p^2$$

Multivariate Normality

- Calculate squared Mahalanobis distance (D^2) of each observed vector of *p* variables from the sample mean vector.

- Order the MD^2 from the smallest to the largest, $D^2_{(1)} < D^2_{(2)} < \ldots < D^2_{(n)}$

- For each ordered D^2, compute the (i-0.5)/n percentile where i is the ith order observed vector.

Multivariate Normality

- The χ^2 values for the percentiles are obtained from the χ^2 with *p* d.f., which can be computed by CINV function in SAS.
- Plot D^2 vs. χ^2 (similar to Q-Q plot).
- The plot should be linear.
- Deviation from linearity indicates non-normality.

Multivariate Normality

- Several tests available, see Seber (1984). However, methods are *ad-hoc* and are not implemented in most of statistical software.
- Transformations

Origial scale	Transformed scale
Counts, x	$\sqrt{x}$
Proportions, $\hat{p}$	$\log\mathrm{it}(\hat{p})=\frac{1}{2}\log(\frac{\hat{p}}{1-\hat{p}})$
Correlations, *r*	Fisher's $z(r)=\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$
Right-skewed	log(X), $\sqrt{x}$
Left-skewed	X^2, X^3, X^4

OUTLIER DETECTION

Outlier Detection

- Univariate outliers
  - For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.
  - For larger sample sizes, increase the threshold value of standard scores up to 4.
  - If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.
- Bivariate outliers
  - Focus their use on specific variable relationships, such as the independent versus dependent variables.
  - Use scatterplots with confidence intervals at a specified alpha level.
- Multivariate outliers
  - Threshold levels for the D^2/df measure should be conservative (.005 or .001), resulting in values of 2.5 (small samples) versus 3 or 4 in larger samples.

Detect outliers in HBAT data

- Univariate, outliers for X6~X19.
- Bivariate outliers for X6~X18 vs. X19.
- Multivariate outliers for X6~X19.

TABLE 3 Description of Database Variables

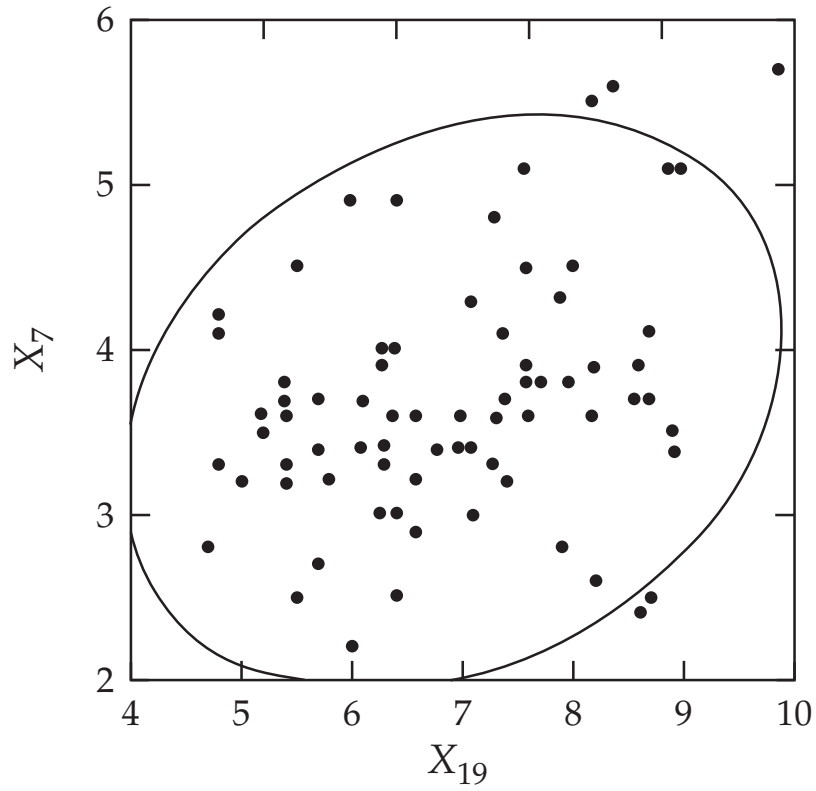
Variable Description	Variable Type
Data Warehouse Classification Variables	
X_1 Customer Type	Nonmetric
X_2 Industry Type	Nonmetric
X_3 Firm Size	Nonmetric
X_4 Region	Nonmetric
X_5 Distribution System	Nonmetric
Performance Perceptions Variables	
X_6 Product Quality	Metric
X_7 E-Commerce Activities/Web Site	Metric
X_8 Technical Support	Metric
X_9 Complaint Resolution	Metric
X_{10} Advertising	Metric
X_{11} Product Line	Metric
X_{12} Salesforce Image	Metric
X_{13} Competitive Pricing	Metric
X_{14} Warranty and Claims	Metric
X_{15} New Products	Metric
X_{16} Ordering and Billing	Metric
X_{17} Price Flexibility	Metric
X_{18} Delivery Speed	Metric
Outcome/Relationship Measures	
X_{19} Satisfaction	Metric
X_{20} Likelihood of Recommendation	Metric
X_{21} Likelihood of Future Purchase	Metric
X_{22} Current Purchase/Usage Level	Metric
X_{23} Consider Strategic Alliance/Partnership in Future	Nonmetric

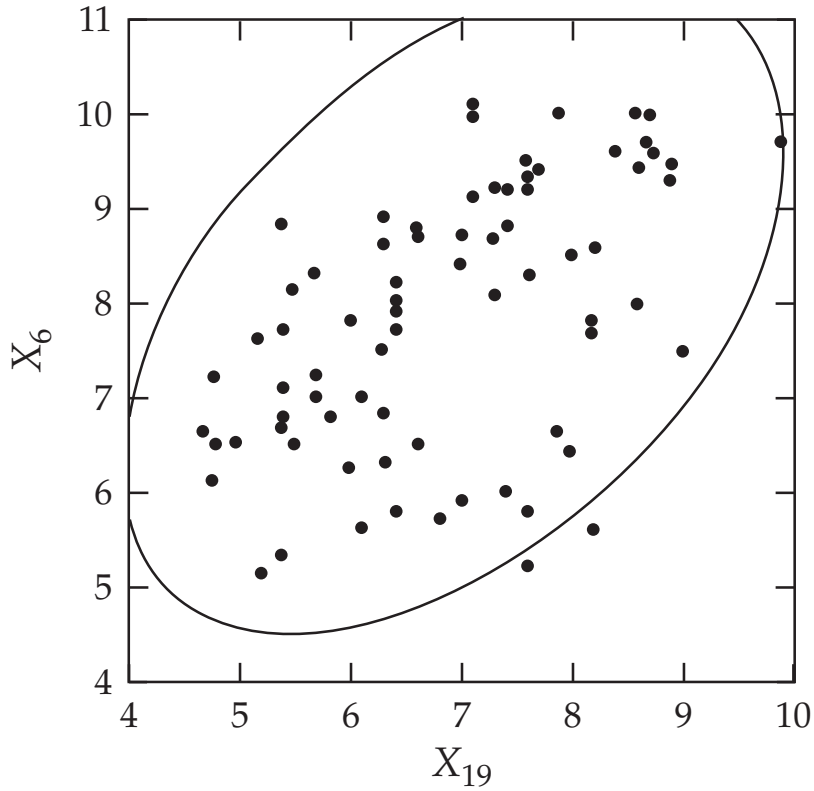
```
HBAT<-read.csv("HBAT.csv",
header=T, sep=",")
x7<-scale(HBAT$X7);
x8<-scale(HBAT$X8);
x19<-scale(HBAT$X19);
#abs(x7>=2.5);
#abs(x8>=2.5);

library(car)
dataEllipse(HBAT$X19,HBAT$X6,
levels=c(0.025,0.975))
dataEllipse(HBAT$X19,HBAT$X7,
levels=c(0.025,0.95))

X<-HBAT[,7:19];  # Read X6 to X18
to X matrix
p<-ncol(X);
Sx <- cov(X)
D2 <- mahalanobis(X, colMeans(X),
Sx);D2
par(mfrow=c(1,1))
chisplot(as.matrix(X))
D2/p; # >2.5 are multivariate
outliers
```

```
#### Chi-Square plot for
chisplot <- function(x) {
  if (!is.matrix(x)) stop("x is
not a matrix")

  ### determine dimensions
  n <- nrow(x)
  p <- ncol(x)
  #
  xbar <- apply(x, 2, mean)
  S <- var(x)
  S <- solve(S)
  index <- (1:n-0.5)/(n)
  #
  xcent <- t(t(x) - xbar)
  di <- apply(xcent, 1,
function(x,S) x %*% S %*% x,S)
  #
  quant <- qchisq(index,p)
  plot(quant, sort(di), ylab =
"Ordered distances",
       xlab = “Chi-square
quantile”, lwd=2,pch=1)
}
```

Scatterplots for Bivariate Detection of Outliers

X7 (E-Commerce Activities) vs. *X19* (Customer Satisfaction)





X6 (Product Quality) vs. *X19* (Customer Satisfaction)

TABLE 10 Univariate, Bivariate, and Multivariate Outlier Detection Results

UNIVARIATE OUTLIERS		BIVARIATE OUTLIERS		MULTIVARIATE OUTLIERS		
Cases with Standardized Values Exceeding ± 2.5		*Cases Outside the 95% Confidence Interval Ellipse*		*Cases with a Value of D^2/df Greater than 2.5 (df = 13)*[a]		
		X_{19} with:		**Case**	**D^2**	**D^2/df**
X_6	No cases	X_6	44, **90**	98	40.0	3.08
X_7	13, 22, **90**	X_7	13, **22**, 24, 53, **90**	36	36.9	2.84
X_8	8, **7**	X_8	**22**, 87			
X_9	No cases	X_9	**2, 22**, 45, 52			
X_{10}	No cases	X_{10}	**22, 24**, 85			
X_{11}	**7**	X_{11}	**2**, 7, **22**, 45			
X_{12}	**90**	X_{12}	**22**, 44, **90**			
X_{13}	No cases	X_{13}	**22**, 57			
X_{14}	77	X_{14}	**22**, 77, 84			
X_{15}	6, 53	X_{15}	6, **22**, 53			
X_{16}	24	X_{16}	**22**, **24**, 48, 62, 92			
X_{17}	No cases	X_{17}	**22**			
X_{18}	**7**, 84	X_{18}	**2**, 7, **22**, 84			
X_{19}	**22**					

[a]*Mahalanobis D^2 value based on the 13 HBAT perceptions (X_6 to X_{18}).*

Multivariate Outliers

$$\mathbf{X_1},\ldots,\mathbf{X_n} \sim N_p(\mu_0,\Sigma)$$

$$\mathbf{X}_i: \text{ a suspected outlier} \sim N_p(\mu = \mu_0 + \Delta, \Sigma)$$

$$H_{0i}: \mathbf{X}_i \sim N_p(\mu_0,\Sigma)$$

$$H_{ai}: \mathbf{X}_i \sim N_p(\mu_0 + \Delta,\Sigma)$$

$$\Downarrow$$

$$H_{0i}: \mu = \mu_0 (\Delta = \mathbf{0})$$

$$H_{ai}: \mu \neq \mu_0 (\Delta \neq \mathbf{0}), \quad \text{for } i = 1 \cdots n$$

Multivariate Outliers

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$n_1 = 1, n_2 = (n-1)$$

- Compare a sample of *n*-1 normal random vectors with a sample of 1 normal vector

$$T_i^2 = \frac{n-1}{n}[\mathbf{X}_i - \overline{\mathbf{X}}_{(-i)}]' \mathbf{S}_{(-i)}^{-1}[\mathbf{X}_i - \overline{\mathbf{X}}_{(-i)}]$$

$$T_i^2 = \frac{(n-2)D_i^2}{[\frac{n-1}{n} - D_i^2]}$$

where D_i^2 is the Mahalanobis distance between $\mathbf{X}_i$ and $\overline{\mathbf{X}}$, $D_i^2 = [\mathbf{X}_i - \overline{\mathbf{X}}]' \mathbf{S}^{-1}[\mathbf{X}_i - \overline{\mathbf{X}}]$, $i = 1,\ldots,n$

Wilks' statistic

- Detection of a single outlier.

$$w = \min_i \frac{\left|(n-2)\mathbf{S}_{-i}\right|}{\left|(n-1)\mathbf{S}\right|}$$

- The statistic w can be expressed in terms of $D^2_{(n)}$, which is the maximum of D^2_i

$$w = 1 - \frac{nD^2_{(n)}}{(n-1)^2},$$

$$D^2_i = [\mathbf{X}_i - \overline{\mathbf{X}}]' \mathbf{S}^{-1} [\mathbf{X}_i - \overline{\mathbf{X}}],\ i = 1,\ldots,n$$

F test of w

$$F_i = \frac{n-p-1}{p}\left[\frac{1}{1-nD_i^2/(n-1)^2}-1\right], \quad i=1,2,\cdots,n.$$

- F_i's are independently and identically distributed as $F_{p,n-p-1}$, and a test can be constructed in terms of $\max_i F_i$.

$$P(\max_i F_i > f) = 1 - P(\text{all } F_i \le f) = 1 - [P(F \le f)]^n.$$

- The test can be carried out using F-table.

$$\max_i F_i = F_{(n)} = \frac{n-p-1}{p}\left(\frac{1}{w}-1\right).$$

Example 4.6.2

- We use the ramus bone data set of Table 3.7 to illustrate a search for multivariate outliers, while at the same time checking for multivariate normality. An examination of each column of Table 3.7 does not reveal any apparent univariate outliers. To check for multivariate outliers, we first calculate Di^2 in (4.27) for each observation vector. The results are given in Table 4.1.

Table 3.7 Ramus Bone Length at Four Ages for 20 Boys

	Age			
Individual	8 yr (y_1)	$8\frac{1}{2}$ yr (y_2)	9 yr (y_3)	$9\frac{1}{2}$ yr (y_4)
1	47.8	48.8	49.0	49.7
2	46.4	47.3	47.7	48.4
3	46.3	46.8	47.8	48.5
4	45.1	45.3	46.1	47.2
5	47.6	48.5	48.9	49.3
6	52.5	53.2	53.3	53.7
7	51.2	53.0	54.3	54.5
8	49.8	50.0	50.3	52.7
9	48.1	50.8	52.3	54.4
10	45.0	47.0	47.3	48.3
11	51.2	51.4	51.6	51.9
12	48.5	49.2	53.0	55.5
13	52.1	52.8	53.7	55.0
14	48.2	48.9	49.3	49.8
15	49.6	50.4	51.2	51.8
16	50.7	51.7	52.7	53.3
17	47.2	47.7	48.4	49.5
18	53.3	54.6	55.1	55.3
19	46.2	47.5	48.1	48.4
20	46.3	47.6	51.3	51.8

Table A.6 Upper Percentiles for Test of Single Multivariate Normal Outlier

Upper percentage points for the test statistic

$$D^2_{(n)} = \max_{l \le i \le n} (\mathbf{y}_i - \overline{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \overline{\mathbf{y}}).$$

This tests for a single outlier in a sample of size n from a multivariate normal distribution. Reject and conclude that the outlier is significant if $D^2_{(n)}$ exceeds the table value.

	$p = 2$		$p = 3$		$p = 4$		$p = 5$	
n	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	3.17	3.19						
6	4.00	4.11	4.14	4.16				
7	4.71	4.95	5.01	5.10	5.12	5.14		
8	5.32	5.70	5.77	5.97	6.01	6.09	6.11	6.12
9	5.85	6.37	6.43	6.76	6.80	6.97	7.01	7.08
10	6.32	6.97	7.01	7.47	7.50	7.79	7.82	7.98
12	7.10	8.00	7.99	8.70	8.67	9.20	9.19	9.57
14	7.74	8.84	8.78	9.71	9.61	10.37	10.29	10.90
16	8.27	9.54	9.44	10.56	10.39	11.36	11.20	12.02
18	8.73	10.15	10.00	11.28	11.06	12.20	11.96	12.98
20	9.13	10.67	10.49	11.91	11.63	12.93	12.62	13.81

Table 4.1 Values of D_i^2 for the Ramus Bone Data in Table 3.7

Observation Number	D_i^2	Observation Number	D_i^2
1	0.7588	11	2.8301
2	1.2980	12	10.5718
3	1.7591	13	2.5941
4	3.8539	14	0.6594
5	0.8706	15	0.3246
6	2.8106	16	0.8321
7	4.2915	17	1.1083
8	7.9897	18	4.3633
9	11.0301	19	2.1088
10	5.3519	20	10.0931

Example

- Husband's age and height, wife's age and height, and husband height at first marriage.

Hage	Hheight	Wage	Wheight	Hagefm
49	1809	43	1590	25
25	1841	28	1560	19
40	1659	30	1620	38
52	1779	57	1540	26
58	1616	52	1420	30
32	1695	27	1660	23
43	1730	52	1610	33
47	1740	43	1580	26
31	1685	23	1610	26
26	1735	25	1590	23

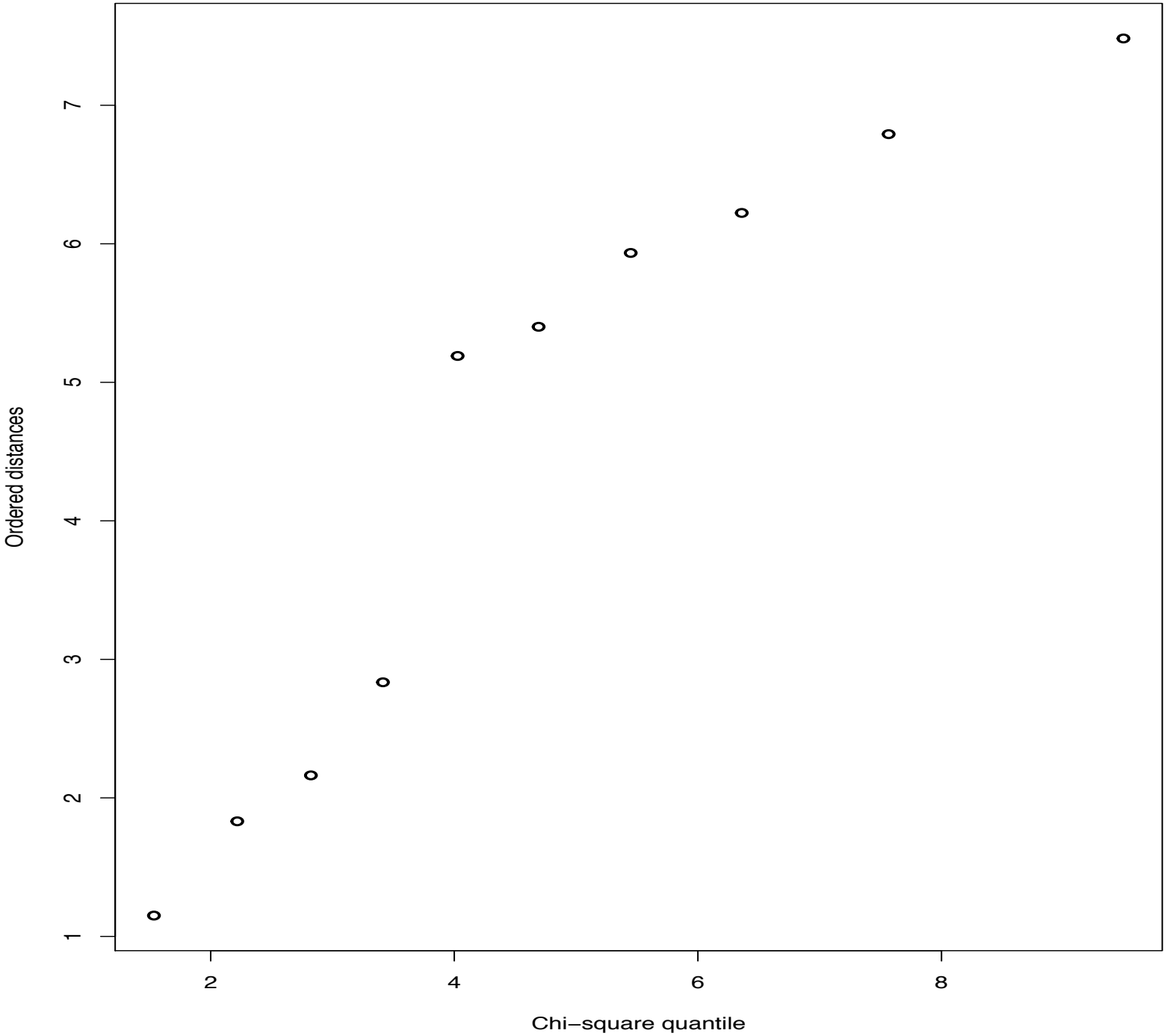
Plotting for Multivariate Normality

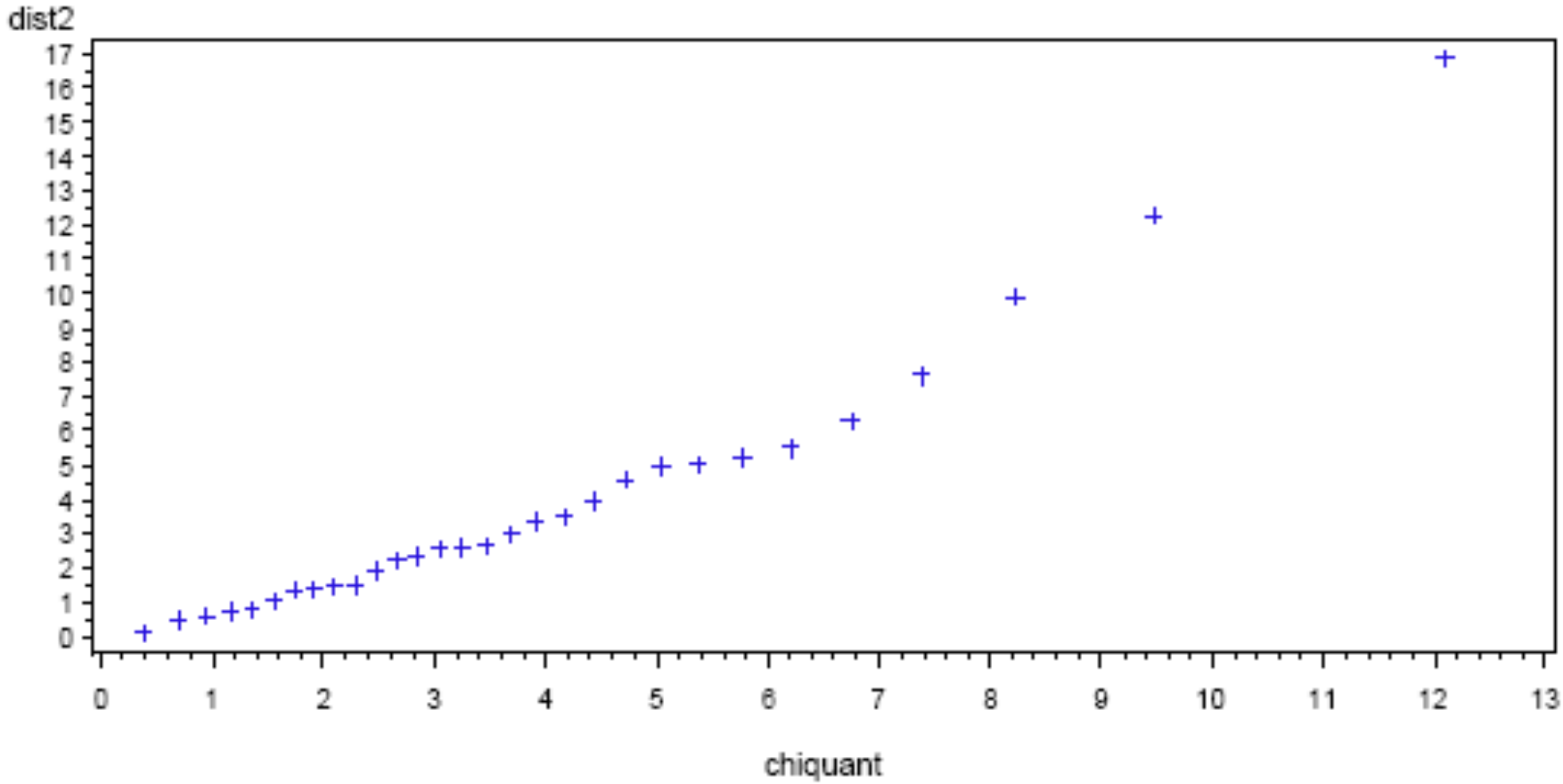
```
par(mfrow=c(1,5))
qqnorm(huswif$Hage, ylab="Ordered Observations Hage")
qqline(huswif[,1])
qqnorm(huswif$Hheight, ylab="Ordered Observations Hheight")
qqline(huswif[,2])
qqnorm(huswif$Wage, ylab="Ordered Observations Wage")
qqline(huswif[,3])
qqnorm(huswif$Wheight, ylab="Ordered Observations Wheight")
qqline(huswif[,4])
qqnorm(huswif$Hagefm, ylab="Ordered Observations Hagefm")
qqline(huswif[,5])
```

##Chi-Square plot

```
chisplot <- function(x) {
  if (!is.matrix(x))
stop("x is not a matrix")
 ### determine dimensions
  n <- nrow(x)
  p <- ncol(x)
  #
  xbar <- apply(x, 2,
mean)
  S <- var(x)
  S <- solve(S)
  index <- (1:n-0.5)/n
 xcent <- t(t(x) - xbar)
```

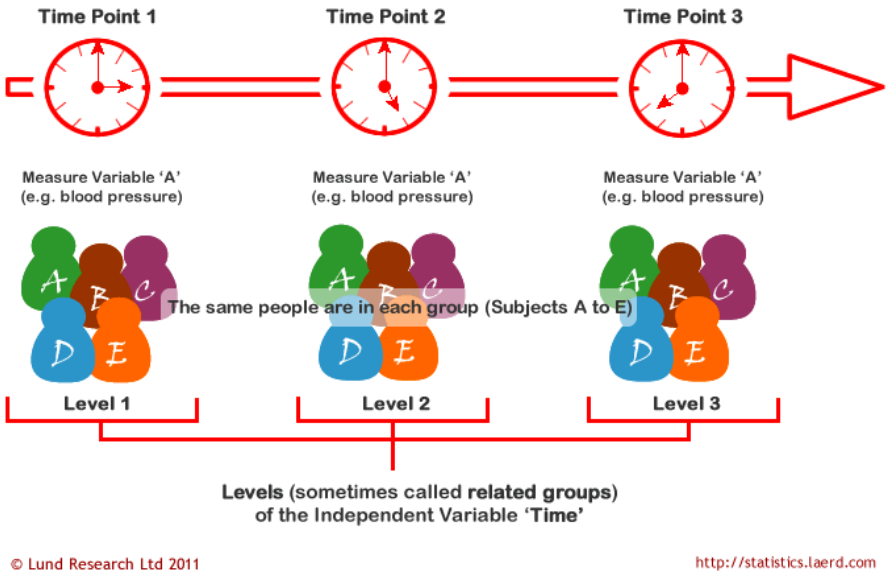
```
di <- apply(xcent, 1,
function(x,S) x %*% S %*%
x,S)

quant <- qchisq(index,p)
  plot(quant, sort(di),
ylab = “Ordered
distances”,
xlab = “Chi-square
quantile”, lwd=2,pch=1)
}
par(mfrow=c(1,1))
chisplot(as.matrix(huswif
))
```


Ordered distances
1
2
3
4
5
6
7
2
4
6
8
Chi−square quantile


dist2
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
chiquant

Profile Analysis

- Multivariate equivalent of repeated measures or mixed ANOVA.
  1. Comparing the same dependent variables between groups over several time-points.
  2. When there are several measures of the same dependent variable
     - Ex. several different psychological tests that all measure depression.



https://statistics.laerd.com/statistical-guides/repeated-measures-anova-statistical-guide.php

Profile Analysis

- Forty-nine elderly men were classified into "senile factor present" (SFP, n_1=12) and "no senile factor" (NSF, n_2=37). The Wechsler Adult Intelligence Scale (WAIS) was administered to all subjects. The WAIS consists of four domains: information, similarity, arithmetic and picture completion. The sample means are given below (Morrison, 2005):

	Group	
	NSF	SFP
Domain	n1 = 37	n2 = 12
information	12.57	8.75
similarity	9.57	5.33
arithmetic	11.49	8.50
picture completion	7.97	4.75

wais.csv

Profile Analysis

The pooled sample covariance matrix

$$\mathbf{C} = \begin{pmatrix} 11.2624 & 9.4060 & 7.1550 & 3.3791 \\ & 13.5265 & 7.3784 & 2.5014 \\ & & 11.5796 & 2.6167 \\ & & & 5.8133 \end{pmatrix}$$

Its inverse is

$$\mathbf{C}^{-1} = \begin{pmatrix} 0.259832 & -0.136477 & -0.058698 & -0.065886 \\ & 0.186734 & -0.038324 & 0.016230 \\ & & 0.150966 & -0.017344 \\ & & & 0.211142 \end{pmatrix}$$

Profile Analysis

- The two-sample Hotelling T^2 statistic is 22.13 $F = (12+37-4-1)(22.13)/[(12+37-2)4] = 5.179$
- $> F_{0.05,\,4,44} = 2.58$
- The 95% simultaneous CI

Domain	CI
Information	(0.12, 7.52)
Similarity	(0.18, 8.30)
Arithmetic	(-0.77, 6.75)
Picture Completion	(0.56, 5.88)

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

Profile Analysis

- Assumptions:
  - The same battery of psychological tests.
  - Measurements of continuous random variables.
  - The responses are commensurable or in comparable units.

Profile Analysis

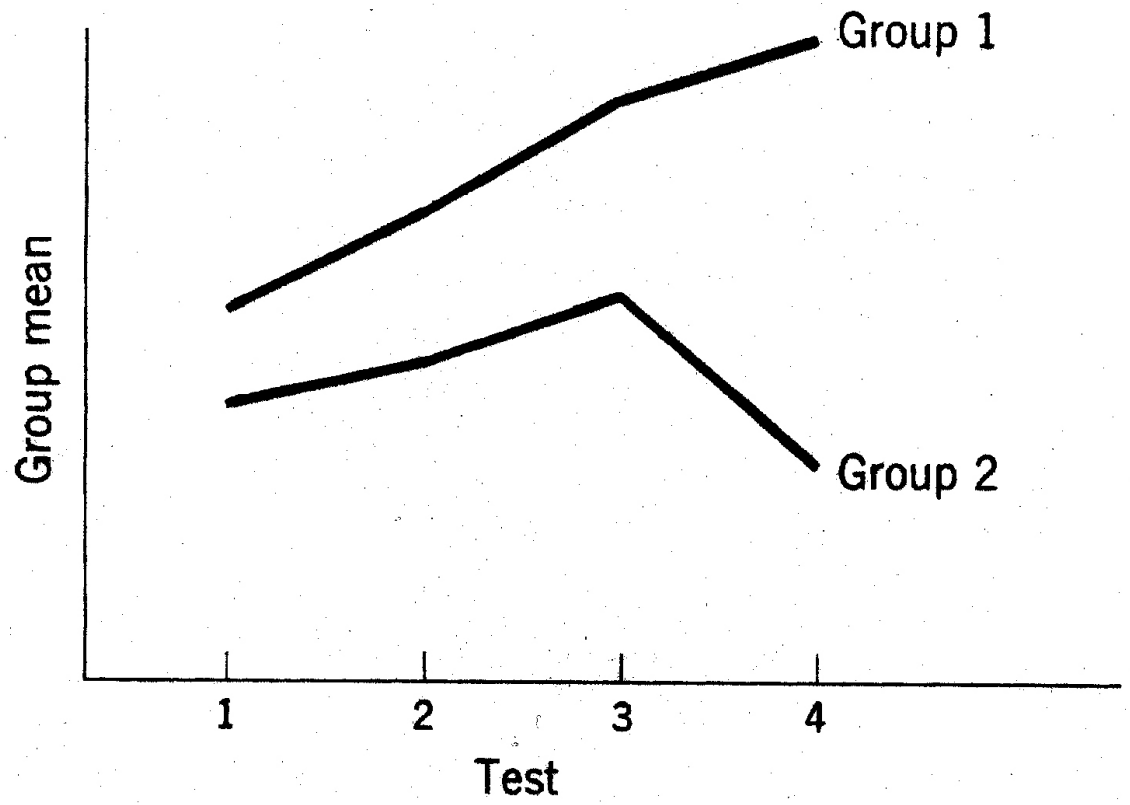


FIGURE 2.2

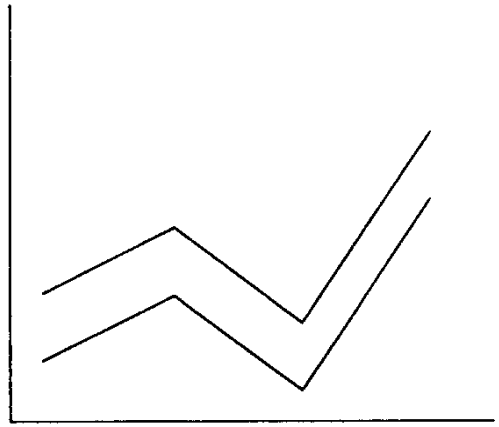
Profile Analysis

- Three questions (in terms of priority):
  - Are the population mean profiles similar, in the sense that the line segments of adjacent tests are parallel? (Test for parallelism: response-by-group interaction)
  - If the two population profiles are indeed parallel, are they also at the same level? (equal group effects)
  - Again assuming parallelism, are the population means of the tests different?

Profile Analysis

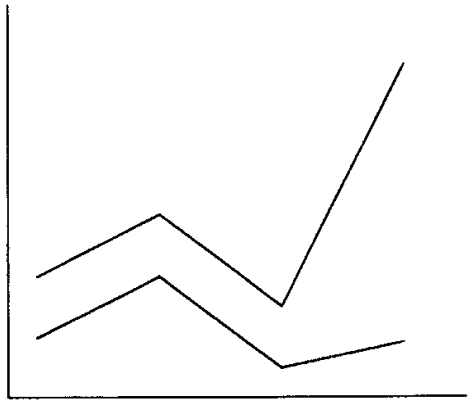
- Parallelism: (response-by-group interaction)
- The slopes of the population profile segments are the same under each condition

$$H_0 : \begin{pmatrix} \mu_{11} - \mu_{12} \\ \cdot \\ \cdot \\ \mu_{1p-1} - \mu_{1p} \end{pmatrix} = \begin{pmatrix} \mu_{21} - \mu_{22} \\ \cdot \\ \cdot \\ \mu_{2p-1} - \mu_{2p} \end{pmatrix}$$



vs.

$$H_a : \begin{pmatrix} \mu_{11} - \mu_{12} \\ \cdot \\ \cdot \\ \mu_{1p-1} - \mu_{1p} \end{pmatrix} \neq \begin{pmatrix} \mu_{21} - \mu_{22} \\ \cdot \\ \cdot \\ \mu_{2p-1} - \mu_{2p} \end{pmatrix}$$



Profile Analysis

Define

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -1 & \dots & 0 & 0 \\ .. & .. & .. & \dots & .. & .. \\ 0 & 0 & 0 & & 1 & -1 \end{pmatrix}$$

The hypothesis of parallelism can be written as

$$H_0 : \mathbf{A}\boldsymbol{\mu}_1 = \mathbf{A}\boldsymbol{\mu}_2 \quad \text{vs.} \quad H_a : \mathbf{A}\boldsymbol{\mu}_1 \neq \mathbf{A}\boldsymbol{\mu}_2$$

Profile Analysis

The test statistic for parallelism is given as

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{A}'(\mathbf{ACA}')^{-1}\mathbf{A}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

Reject the null hypothesis of parallelism at the α significance level if

$$F = \frac{n_1 + n_2 - p}{(n_1 + n_2 - 2)(p-1)} T^2$$

$$> F_{\alpha,\, p\text{-}1,\, n_1+n_2-p}$$

Profile Analysis

- If the null hypothesis of parallelism is not rejected at the α significance level, we can test the hypothesis of the same level.

Define

$$\mathbf{1}' = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

The hypothesis of equal treatment effects can be written as

$$H_0 : \mathbf{1}'\boldsymbol{\mu}_1 = \mathbf{1}'\boldsymbol{\mu}_2 \quad \text{vs.} \quad H_a : \mathbf{1}'\boldsymbol{\mu}_1 \neq \mathbf{1}'\boldsymbol{\mu}_2$$

Reject the null hypothesis of the same level

$$\text{if } |t| = \left| \frac{\mathbf{1}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\sqrt{\mathbf{1}'\mathbf{C}\mathbf{1}(1/n_1 + 1/n_2)}} \right| > t_{\alpha/2, n_1+n_2-2}$$

Profile Analysis

If the null hypothesis of parallelism is not rejected at the α significance level, we can test the hypothesis of the equal response

The hypothesis of equal mean responses can be written as

$$H_0: \ \mathbf{A}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = 0 \ \text{ vs. } H_a: \ \mathbf{A}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \neq 0$$

The grand mean vector

$$\bar{\mathbf{x}} = \frac{n_1}{n_1 + n_2}\bar{\mathbf{x}}_1 + \frac{n_2}{n_1 + n_2}\bar{\mathbf{x}}_2$$

$$T^2 = (n_1 + n_2)\bar{\mathbf{x}}'\mathbf{A}'(\mathbf{ACA}')^{-1}\mathbf{A}\bar{\mathbf{x}}$$

Reject the null hypothesis of equal mean responses

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

Profile Analysis

- Example: WAIS Dataset
  - Hypothesis of parallelism

The vectors of mean differences

- $\bar{\mathbf{x}}_1\mathbf{A}' = (3.00 \;\; -1.92 \;\; 3.52)$ and $\bar{\mathbf{x}}_2\mathbf{A}' = (3.42 \;\; -3.17 \;\; 3.75)$

$$\mathbf{ACA}' = \begin{pmatrix} 5.98 & -3.90 & -1.10 \\ & 10.35 & -4.09 \\ & & 12.16 \end{pmatrix} \text{ and } (\mathbf{ACA}')^{-1} = \begin{pmatrix} 0.2598 & 0.1235 & 0.0065 \\ & 0.1701 & 0.0683 \\ & & 0.1111 \end{pmatrix}$$

$$T^2 = 1.464$$

$$F = \frac{(45)(1.464)}{(47)(3)} = 0.47 < F_{0.05,3,45} = 2.80$$

Fail to reject the null hypothesis of parallelism at the 5% significance level.

Profile Analysis

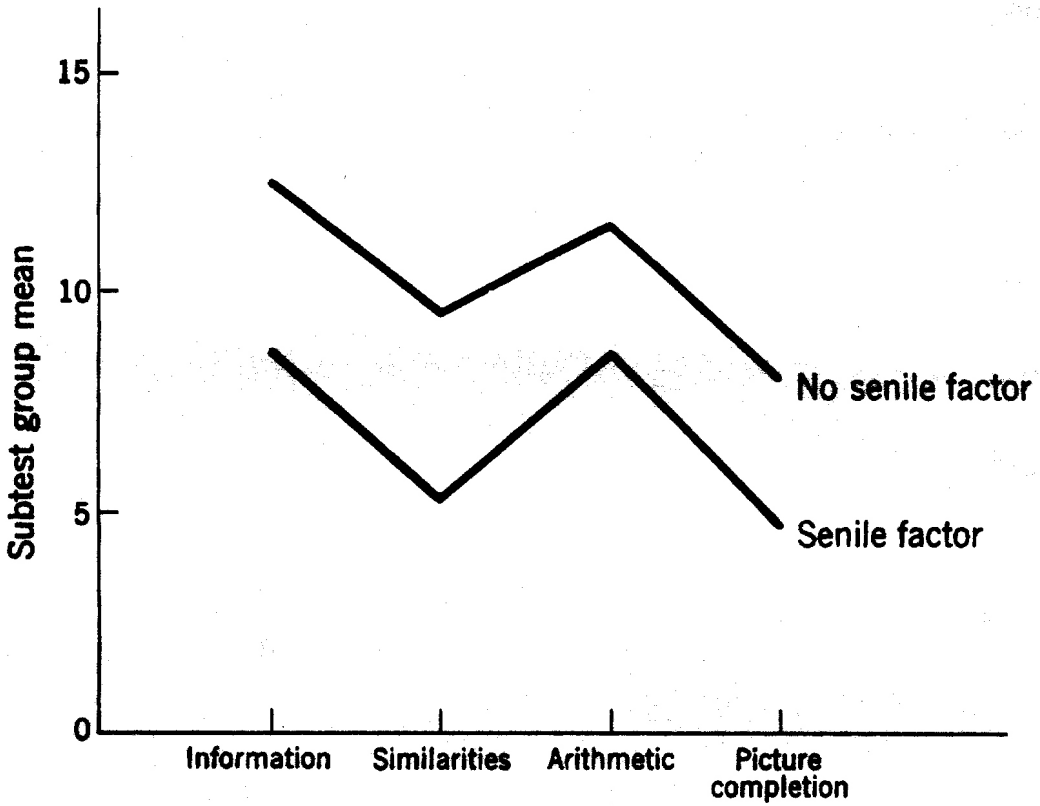


FIGURE 2.3 WAIS subtest profiles.

Profile Analysis

- Example: WAIS Dataset
  - Hypothesis of the equal mean responses

The grand mean vector

$$\bar{\mathbf{x}}' = \begin{pmatrix} 11.63 & 8.53 & 10.76 & 7.18 \end{pmatrix}$$

$$\bar{\mathbf{x}}'\mathbf{A}' = (3.10 \;\; -2.33 \;\; 3.58)$$

$$T^2 = 49(3.10 \;\; -2.33 \;\; 3.58)(\mathbf{ACA}')^{-1}\begin{pmatrix} 3.10 \\ -2.33 \\ 3.58 \end{pmatrix} = 167.15$$

Profile Analysis

- Example: WAIS Dataset
  - Hypothesis of the equal mean responses

$$F = \frac{(37+12-4)167.15}{(37+12-2)3} = 53.35$$

$$> F_{0.05,3,45} = 2.81$$

Reject the null hypothesis of the equal mean responses at the 5% significance level and conclude that the mean responses of WAIS score are different for different domains.

Profile Analysis

- Example: WAIS Dataset
  - Hypothesis of the same level

The overall mean for NSF is $\mathbf{1}'\bar{\mathbf{x}}_1$=41.60 and

the overall mean for SFP is $\mathbf{1}'\bar{\mathbf{x}}_2$=27.33, where

$\mathbf{1}' = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}$

In addition, $\mathbf{1}'\mathbf{S}\mathbf{1} = 107.06$

$$|t| = \left| \frac{41.60 - 27.33}{\sqrt{107.06(1/37 + 1/12)}} \right| = 4.15 > t_{0.025,47} = 2.02$$

Reject the null hypothesis of the same level at the 5% significance level and conclude that the mean overal WAIS score for NSF group is higher than that of SFP group.

Assumptions

- Multivariate normality
- Equal covariance matrices
- Independent samples

Likelihood Function

- Definition:
- Let $f(x_1,\ldots,x_n;\theta), \theta \in \Theta \subseteq R^k$, be the joint probability (or density) function of n random variables $X_1,\ldots,X_n$: with sample values $x_1,\ldots,x_n$

$$L(\theta, x_1,\ldots,x_n) = f(x_1,\ldots,x_n;\theta)$$

- The likelihood function of the sample is given by:

- If $X_1,\ldots,X_n$ are <u>discrete</u> iid random variable with probability function $p(x,\theta)$,

then, the **likelihood function** is given by

$$L(\theta) = P(X_1 = x_1,\ldots,X_n = x_n)$$

$$= \prod_{i=1}^{n} P(X_i = x_i)$$

$$= \prod_{i=1}^{n} P(x_i,\theta)$$

- In the <u>continuous</u> case, if the density is $f(x,\theta)$ then, **the likelihood function** is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

i.e. Let $X_1,...,X_n$ be $N(\mu,\sigma^2)$ iid random variables.

Find the Likelihood function?

$$L(\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x_i-\mu)^2}{2\sigma^2}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2})$$

The Likelihood Function

$X \sim f(x, \theta)$ pdf. parameter θ

Likelihood function

$$L(\mathcal{X}; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

MLE

$$\widehat{\theta} = \arg\max_{\theta} L(\mathcal{X}; \theta)$$

log-likelihood

$$\ell(\mathcal{X}; \theta) = \log L(\mathcal{X}; \theta)$$

Ex: Let $X_1,\ldots,X_n$ *be* $N(\mu,\sigma^2)$

- if μ is unknown and $\sigma^2=\sigma_0^2$ is known, find the MLE for μ.
- if $\mu=\mu_0$ is known and σ^2 unknown, find the MLE for σ^2.
- if μ and σ^2 are both unknown, find the MLE for $\theta(\mu,\sigma^2)$.
- Ans:

Let $\theta=\sigma^2$, so the likelihood function is:

$$L(\mu,\theta)=(2\pi\theta)^{-n/2}\exp\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta}\right)$$

- b) When $\mu = \mu_0$ is known, so we only need to solve one parameter $\theta = \sigma^2$:

$$\frac{\partial}{\partial\theta}(\ln L(\mu,\theta)) = -\frac{n}{2\theta} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\theta^2} = 0$$

$$\Rightarrow \hat{\theta} = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{n}$$

- c) When both μ and θ unknown, we need to differentiate both parameters, and mostly follow the same steps by part a). and b).

Likelihood ratio

- Define L^*_j = max $L(X\ ;\ \theta)$, the maxima of the likelihood for each of the hypotheses.

$$\lambda(\mathcal{X}) = \frac{L_0^*}{L_1^*}$$

- Likelihood ratio test
- Rejection region: $R = \{x : \lambda(x) < c\}$

$$\sup_{\theta \in \Omega_0} P_\theta(x \in R) = \alpha$$

Theorem (Wilks)

- If $\Omega_1 \subset R^q$ is a q-dimensional space and if $\Omega_0 \subset \Omega_1$ is an r-dimensional subspace, then under regularity conditions for $n \to \infty$

$$\forall\ \theta \in \Omega_0 : -2 \log \lambda \xrightarrow{\mathcal{L}} \chi^2_{q-r}.$$

Test problem 1

$X_1, \ldots, X_n$, i.i.d. with $X_i \sim N_p(\mu, \Sigma)$

$$H_0 : \mu = \mu_0, \quad \Sigma \text{ known}, \qquad H_1 : \text{ no constraints.}$$

$\Omega_0 = \{\mu_0\}, r = 0, \Omega_1 = \mathbb{R}^p, q = p$

$$-2\log\lambda = 2(\ell_1^* - \ell_0^*) = n(\overline{x} - \mu_0)^\top \Sigma^{-1}(\overline{x} - \mu_0)$$

$$-2\log\lambda \sim \chi_p^2$$

Rejection region R: $\{x \in \mathbb{R}^n \text{ such that } -2\log\lambda > \chi^2_{0.95;p}\}$

Example

設一樣品變積矩陣為

$$\hat{\Sigma} = \begin{pmatrix} 3.42 & 2.60 & 1.89 \\ & 8.00 & 6.51 \\ \text{對稱} & & 9.62 \end{pmatrix}$$

其樣品大小 $n = 20$， $p = 3$，其假設檢定為

$$H_0 : \Sigma = \Sigma_0 = \begin{pmatrix} 4 & 3 & 2 \\ & 6 & 5 \\ & & 10 \end{pmatrix}$$

利用 (7.1) 及 (7.2) 式

$$|\Sigma_0| = 86,\ \left|\hat{\Sigma}\right| = 88.6355,\ v = 19,\ tr\hat{\Sigma}\Sigma_0^{-1} = 3.2217$$

$$L = 19(\ln 86 - \ln 88.6355 + 3.2217 - 3) = 3.6388$$

$$L' = \left[1 - \frac{1}{6 \times 19 - 1}\left(2 \times 3 + 1 - \frac{2}{3+1}\right)\right] \times 3.6388$$

$$= 0.9425 \times 3.6388 = 3.4296 < \chi^2_{0.05,6} = 12.592$$

Homogeneity of Covariance Matrix

- Hypothesis of Equal Covariance Matrices of m p-dimensional multivariate normal distributions

H_o: $\Sigma_1 = \ldots = \Sigma_m$

Homogeneity of Covariance Matrix

$\mathbf{S}_j$ be the pxp sample covaraince matrix of sample j with sample size n_j and degrees of freedom n_j-1

The pooled pxp covariance matrix is given as

$$\mathbf{S} = \frac{1}{\sum_{j=1}^{m}(n_j-1)} \sum_{j=1}^{m}(n_j-1)\mathbf{S}_j$$

Homogeneity of Covariance Matrix

$$M=\sum_{j=1}^{m}(n_j-1)\ln|\mathbf{S}|-\sum_{j=1}^{m}(n_j-1)\ln|\mathbf{S}_j|$$

Scale factor

$$K^{-1}=1-\frac{2p^2+3P-1}{6(p+1)(m-1)}\left[\sum_{j=1}^{m}\frac{1}{(n_j-1)}-\sum_{j=1}^{m}\frac{1}{\sum_{j=1}^{m}(n_j-1)}\right]$$

$$X=MK^{-1}\sim\chi^2_{(m-1)p(p+1)/2}$$

Homogeneity of Covariance Matrix

- Reaction Times

  32 male and 32 female normal subjects reacted to visual stimuli preceded by warning intervals of different lengths in 0.5 and 0.15 seconds

Homogeneity of Covariance Matrix

The sample covariance matrices for male and female

$$\mathbf{S}_M = \begin{pmatrix} 4.32 & 1.88 \\ & 9.18 \end{pmatrix} \text{ and } \mathbf{S}_F = \begin{pmatrix} 2.52 & 1.90 \\ & 10.06 \end{pmatrix}$$

The pooled covariance matrix

$$\mathbf{S} = \begin{pmatrix} 3.42 & 1.89 \\ & 9.62 \end{pmatrix}$$

$$M=2(31-1)\ln|\mathbf{S}|-(31-1)[\ln|\mathbf{S}_\mathbf{M}|+\ln|\mathbf{S}_\mathbf{M}|]$$

$$=62\ln(29.328)-30[\ln(36.123)+\ln(21.741)]$$

$$=2.82$$

$K^{-1}=0.965$ and $X=2.82\text{x}0.965=2.72 < \chi^2_{0.05,3} = 7.81$

Fail to reject the null hypothesis of equal covariance matrix

Summary

- Hotelling T^2 statistics
  - One sample
  - Two independent samples (unpaired)
  - Paired samples
  - Confidence regions
  - Simultaneous confidence interval
  - Profile analysis