A course in

MULTIVARIATE ANALYSIS

LECTURE NOTES FOR

MATHEMATICS 204 A-B

at

University of California, Irvine

by

Howard G. Tucker[*]

July, 1993

TABLE OF CONTENTS

/k

CHAPTER 1. MULTIVARIATE NORMAL DISTRIBUTION.

§1. Fundamental Facts Concerning the Multivariate Normal Distribution. The strong prerequisite for multivariate analysis is a knowledge of the multivariate normal distribution and its properties. In this section we present this strong prerequisite in considerable detail.

DEFINITION: The n random variables $X_1, \cdots, X_n$, where $n \geq 2$, are said to be jointly normal (or Gaussian) or are said to have a multivariate normal (Gaussian) distribution if there exist n independent random variables $Z_1, \cdots, Z_n$ which are each $\mathcal{N}(0,1), n$ constants $\mu_1, \cdots, \mu_n$ and an $n \times n$ nonsingular matrix $A = (a_{ij})$ of real numbers such that

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} a_{11} \cdots a_{1n} \\ \vdots \\ a_{n1} \cdots a_{nn} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$$

This can be written in vector form as

$$\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z},$$

where

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, A = \begin{pmatrix} a_{11} \cdots a_{1n} \\ \vdots \\ a_{n1} \cdots a_{nn} \end{pmatrix} \; and \; \mathbf{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}.$$

We shall adhere to the following notation. The point or vector $\mathbf{x}$ in $\mathbf{R}^n$ will be a vertical vector, $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. It will be treated like a matrix, so that $\mathbf{x}^t$ will denote its transpose $(x_1 \, x_2 \, \cdots \, x_n)$. In general, if B is any matrix, then B^t denotes its transpose. We shall have use for the following three lemmas

LEMMA 1. *If $X_1, \cdots, X_n$ are random variables with a joint absolutely continuous distribution with joint density $f(x_1 \cdots, x_n)$, and if $W_i = a_i X_i + b_i, 1 \leq i \leq n$, where $a_i > 0$ and b_i are constants, then $W_1, W_2, \cdots, W_n$ have a joint absolutely continuous distribution with density*

$$f_{W_1,\cdots,W_n}(w_1, \cdots, w_n) = \frac{1}{\prod_{i=1}^n a_i} f\left(\frac{w_1 - b_1}{a_1}, \cdots, \frac{w_n - b_n}{a_n}\right)$$

Proof: We note that

$$\begin{aligned} P[W_1 \leq w_1, \cdots, W_n \leq w_n] &= P\left[X_1 \leq \frac{w_1 - b_1}{a_1}, \cdots, X_n \leq \frac{w_n - b_n}{a_n}\right] \\ &= \int_{-\infty}^{(w_1-b_1)/a_1} \cdots \int_{-\infty}^{(w_n-b_n)/a_n} f(t_1, \cdots, t_n) dt_1 \cdots dt_n. \end{aligned}$$

Now make the change of variables (one at a time) $z_i = a_i t_i + b_i, 1 \leq i \leq n$, and obtain

$$= \int_{-\infty}^{w_1} \cdots \int_{-\infty}^{w_n} f\left(\frac{z_1 - b_1}{a_1}, \cdots, \frac{z_n - b_n}{a_n}\right) \cdot \frac{1}{\prod_{i=1}^n a_i} dz_1 \cdots dz_n.$$

By the definition of density we conclude

$$f_{W_1,\cdots,W_n}(w_1, \cdots, w_n) = f\left(\frac{w_1 - b_1}{a_1}, \cdots, \frac{w_n - b_n}{a_n}\right) \frac{1}{\prod_{i=1}^n a_i}.$$

Q.E.D.

LEMMA 2. *Let A be a non-singular $n \times n$ matrix, and denote $S = \{\mathrm{x} \in \mathbf{R}^n : A\mathrm{x} \in I\}$, where $I = \times_{j=1}^n [a_j, b_j]$. Let H be an integrable function over S. Then*

$$\int \cdots \int_I H(A\mathrm{u}) |det A| du_1, \cdots du_n = \int \cdots \int_S H(\mathrm{x}) d\mathrm{x}.$$

This lemma is a special case of a deep theorem in multivariable calculus whose proof is beyond the scope of this course.

Lemma 3. *Let $X_1 \cdots, X_n$ be random variables with a joint absolutely continuous distribution with joint density $f_{\mathbf{X}}(\mathbf{x})$. Let A be a non-singular $n \times n$ matrix, and define $\mathbf{U} = A\mathbf{X}$. Then $\mathbf{U}$ has a joint absolutely continuous distribution with joint density given by*

$$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{X}}(A^{-1}\mathbf{u})|detA^{-1}|.$$

Proof: For arbitrary $\mathbf{u} \in \mathbf{R}^n$, define

$$S = \{\mathbf{x} \in \mathbf{R}^n : A\mathbf{x} \in I\},$$

where $I = \times_{j=1}^{n}(-\infty, u_j]$. Then by Lemma 2 above, we have

$$\begin{aligned} F_{\mathbf{U}}(\mathbf{u}) &= P[\mathbf{U} \in I] = P[\mathbf{X} \in S] \\ &= \int_S \cdots \int f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} = \int_{-\infty}^{u_1} \cdots \int_{-\infty}^{u_n} f_{\mathbf{X}}(A^{-1}\mathbf{u})|detA^{-1}|d\mathbf{u}. \end{aligned}$$

By the definition of joint density, it follows that $f_{\mathbf{X}}(A^{-1}\mathbf{u})|detA^{-1}|$ is the density of $\mathbf{U}$. Q.E.D.

Definition: If $U = (U_{ij})$ is a matrix of random variables, each having finite expectation, then we define EU as the matrix of expectations $(E(U_{ij}))$. If $G(x) = (g_{ij}(x))$ is a matrix of integrable functions defined over some interval $[a, b]$, then $\int_a^b G(x)dx$ will denote the matrix of integrals $(\int_a^b g_{ij}(x)dx)$.

LEMMA 4. *If $U = (U_{ij})$ is an $m \times n$ matrix of random variables, if $A = (a_{ij})$ and $B = (b_{ij})$ are $r \times m$ and $n \times s$ matrices respectively of real numbers, then*

$$E(AUB) = AE(U)B.$$

Proof: The element in the ith row and the kth column of AU is $\sum_{r=1}^{m} a_{ir}U_{rk}$, and the element in the ith row and jth column of AUB is $\sum_{k=1}^{n}\sum_{r=1}^{m} a_{ir}U_{rk}b_{kj}$. Hence by the definition given above of the expectation of a matrix of random variables, the element in the ith row and jth column of $E(AUB)$ is $\sum_{k=1}^{n}\sum_{r=1}^{m} a_{ir}E(U_{rk})b_{kj}$. By standard matrix multiplication, this is the element in the ith row and jth column of $AE(U)B$. Q.E.D.

Definition. If $\mathbf{U}$ and $\mathbf{V}$ are $m-$ and n-dimensional random vectors respectively, and if each of the coordinates of each of them has finite second moment, then we define the covariance matrix of $\mathbf{U}, \mathbf{V}$ to be

$$Cov(\mathbf{U}, \mathbf{V}) = E((\mathbf{U} - E(\mathbf{U}))(\mathbf{V} - E(\mathbf{V}))^t),$$

i.e., $Cov(\mathbf{U}, \mathbf{V}) = (c_{ij})$, an $m \times n$ matrix, where $C_{ij} = Cov(U_i, V_j)$. The covariance matrix of the random vector $\mathbf{U}$ is defined to be $Cov(\mathbf{U}) = Cov(\mathbf{U}, \mathbf{U})$. I.e., $Cov(\mathbf{U}) = (d_{ij})$ is an $m \times m$ matrix, where $d_{ij} = Cov(U_i, U_j)$ for $i \neq j$, and $d_{ii} = Var(U_i)$.

Theorem 1. *If $X_1, \cdots, X_n$ are multivariate normal, then their joint density is*

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t C^{-1}(\mathbf{x} - \boldsymbol{\mu})),$$

where $\boldsymbol{\mu} = E\mathbf{X}$ and C is the covariance matrix, i.e., $C = E((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t)$.

Proof: From the definition of joint normality of $\mathbf{X}$, we know $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$, where $Z_1, \cdots, Z_n$ are independent and $N(0,1)$, and A is a non-singular matrix of real numbers. Thus

$$f_{\mathbf{Z}}(\mathbf{z}) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\mathbf{z}^t\mathbf{z}}$$

Note that $\mathbf{Z} = A^{-1}(\mathbf{X} - \boldsymbol{\mu})$; hence by Lemma 1 and Lemma 3 above,

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t (A^{-1})^t A^{-1}(\mathbf{x} - \boldsymbol{\mu})\}|detA^{-1}|.$$

Let us define the matrix C by $C = AA^t$. Since $(A^t)^{-1} = (A^{-1})^t$, we have $C^{-1} = (AA^t)^{-l} = (A^t)^{-1}A^{-1} = (A^{-1})^t A^{-1}$. From here on, $|D|$ will denote the determinant of a matrix D, and $|detD|$ will denote the absolute value of the determinant of D .) Thus, $|C^{-1}| = |(A^t)^{-1}||A^{-1}| = |A^{-1}|^2$, and hence $|detA^{-1}| = \sqrt{|C^{-1}|}$. We now have

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t C^{-1}(\mathbf{x} - \boldsymbol{\mu})).$$

We have yet to determine $\boldsymbol{\mu}$ and C . We easily check from the definition that $E(\mathbf{X}) = \boldsymbol{\mu}$. Thus

$$\begin{aligned} Cov(X) &= E((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^t) = E(A\mathbf{Z}\mathbf{Z}^t A^t) \\ &= AE(\mathbf{Z}\mathbf{Z}^t)A^t = AIA^t = AA^t = C. \end{aligned}$$

Q. E.D.

From Theorem 1, we see that the joint density and hence the distribution of a multivariate normal distribution is determined by its expectation vector $\boldsymbol{\mu}$ and covariance matrix C. Thus, we shall write: $\mathbf{X}$ is $\mathcal{N}(\boldsymbol{\mu}, C)$, which means: $\mathbf{X}$ is an n-dimensional random vector whose expectation (or mean) vector is $\boldsymbol{\mu}$ and whose covariance matrix is C .

We now relate the usual definition of the multivariate normal distribution to that given above. Also, a number of properties of the multivariate normal distribution that are most frequently used in multivariate analysis and linear regression analysis will be obtained.

LEMMA 5. *If* $\mathbf{X}$ *is* $\mathcal{N}_n(\boldsymbol{\mu}, C)$ *, then* C *is positive definite.*

Proof: By definition, $\mathbf{X} = A\mathbf{Z}+\boldsymbol{\mu}$, where $Z_1 \cdots, Z_n$ are *i.i.d.*$\mathcal{N}(0,1)$ and A is a non-singular $n \times n$ matrix. By the proof of Theorem 1 in §1, $C = AA^t$. Let $\mathbf{x} \in \mathbf{R}^n, \mathbf{x} \neq \mathbf{0}$. Then $A^t\mathbf{x} \neq \mathbf{0}$ and $\mathbf{x}^tC\mathbf{x} = \mathbf{x}^tAA^t\mathbf{x} = (A^t\mathbf{x})^tA^t\mathbf{x} > 0$. Q.E.D.

THEOREM 2. *If* $X_1, \cdots, X_n$ *are random variables with a joint absolutely continuous distribution with density*

$$f_{\mathbf{X}}(\mathbf{x}) = K\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^tD(\mathbf{x}-\boldsymbol{\mu})\} \tag{1}$$

for all $\mathbf{x} \in \mathbf{R}^n$ *, where* $\boldsymbol{\mu}$ *is a vector of constants, and* D *is a symmetric* $n \times n$ *positive definite matrix, then* $\mathbf{X}$ *is* $\mathcal{N}_n(\boldsymbol{\mu}, D^{-1})$.

Proof: Since D is positive definite, so is D^{-1} . Now given the positive definite symmetric matrix D^{-1}, it is known that there exists a symmetric positive definite matrix $D^{-1/2}$ which satisfies $D^{-1/2}D^{-1/2} = D^{-1}$. Let $Z_1, \cdots, Z_n$ be n independent $\mathcal{N}(0,1)$ random variables, and let us write

$$\mathbf{Y} = D^{-1/2}\mathbf{Z} + \boldsymbol{\mu}.$$

By Theorem 1 of §1, we know that $\mathbf{Y}$ is multivariate normal with density

$$f_{\mathbf{Y}}(\mathbf{x}) = \frac{\sqrt{|D|}}{(2\pi)^{n/2}}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^tD(\mathbf{x}-\boldsymbol{\mu})}.$$

Hence $X_1, \cdots, X_n$ are multivariate normal, and $K = (|D|/(2\pi)^n)^{1/2}$. Q.E.D.

LEMMA 6. *Let C be the covariance matrix of n jointly normal random variables, and let C be partitioned as follows:*

$$C = \left(\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{array}\right)$$

where C_{11} is a $k \times k$ submatrix, $1 < k < n$. Then C_{11} is non-singular, symmetric and positive definite.

Proof: We need only prove that C_{11} is positive definite. Indeed, let $\mathbf{x} \in \mathbf{R}^k$ be such that $\mathbf{x} \neq \mathbf{0}$. Define $\mathbf{y} = \binom{\mathbf{x}}{\mathbf{0}}$. Then $\mathbf{y} \neq \mathbf{0}$. Since C is positive definite, then

$$0 < \mathbf{y}^t C \mathbf{y} = \mathbf{x}^t C_{11} \mathbf{x}.$$

Q.E.D.

Lemma 7. *If $\mathbf{X}$ is $\mathcal{N}_n(\mu, C)$, and B is a non-singular $n\times n$ matrix, then $B\mathbf{X}$ is $\mathcal{N}_n(B\mu, BCB^t)$*

.

Proof: By definition $\mathbf{X} = A\mathbf{Z} + \mu$ where $Z_l \cdots, Z_n$ are *i.i.d.* and $\mathcal{N}(0,1)$. By the proof of Theorem 1 in Section 1, $C = AA^t$. Now $B\mathbf{X} = BA\mathbf{Z}+B\mu$, and since BA is non-singular, we obtain by the definition that $B\mathbf{X}$ is $\mathcal{N}_n(B\mu, D)$, where $D = BA(BA)^t = BAA^tB^t = BCB^t$.

Q E D

Our next aim is to show that marginals of multivariate normal distributions are univariate or multivariate normal. We first introduce a useful marginal matrix notation.

Notation. Let $C = (C_{ij})$ be an $n \times n$ matrix, let $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in R^n$, and let $S = \{r_1, \cdots, r_k\}$, where $1 \le r_1 < \cdots < r_k \le n$. Then we define the $k \times k$ matrix C_S and the k-dimensional vector $\mathbf{x}_S$ by $C_S = (c_{r_i r_j})$ and $\mathbf{x}_S = \begin{pmatrix} x_{r_1} \\ \vdots \\ x_{r_k} \end{pmatrix}$.

THEOREM 3. *If* $\mathbf{X}$ *is* $\mathcal{N}(\mu, C)$, *and if* $S = \{r_1, \cdots, r_n\}$, *where* $1 \le r_1 < \cdots < r_k \le n$, *then* $\mathbf{X}_S$ *is* $\mathcal{N}_k(\mu_S, C_S)$.

Proof: We first prove the theorem in the special case where $r_i = i, 1 \le i \le k$. By Lemma 5, C is positive definite. Let us partition C as follows:

$$C = \left(\begin{array}{c:c} C_{11} & C_{12} \\ \hdashline C_{21} & C_{22} \end{array}\right)$$

where C_{11} is a $k \times k$ submatrix. By Lemma 6, C_{11} is nonsingular. Let B denote the $k \times n$ partitioned matrix, $B = (I_k | 0)$, where I_k is the $k \times k$ identity matrix. Let A_1 denote the $(n-k) \times k$ matrix

$$A_1 = (C_{11}^{-1}(-C_{12}))^t,$$

and let A_2 denote the partitioned matrix

$$A_2 = (A_1 | I_{n-k}),$$

where I_{n-k} is the $(n-k) \times (n-k)$ identity matrix. Then

$$\begin{aligned} \left(\frac{B}{A_2}\right) C(B^t | A_2^t) &= \left(\begin{array}{c|c} I_k & 0 \\ \hline (C_{11}^{-1}(-C_{\cdot 1}))^t & I_{n-k} \end{array}\right) \left(\begin{array}{c|c} C_{11} & C_{12} \\ \hline C_{21} & C_{22} \end{array}\right) \left(\begin{array}{c|c} I_k & C_{11}^{-1}(-C_{12}) \\ \hline 0 & I_{n-k} \end{array}\right) \\ &= \left(\begin{array}{c|c} C_{11} & C_{12} \\ \hline 0 & -C_{21}C_{11}^{-1}C_{12} + C_{12} \end{array}\right) \left(\begin{array}{c|c} I_k & C_{11}^{-1}(-C_{12}) \\ \hline 0 & I_{n-k} \end{array}\right) \end{aligned}$$

$$= \left(\begin{array}{c:c} C_{11} & 0 \\ \hdashline 0 & D \end{array}\right)$$

where $D = C_{22} - C_{21}C_{11}^{-1}C_{12}$ is an $(n-k) \times (n-k)$ symmetric positive definite matrix. We now consider the vector random variables

$$\mathbf{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_k \end{pmatrix} \; and \, \mathbf{V} = \begin{pmatrix} V_1 \\ \vdots \\ V_{n-k} \end{pmatrix}$$

defined by

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} = \left(\frac{B}{A_2}\right)(\mathbf{X} - \boldsymbol{\mu}) \; or \; \mathbf{X} = \left(\frac{B}{A_2}\right)^{-1} \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} + \boldsymbol{\mu}$$

with the associated transformation

$$\begin{pmatrix} \mathbf{u} \\ \cdots \\ \mathbf{v} \end{pmatrix} = \left(\frac{B}{A_2}\right)(\mathbf{x} - \boldsymbol{\mu}) \; or \; \mathbf{x} = \left(\frac{B}{A_2}\right)^{-1} \begin{pmatrix} \mathbf{u} \\ \cdots \\ \mathbf{v} \end{pmatrix} + \boldsymbol{\mu}.$$

We note that

$$(\mathbf{x}^t - \boldsymbol{\mu}^t)C^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{u}^t|\mathbf{v}^t)(B^t|A_2^t)^{-1}C^{-1}\left(\frac{B}{A_2}\right)^{-1} \begin{pmatrix} \mathbf{u} \\ \cdots \\ \mathbf{v} \end{pmatrix}$$

$$= (\mathbf{u}^t|\mathbf{v}^t)\left(\left(\frac{B}{A_2}\right)C(B^t|A_2^t)\right)^{-1}\left(\frac{\mathbf{u}}{\mathbf{v}}\right) = \begin{pmatrix} \mathbf{u} \\ \cdots \\ \mathbf{v} \end{pmatrix} \left(\begin{array}{c|c} C_{11}^{-1} & 0 \\ \hline 0 & D^{-1} \end{array}\right) \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$$

$$= \mathbf{u}^t C_{11}^{-1}\mathbf{u} + \mathbf{v}^t D^{-1}\mathbf{v}.$$

Also one verifies that

$$\left|\left(\frac{B}{A_2}\right)\right| = 1 \; and \; |C^{-1}| = |C_{11}^{-1}||D^{-1}|.$$

Thus by lemma 6 and lemma 7 we obtain

$$f_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v}) = \frac{\sqrt{|C_{11}^{-1}|}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\mathbf{u}^t C_{11}^{-1}\mathbf{u}\right) \frac{\sqrt{|D^{-1}|}}{(2\pi)^{n-k}} \exp\left(-\frac{1}{2}\mathbf{v}^t D^{-1}\mathbf{v}\right).$$

The marginal distribution of **U** is obtained by integrating out the v's above, and we obtain

$$f_{\mathbf{U}}(\mathbf{u}) = \frac{\sqrt{|C_{11}^{-1}|}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\mathbf{u}^t C_{11}^{-1}\mathbf{u}\right).$$

Thus, the joint distribution of $X_1, \cdots, X_k$ has a density

$$f_{X_1,\cdots,X_k}(\mathbf{w}) = \frac{\sqrt{|C_{11}^{-1}|}}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\nu})^t C_{11}^{-1}(\mathbf{w}-\boldsymbol{\nu})\right),$$

where $\boldsymbol{\nu}^t = (\mu_1 \cdots \mu_k)^t$. We now prove the theorem for arbitrary S. Let D be an $n \times n$ permutation matrix such that the first k rows and first k columns of DCD^t are the $r_1\underline{\text{th}}, \cdots, r_k$th rows and r_1th$, \cdots, r_k$th columns respectively of C . Clearly the submatrix of DCD^t determined by its first k rows and first k columns is C_S , and the first k coordinates of $D\boldsymbol{\mu}$ are the coordinates of $\boldsymbol{\mu}_S$. By the special case proved above, we now have the general conclusion. Q.E.D.

In the particular case when S consists of one number between 1 and n , say $S = \{i\}$, then

$$X_S = X_i \ is \ \mathcal{N}(\mu_i, c_{ii}).$$

It should be pointed out that it is possible for each of n random variables $X_1, \cdots, X_n$ to have a normal distribution, and yet together they do not have a joint normal distribution. As an example, consider the random variables X and Y which have a joint absolutely continuous distribution with density

$$f_{X,Y}(x,y) = \begin{cases} (1/\pi)\exp -\frac{1}{2}(x^2+y^2) & \text{if } xy \geq 0 \\ 0 & \text{if } xy < 0 \end{cases}$$

It is easy to verify that X and Y are each $\mathcal{N}(0,1)$, they are not independent, and they are not jointly normal.

THEOREM 4. *If* $\mathbf{X}$ *is* $\mathcal{N}_n(\mu, C)$, *and if* A *is a* $k \times n$ *matrix n of rank* $k \leq n$ *, then* $A\mathbf{X}$ *is* $\mathcal{N}_k(A\mu, ACA^t)$.

Proof: By Lemma 7 we need only prove the theorem in the case $k < n$. In this case there exists an $(n-k) \times n$ matrix B such that the partitioned matrix $\left(\frac{A}{B}\right)$ is non-singular . By Lemma 7, $\left(\frac{A}{B}\right)\mathbf{X} = \left(\frac{A\mathbf{X}}{B\mathbf{X}}\right)$ is multivariate normal with mean vector $\left(\frac{A\mu}{B\mu}\right)$ and covariance matrix

$$\left(\frac{A}{B}\right) C \left(A^t | B^t\right) = \left(\frac{AC}{BC}\right)(A^t|B^t) = \left(\begin{array}{c|c} ACA^t & ACB^t \\ \hline BCA^t & BCB^t \end{array}\right)$$

We now apply Theorem 3 to obtain our conclusion. Q.E.D.

If k is any positive integer, we shall let I_k denote the $k \times k$ identity matrix. Thus, if the n coordinates of $\mathbf{X}$ are independent and $\mathcal{N}(0,1)$, then easily $\mathbf{X}$ is $\mathcal{N}_n(\mathbf{0}, I_n)$, and conversely.

Theorem 5. *If* $\mathbf{X}$ *is* $\mathcal{N}_n(\mu, C)$, *if* $\mathbf{Y}$ *is* $\mathcal{N}_n(\nu, D)$ *, and if* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent, then* $\mathbf{X}+\mathbf{Y}$ *is* $\mathcal{N}(\mu+\nu, C+D)$.

Proof: By the definition of the multivariate normal distribution, there exist non-singular $n \times n$ matrices A and B such that $\mathbf{X} = A\mathbf{U} + \mu$, and $\mathbf{Y} = B\mathbf{V} + \nu$, where $\mathbf{U}$ and $\mathbf{V}$ are each $\mathcal{N}_n(\mathbf{0}, I_n)$. Since $\mathbf{X}$ and $\mathbf{Y}$ are independent, one can easily prove that $\mathbf{U}$ and $\mathbf{V}$ are independent. Hence $\binom{\mathbf{U}}{\mathbf{V}}$ is $\mathcal{N}_{2n}(\mathbf{0}, I_{2n})$. Since $\binom{\mathbf{X}}{\mathbf{Y}} = \left(\begin{array}{cc} A & 0 \\ \hdashline 0 & B \end{array}\right) \binom{\mathbf{U}}{\mathbf{V}} + \binom{\boldsymbol{\mu}}{\boldsymbol{\nu}}$, it follows that $\binom{\mathbf{X}}{\mathbf{Y}}$ is multivariate normal with mean vector $\binom{\boldsymbol{\mu}}{\boldsymbol{\nu}}$ and covariance matrix given by

$$\left(\begin{array}{c|c} A & 0 \\ \hdashline 0 & B \end{array}\right) \left(\begin{array}{c|c} A^t & 0 \\ \hdashline 0 & B^t \end{array}\right) = \left(\begin{array}{c|c} AA^t & 0 \\ \hdashline 0 & BB^t \end{array}\right) = \left(\begin{array}{c|c} C & 0 \\ \hdashline 0 & D \end{array}\right).$$

But by Theorem 4, $(I_n|I_n)\binom{\mathbf{X}}{\mathbf{Y}}$ is multivariate normal with mean vector given by $(I_n|I_n)\binom{\boldsymbol{\mu}}{\boldsymbol{\nu}} = \mathbf{u} + \boldsymbol{\nu}$ and covariance matrix given by

$$(I_n|I_n) \left(\begin{array}{c|c} C & 0 \\ \hdashline 0 & D \end{array}\right) \left(\frac{I_n}{I_n}\right) = C + D.$$

Q.E.D.

THEOREM 6. *If* $\mathbf{U}$ *and* $\mathbf{V}$ *are* $m-$ *and* n*-dimensional random vectors and if* $\binom{\mathbf{U}}{\mathbf{V}}$ *is* $\mathcal{N}_{m+n}\left(\binom{\boldsymbol{\mu}}{\boldsymbol{\nu}}, \left(\begin{array}{c|c} C_{11} & C_{12} \\ \hdashline C_{21} & C_{22} \end{array}\right)\right)$, *where* C_{11} *is* $m \times m$, *then* $\mathbf{U}, \mathbf{V}$ *are independent if and only if* $C_{12} = C_{21}^t = 0$, *i.e.,* $Cov(\mathbf{U}, \mathbf{V}) = 0$.

Proof: By Theorem 3, $\mathbf{U}$ is $\mathcal{N}_m(\boldsymbol{\mu}, C_{11})$ and $\mathbf{V}$ is $\mathcal{N}_n(\boldsymbol{\nu}, C_{22})$. If $\mathbf{U}$ and $\mathbf{V}$ are independent, then their joint density equals the product of the two densities, i.e., the product of

$$\frac{\sqrt{|C_{11}^{-1}|}}{(2\pi)^{m/2}} \exp -\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^t C_{11}^{-1}(\mathbf{u} - \boldsymbol{\mu})$$

and

$$\frac{\sqrt{|C_{22}^{-1}}}{(2\pi)^{n/2}} \exp -\frac{1}{2}(\mathbf{v} - \boldsymbol{\nu})^t C_{22}^{-1}(\mathbf{v} - \boldsymbol{\nu})$$

which is easily seen to be

$$\frac{\sqrt{|C^{-1}|}}{(2\pi)^{(m+n)/2}} \exp -\frac{1}{2}\left(\binom{\mathrm{u}}{\mathrm{v}} - \binom{\mu}{\nu}\right)^t C^{-1}\left(\binom{\mathrm{u}}{\mathrm{v}} - \binom{\mu}{\nu}\right),$$

where $C = \left(\begin{array}{c|c} C_{11} & 0 \\ \hline 0 & C_{22} \end{array}\right)$, which proves that $C_{12} = C_{21}^t = 0$. If, conversely, $C_{12} = C_{21}^t = 0$, then by a little algebra, one shows that the joint density of **U** and **V** factors into the product of their densities. Q.E.D.

THEOREM 7. *If $\binom{X}{Y}$ is $\mathcal{N}_2(\mu, C)$, then there exist constants a and b and a random variable Z such that*

$$Y = a + bX + Z$$

where $E(Z) = 0$ and X and Z are independent.

Proof: Whatever the value of b , we see that the value of a must be taken as $a = E(Y) - bE(X)$. Let us define Z by

$$Z = Y - a - \frac{Cov(X,Y)}{Var(X)}X.$$

Then

$$\binom{X}{Z} = \binom{0}{-a} + \left(\begin{array}{c|c} 1 & 0 \\ \hline -\frac{Cov(X,Y)}{Var(X)} & 1 \end{array}\right)\binom{X}{Y}.$$

By Lemma 7, the random vector $\binom{X}{Z}$ is bivariate normal. An easy computation yields: $Cov(X, Z) = 0$. By Theorem 6, we obtain that Z and X are independent. Thus if we take $b = \frac{Cov(X,Y)}{Var(X)}$ and $a = E(Y) - bE(X)$, we obtain the conclusion. Q.E.D.

If we denote σ_X and σ_Y as the standard deviations of X and Y respectively, and if $\rho_{x,y}$ denotes the correlation coefficient of X and Y , then, if $\binom{X}{Y}$ is $\mathcal{N}_2(\mu, C)$, we may write

$$Y - E(Y) = \frac{\sigma_Y}{\sigma_X}\rho_{X,Y}(X - E(X)) + Z,$$

where Z and X are independent and $E(Z) = 0$. This is a restatement of Theorem 7.

We conclude this long section with a simple application. The following problem frequently arises in bio-medical research. One has n independent observation, $\binom{X_1}{Y_1}, \cdots, \binom{X_n}{Y_n}$ on a random vector $\binom{X}{Y}$ whose joint distribution is bivariate normal with mean vector $\binom{\mu}{\nu}$. One wishes to test the null hypothesis $H_0 : \mu = \nu$ against the alternative hypothesis $H_1 : \mu \neq \nu$ with level of significance α. A common mistake made is that of doing a two-sample t-test. However, for each i, X_i and Y_i are not necessarily independent; this happens when X_i and Y_i are, e.g., two particular measurements on the same patient. In such a case one must do a paired comparison test.

PROPOSITION 1. *If $\binom{X_1}{Y_1}, \cdots, \binom{X_n}{Y_n}$ are independent $\mathcal{N}_1(\mu, C)$ random vectors, then $X_1 - Y_1, \cdots, X_n - Y_n$ are independent $\mathcal{N}(\mu_1 - \mu_2, \tau^2)$ random variables, where $\tau^2 = Var(X_1) + Var(Y_1) - 2Cov(X_1, Y_1)$.*

Proof: Let A be the $n \times 2n$ matrix defined as follows:

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & & & \\ 0 & 0 & 0 & 0 & 1 & -1 & & & \\ \vdots & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}$$

and let $\mathbf{W}$ be the $2n$-dimensional random vector define by

$$\mathbf{W}^t = (X_1 Y_1 X_2 Y_2 \cdots X_n Y_n)^t.$$

Then

$$\begin{pmatrix} X_1 - Y_1 \\ \vdots \\ X_n - Y_n \end{pmatrix} = A\mathbf{W},$$

and $\mathbf{W}$ is $\mathcal{N}_{2n}(\mathbf{1}_n \otimes \boldsymbol{\mu}, I_n \otimes C)$, where $\mathbf{1}_n \otimes \boldsymbol{\mu}$ is the $2n$-dimensional vector,

$$\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{pmatrix}.$$

and $I_n \otimes C$ is the $2n \times 2n$ matrix

$$\left(\begin{array}{c|c|c|c} \mathrm{C} & 0 & & 0 \\ \hline 0 & \mathrm{C} & & 0 \\ \hline & & & \\ \hline 0 & 0 & & \mathrm{C} \end{array}\right)$$

It is easy to see that $rank(A) = n$, and thus by Lemma 6 we have that $A\mathbf{W}$ is $\mathcal{N}_n((\mu_1 - \mu_2)\mathbf{1}_n, A(I_n \otimes C)A^t)$. It is easy to verify that

$$A(I_n \otimes C)A^t = diag\{\tau^2, \cdots, \tau^2\}.$$

Thus $X_1 - X_1, \cdots, X_n - Y_n$ are independent, each being $\mathcal{N}(\mu_1 - \mu_2, \tau^2)$, and H_0 is true if and only if $\mu_1 - \mu_2 = 0$. If we let $Z_i = X_i - Y_i, 1 \leq i \leq n, \bar{Z}_n = \bar{X}_n - \bar{Y}_n$ and

$$s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2,$$

then, if H_0 is true, the statistic T defined by

$$T = \frac{\sqrt{n}\bar{Z}_n}{s_n}$$

has the t-distribution with $n-1$ degrees of freedom. We would accordingly reject H_0 if $|T| \geq C$, where C is obtained from the $(n-1)$th row and the $1-\alpha/2$ column of the tables of the t-distribution.

EXERCISES

1. Prove: $\mathbf{X}$ is $\mathcal{N}_n(0, I_n)$, where I_n is the $n \times n$ identity matrix, if and only if $X_1, \cdots, X_n$ are independent and each $\mathcal{N}(0,1)$.

2. Prove: If A is an $n \times n$ non-singular matrix, then

$$(A^t)^{-1} = (A^{-1})^t.$$

3. Prove:

$$(2\pi)^{-n/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{z}\mathbf{z}^t e^{-\frac{1}{2}\mathbf{z}^t\mathbf{z}} d\mathbf{z} = I_n.$$

4. Let A be an $n \times n$ symmetric non-singular matrix, and let P be an $n \times n$ orthogonal matrix such that P^tAP is a diagonal matrix, i.e., $P^tAP = (\lambda_i \delta_{ij})$, where $\lambda_1, \cdots, \lambda_n$ are the characteristic roots of A and δ_{ij} is the Kronecker delta. Prove that

$$A^{-1} = P(\lambda_i^{-1}\delta_{ij})P^t.$$

5. If $X, \cdots, X_n$ is multivariate normal, and if $\{k_1, \cdots, k_n\}$ is a permutation of the integers $\{1 \cdots, n\}$, then $X_{k_1}, \cdots, X_{k_n}$ is multivariate normal.

6. Prove: An n-dimensional random vector $\mathbf{X}$ has a multivariate normal distribution if and only if there exist a vector $\boldsymbol{\mu} \in R^n$, a positive definite (symmetric) $n \times n$ matrix A and independent random variables $Z_1, \cdots, Z_n$, each being $\mathcal{N}(0,1)$, such that $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$.

7. Verify the statements made about the example given after Theorem 3.

8. Let U, V be two independent $\mathcal{N}(0,1)$ random variables. Define X, Y by

$$X = U, Y = |V| I_{[U \geq 0]} - |V| I_{[U<0]}.$$

Find the joint density of X, Y.

9. Prove: If A is a positive definite $n \times n$ matrix, then

$$|(A^{1/2})^{-1}| = |(A^{-1})^{1/2}| = 1/\sqrt{|A|}.$$

10. Prove: If $Z_1, \cdots, Z_n$ are independent $\mathcal{N}(0,1)$ random variables, if P is an $n \times n$ orthogonal matrix, and if $W_1, \cdots, W_n$ is defined by

$$\mathbf{W} = P\mathbf{Z},$$

then $W_1, \cdots, W_n$ are independent and $\mathcal{N}(0,1)$.

11. Prove: If $X_1, \cdots, X_n$ are multivariate normal, if $c_1, \cdots, c_n$ are constants, not all zero, then $c_1 X_1 + \cdots + c_n X_n$ has a normal distribution.

12. Prove the converse of Problem 10: If $\mathbf{Z}$ is $\mathcal{N}_n(\mathbf{0}, I_n)$, and if P is an $n \times n$ matrix such that $\mathbf{W}$ defined by $\mathbf{W} = P\mathbf{Z}$ is $\mathcal{N}_n(0, I_n)$, then P is an orthogonal matrix.

13. Prove: If $\mathbf{X}$ is $\mathcal{N}_n(\boldsymbol{\mu}, C)$ and if $\mathbf{c} \in R^n, \mathbf{c} \neq \mathbf{0}$, then $\mathbf{c}^t\mathbf{X}$ is $\mathcal{N}(\mathbf{c}^t\boldsymbol{\mu}, \mathbf{c}^t C\mathbf{c})$.

14. Prove: If $\mathbf{X}$ is $\mathcal{N}_n(\boldsymbol{\mu}, C)$, then $X_1, \cdots, X_n$ are independent if and only if C is a diagonal matrix.

15. Let $\mathbf{X}$ be $\mathcal{N}_n(\boldsymbol{\mu}, C)$ and $\mathbf{Y}$ be $\mathcal{N}_n(\boldsymbol{\nu}, D)$, and assume that $\mathbf{X}$ and $\mathbf{Y}$ are independent. By the definition of multivariate normal distribution one can write $\mathbf{X} = A\mathbf{U} + \boldsymbol{\mu}$ and

$\mathbf{Y} = B\mathbf{V} + \nu$, where A and B are non-singular $n \times n$ matrices, $\mathbf{U}$ is $\mathcal{N}_n(\mathbf{0}, I_n)$ and $\mathbf{V}$ is $\mathcal{N}_n(\mathbf{0}, I_n)$. Prove that U and V are independent and

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \textit{ is } \mathcal{N}_{2n}(\mathbf{0}, I_{2n}).$$

16. If $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are independent $\mathcal{N}_k(\boldsymbol{\mu}, C)$ random vectors, and if $\boldsymbol{\alpha} \in R^k \setminus \{\mathbf{0}\}$, then $\boldsymbol{\alpha}^t\mathbf{X}_1, \cdots, \boldsymbol{\alpha}^t\mathbf{X}_n$ are independent $\mathcal{N}(\boldsymbol{\alpha}^t\boldsymbol{\mu}, \boldsymbol{\alpha}^t C\boldsymbol{\alpha})$ random variables.

17. Prove: If $\binom{X}{Y}$ is $\mathcal{N}_2(\boldsymbol{\mu}, C)$, then $Var(X) = Var(Y)$ if and only if $X - Y$ and $X + Y$ are independent.

§2. Conditional Densities and Conditional Expectations. We develop these topics in this section only for random variables which have a joint absolutely continuous distribution function, i.e., which have a joint density.

DEFINITION 1. *If $X_1, \cdots, X_m, Y_1, \cdots, Y_n$ are random variables with a joint absolutely continuous distribution function and joint density $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$, then we define the conditional density of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$ by*

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \begin{cases} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})/f_{\mathbf{Y}}(\mathbf{y}) & \text{if } f_{\mathbf{Y}}(\mathbf{y}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

PROPOSITION 1. *In Definition 1, the conditional density $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ is a density in $\mathbf{x}$ for every fixed $\mathbf{y}$ at which $f_{\mathbf{Y}}(\mathbf{y}) > 0$.*

Proof: Since $f_{\mathbf{Y}}(\mathbf{y}) > 0$, it follows that $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \geq 0$. We need only prove that $\int \underset{\mathbf{R}^m}{\cdots} \int f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})d\mathbf{x} = 1$. But

$$\begin{aligned} \underset{\mathbf{R}^m}{\int \cdots \int} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})d\mathbf{x} &= \underset{\mathbf{R}^m}{\int \cdots \int} \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}d\mathbf{x} \\ &= \frac{1}{f_{\mathbf{Y}}(\mathbf{y})} \underset{\mathbf{R}^m}{\int \cdots \int} f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})d\mathbf{x}, \end{aligned}$$

and the conclusion follows due to the fact that $\int \underset{\mathbf{R}^m}{\cdots} \int f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})d\mathbf{x} = f_{\mathbf{Y}}(\mathbf{y})$. Q.E.D.

PROPOSITION 2. *In Definition 1, $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})$ for all y at which $f_{\mathbf{Y}}(\mathbf{y}) > 0$.*

Proof: $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y})$ at all $\mathbf{x}, \mathbf{y}$, which is true if and only if

$$f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} = f_{\mathbf{X}}(\mathbf{x})$$

at all $\mathbf{x}$ and all $\mathbf{y}$ at which $f_{\mathbf{Y}}(\mathbf{y}) > 0$. Q.E.D.

The following corollary will be used frequently.

COROLLARY TO PROPOSITION 2. *In Proposition 2,* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent if and only if* $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ *does not depend on* $\mathbf{y}$ *for each fixed* $\mathbf{x}$.

Proof: If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then, by Proposition 2, $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})$ which does not depend on $\mathbf{y}$. Conversely, if $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ does not depend on $\mathbf{y}$, then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \int\limits_{\mathbf{R}^m}\!\!\cdots\!\int f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})d\mathbf{y} \\ &= \int\limits_{\mathbf{R}^m}\!\!\cdots\!\int f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} \\ &= f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})\int\limits_{\mathbf{R}^m}\!\!\cdots\!\int f_{\mathbf{Y}}(\mathbf{y})d\mathbf{y} \\ &= f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}), \end{aligned}$$

and thus by Proposition 2, $\mathbf{X}$ and $\mathbf{Y}$ are independent. Q.E.D.

DEFINITION 2. *If, in Definition 1,* $m = 1$, *we define the conditional expectation of* X_1 *given* $\mathbf{Y} = \mathbf{y}$ *by*

$$E(X_1|\mathbf{Y} = \mathbf{y}) = \int_{-\infty}^{\infty} x f_{X_1|\mathbf{Y}}(x|\mathbf{y})dx$$

at all $\mathbf{y} \in \mathbf{R}^n$ *at which* $f_{\mathbf{Y}}(\mathbf{y}) > 0$.

PROPOSITION 3. *If, in Definition 1, $m = 2$, and if $a \neq 0$ is a constant, then*

$$E(aX_1 + X_2|\mathbf{Y} = \mathbf{y}) = aE(X_1|\mathbf{Y} = \mathbf{y}) + E(X_2|\mathbf{Y} = \mathbf{y}).$$

Proof: Let $U_1 = aX_1 + X_2, U_2 = X_2$ and $\mathbf{V} = \mathbf{Y}$. Then by Theorem 1 in Section 1 we have

$$f_{U_1,U_2,\mathbf{V}}(u_1, u_2, \mathbf{v}) = f_{X_1,X_2,\mathbf{Y}}(\frac{1}{a}(u_1 - u_2), u_2, \mathbf{v})\frac{1}{|a|}.$$

Thus

$$\begin{aligned}
E(aX_1 + X_2|\mathbf{Y} = \mathbf{y}) &= E(U_1|\mathbf{V} = \mathbf{y}) \\
&= \int_{-\infty}^{\infty} u_1 f_{U_1|\mathbf{V}}(u_1|\mathbf{y})du_1 \\
&= \int_{-\infty}^{\infty} u_1 \frac{1}{f_{\mathbf{Y}}(\mathbf{y})}(\int_{-\infty}^{\infty} f_{U_1,U_2,\mathbf{V}}(u_1, u_2, \mathbf{y})du_2)du_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u_1 \frac{1}{f_Y(y)} f_{X_1,X_2,\mathbf{Y}}(\frac{1}{a}(u_1 - u_2), u_2, \mathbf{y})\frac{1}{|a|}du_2du_1.
\end{aligned}$$

Now make the following change of variables: $s = \frac{1}{a}(u_1 - u_2), t = u_2$. If this transformation is denoted by g, then $|\det g^{-1}(s,t)| = |a|$. Hence

$$\begin{aligned}
E(aX_1 + X_2|\mathbf{Y} = \mathbf{y}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(as + t)\frac{f_{X_1,X_2,\mathbf{Y}}(s, t, \mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}dsdt \\
&= a\int_{-\infty}^{\infty} sf_{X_1|\mathbf{Y}}(s|\mathbf{y})ds + \int_{-\infty}^{\infty} tf_{X_2}(t|\mathbf{y})dt \\
&= aE(X_1|\mathbf{Y} = \mathbf{y}) + E(X_2|\mathbf{Y} = \mathbf{y}).
\end{aligned}$$

Q.E.D.

DEFINITION 3. *Suppose $m = 1$ in Definition 1, and define $\varphi(\mathbf{y}) = E(X_1|\mathbf{Y} = \mathbf{y})$. We define the conditional expectation of X_1 given $\mathbf{Y}$, and denoted by $E(X_1|\mathbf{Y})$, by*

$$E(X_1|\mathbf{Y}) = \varphi(\mathbf{Y}).$$

It should be noted that $E(X_1|\mathbf{Y})$ is not a number as $E(X_1|\mathbf{Y} = \mathbf{y})$ was but is a function of $\mathbf{Y}$.

PROPOSITION 4. *If $m = 1$ in Definition 1, and if $E|X_1| < \infty$, then $E(E(X_1|\mathbf{Y})) = E(X_1)$.*

Proof: Let $\varphi(\cdot)$ be as defined in Definition 3. Then

$$\begin{aligned}
E(E(X_1|\mathbf{Y})) &= E\varphi(\mathbf{Y}) = \underset{\mathbf{R}^m}{\int\dots\int} \varphi(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \\
&= \underset{\mathbf{R}^m}{\int\dots\int} (\int_{-\infty}^{\infty} x f_{X|\mathbf{Y}}(x|\mathbf{y}) dx) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} \\
&= \int_{-\infty}^{\infty} x (\underset{\mathbf{R}^m}{\int\dots\int} f_{X_1|\mathbf{Y}}(x|\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}) dx \\
&= \int_{-\infty}^{\infty} x (\underset{\mathbf{R}^m}{\int\dots\int} f_{X_1,\mathbf{Y}}(x,\mathbf{y}) d\mathbf{y}) dx \\
&= \int_{-\infty}^{\infty} x f_{X_1}(x) dx = E(X_1).
\end{aligned}$$

Q.E.D.

PROPOSITION 5. *If X is a random variable and $\mathbf{Y}$ is an n-dimensional random vector, if X and $\mathbf{Y}$ are independent, and if $X, \mathbf{Y}$ have a joint absolutely continuous distribution function, then $E(X|\mathbf{Y} = y) = EX$ and $E(X|\mathbf{Y}) = E(X)$.*

Proof: By the hypothesis of independence, $f_{X,\mathbf{Y}}(x,\mathbf{y}) = f_X(x) f_{\mathbf{Y}}(\mathbf{y})$, which implies $f_{X|\mathbf{Y}}(x|\mathbf{y}) = f_X(x)$. Now by the definition,

$$\begin{aligned}
E(X|\mathbf{Y} = \mathbf{y}) &= \int_{-\infty}^{\infty} x f_{X|\mathbf{Y}}(x|\mathbf{y}) dx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx = E(X),
\end{aligned}$$

from which the conclusion follows. Q.E.D.

The proof of the next result is beyond the scope of this course.

PROPOSITION 6. *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent* $m-$ *and* n*- dimensional random vectors, if* $g(\mathbf{X},\mathbf{Y})$ *is a function of* $\mathbf{X}$ *and* $\mathbf{Y}$ *such that* $g(\mathbf{X},\mathbf{Y})$ *is a random variable and such that the joint distribution function of* $g(\mathbf{X},\mathbf{Y}),\mathbf{Y}$ *is absolutely continuous, and if* $E|g(\mathbf{X},\mathbf{Y})| < \infty$*, then* $E(g(\mathbf{X},\mathbf{Y})|\mathbf{Y}=\mathbf{y}) = Eg(\mathbf{X},\mathbf{y})$*.*

Although a proof of this proposition is beyond the scope of this course, we are able to motivate it by presenting a short proof of it in the case that the joint distribution function of $\mathbf{X},\mathbf{Y}$ is discrete. In this case,

$$\begin{aligned} E(g(\mathbf{X},\mathbf{Y})|\mathbf{Y}=\mathbf{y}) &= \sum_z zP([g(\mathbf{X},\mathbf{Y})=z]|[\mathbf{Y}=\mathbf{y}]) \\ &= \sum_z zP([g(\mathbf{X},\mathbf{y})=z]|[\mathbf{Y}=\mathbf{y}]). \end{aligned}$$

Since $\mathbf{X}$ and $\mathbf{Y}$ are independent, we obtain

$$\begin{aligned} E(g(\mathbf{X},\mathbf{Y})|\mathbf{Y}=\mathbf{y}) &= \sum_z zP[g(\mathbf{X},\mathbf{y})=z] \\ &= Eg(\mathbf{X},\mathbf{y}), \end{aligned}$$

which concludes the proof.

PROPOSITION 7. *If* X,Y *are random variables with a joint absolutely continuous distribution function, and if* $E|XY| < \infty$*, then* $E(XY|Y) = YE(X|Y)$*.*

Proof: Let us consider random variables $Z = XY$ and $W = Y$ and the corresponding

$1-1$ continuously differentiable mapping of $\mathbf{R}^2 \to \mathbf{R}^2$ defined by $z = xy, w = y$. Then $x = z/w, y = w$, and $|\det \frac{\partial(x,y)}{\partial(z,w)}| = \frac{1}{|w|}$ for $w \neq 0$. Hence $f_{Z,W}(x,w) = f_{X,Y}(z/w,w)\frac{1}{|w|}$, and

$$\begin{aligned} E(XY|Y=y) &= E(Z|W=y) = \int_{-\infty}^{\infty} z f_{Z|W}(z|y)dz \\ &= \frac{1}{|y|}\int_{-\infty}^{\infty} z\frac{f_{X,Y}(z/y,y)}{f_Y(y)}dz. \end{aligned}$$

Making the change of variable $z = xy$ in this integral (Remember: y is fixed.), we have

$$\begin{aligned} E(XY|Y=y) &= \frac{1}{|y|}\int_{-\infty}^{\infty} xy\frac{f_{X,Y}(x,y)}{f_Y(y)}|y|dx \\ &= y\int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx = yE(X|Y=y). \end{aligned}$$

Hence $E(XY|Y = y) = yE(X|Y = y)$, from which we obtain $E(XY|Y) = YE(X|Y)$. Q.E.D.

The following proposition is a generalization of Proposition 7. Its proof is beyond the scope of this course. The student should supply a proof of it in the discrete case.

PROPOSITION 8. *If X is a random variable and $\mathbf{Y}$ is a p-dimensional random vector, if $f : \mathbf{R}^p \to \mathbf{R}^1$ is a function such that $f(\mathbf{Y})$ is a random variable, and if $E(Xf(\mathbf{Y})) < \infty$, then $E(Xf(\mathbf{Y})|\mathbf{Y}) = f(\mathbf{Y})E(X|\mathbf{Y})$.*

The important property about $E(X|\mathbf{Y})$ is that it is that unique function of $\mathbf{Y}$ which minimizes $E((X - f(\mathbf{Y}))^2)$. Let us state this more precisely.

PROPOSITION 9. *If $X, \mathbf{Y}$ and f are as in Proposition 8, then*

$$E((X - E(X|\mathbf{Y}))^2) \leq E((X - f(\mathbf{Y}))^2).$$

Proof: We first observe that

$$\begin{aligned} E((X - f(\mathbf{Y}))^2) &= E(((X - E(X|\mathbf{Y})) + (E(X|\mathbf{Y}) - f(\mathbf{Y})))^2) \\ &= E((X - E(X|\mathbf{Y}))^2) + E((E(X|\mathbf{Y}) - f(\mathbf{Y}))^2) \\ &\quad +2E((X - E(X|\mathbf{Y}))(E(X|\mathbf{Y}) - f(\mathbf{Y}))). \end{aligned}$$

Note that $E(X|\mathbf{Y}) - f(\mathbf{Y})$ is a function of $\mathbf{Y}$. Hence by Propositions 4 and 8 we have

$$\begin{aligned} E((X - E(X|\mathbf{Y}))(E(X|\mathbf{Y}) - f(\mathbf{Y}))) &= E(E((X - E(X|\mathbf{Y}))(E(X|\mathbf{Y}) - f(\mathbf{Y}))|\mathbf{Y})) \\ &= E((E(X|\mathbf{Y}) - f(\mathbf{Y}))E((X - E(X|\mathbf{Y}))|\mathbf{Y})) \\ &= E((E(X|\mathbf{Y}) - f(\mathbf{Y}))\{E(X|\mathbf{Y}) - E(E(X|\mathbf{Y})|\mathbf{Y})\}). \end{aligned}$$

By Proposition 8,

$$E(E(X|\mathbf{Y})|\mathbf{Y}) = E(X|\mathbf{Y}).$$

Thus $E((X - f(\mathbf{X}))^2 = E((X - E(X|\mathbf{Y}))^2) + E((E(X|\mathbf{Y}) - f(\mathbf{Y}))^2 \geq E((X - E(X|Y))^2)$.

Q.E.D.

EXERCISES

1. Let X, X, Z be independent random variables, each one being $\mathcal{N}(0,1)$. Find $E(X^2 + Y^2 + Z^2 | Z = z)$.

2. Let X, Y be random variables whose joint distribution is uniform over the unit disk in $\mathbf{R}^2$, i.e., their joint density is

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{if } x^2 + y^2 > 1 \end{cases}.$$

(i) Prove that X and Y are not independent.

(ii) Determine $f_{X|Y}(x|y)$ when $-1 < y < 1$.

(iii) Determine the univariate marginal densities $f_X(X)$ and $f_Y(y)$.

(iv) Determine $E(X|Y = y)$ when $-1 < y < 1$.

(v) Determine $E(X^2 + Y^2 | Y = y)$ when $-1 < y < 1$.

§3. **Regression and Independence.** We now pursue a deeper study of the multivariate normal distribution.

THEOREM 1. *Let* $\mathbf{X}$ *be* $\mathcal{N}_{p+q}(\mu, \sum)$, *where* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} \end{pmatrix}$ *and*

$$\sum = \left(\begin{array}{c|c} \sum_{11} & \sum_{12} \\ \hline \sum_{21} & \sum_{22} \end{array}\right)$$

where $\mathbf{X}_{(1)}$ *and* $\mu_{(1)}$ *are p-dimensional, and where* $\sum_{11}$ *is* $p \times p$. *Define* $\sum_{22.1} = \sum_{22} - \sum_{21} \sum_{11}^{-1} \sum_{12}$. *Then* $\mathbf{X}_{(1)}$ *and* $\mathbf{X}_{(2)} - \sum_{21} \sum_{11}^{-1} \mathbf{X}_{(1)}$ *are independent, and their distributions are* $\mathcal{N}_p(\mu_{(1)}, \sum_{11})$ *and* $\mathcal{N}_q(\mu_{(2)} - \sum_{21} \sum_{11}^{-1} \mu_{(1)}, \sum_{22.1})$ *respectively.*

Proof: Define

$$C = \left(\begin{array}{c|c} I_p & 0 \\ \hline -\sum_{21} \sum_{11}^{-1} & I_q \end{array}\right)$$

Clearly C is nonsingular. Next, define

$$\mathbf{Y} = C\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} - \sum_{21} \sum_{11}^{-1} \mathbf{X}_{(1)} \end{pmatrix}.$$

Then $\mathbf{Y}$ is $\mathcal{N}_{p+q}(C\mu, C \sum C^t)$. However,

$$C\mu = \begin{pmatrix} \mu_{(1)} \\ \mu_{(2)} - \sum_{21} \sum_{11}^{-1} \mu_{(1)} \end{pmatrix} \text{ and } C \sum C^t = \left(\begin{array}{c|c} \sum_{11} & 0 \\ \hline 0 & \sum_{22.1} \end{array}\right)$$

which yields our conclusions. Q.E.D.

Next, we look at conditional density.

THEOREM 2. *If* $\mathbf{X}$ *is as in Theorem 1, and using the same notation as in Theorem 1, then the conditional density of* $\mathbf{X}_{(2)}$*, given* $\mathbf{X}_{(1)} = \mathbf{x}_{(1)}$*, is the same as the density of a random vector which is* $\mathcal{N}_q(\boldsymbol{\mu}_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}), \Sigma_{22.1})$.

Proof: We first recall that $f_{\mathbf{X}_{(2)}|\mathbf{X}_{(1)}}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)}) = f_{\mathbf{X}}(\mathbf{x})/f_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)})$. Now $f_{\mathbf{X}_{(1)}}(\mathbf{x}_{(1)}) = \mathit{const}\exp -\frac{1}{2}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^t\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})$. Let $\mathbf{Y} = C\mathbf{X}$, where C is as defined in the proof of Theorem 1. Then by Theorem 1,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}_{(1)}, \mathbf{y}_{(2)}) &= \mathit{const}\exp -\frac{1}{2}\{(\mathbf{y}_{(1)} - \boldsymbol{\mu}_{(1)})^t\Sigma_{11}^{-1}(\mathbf{y}_{(1)} - \boldsymbol{\mu}_{(1)}) \\ &\quad + (\mathbf{y}_{(2)} - \boldsymbol{\mu}_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{(1)})^t\Sigma_{22.1}^{-1}(\mathbf{y}_{(2)} - \boldsymbol{\mu}_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}\boldsymbol{\mu}_{(1)})\}. \end{aligned}$$

Then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{Y}}(C\mathbf{x})|detC| \\ &= \mathit{const}\exp -\frac{1}{2}\{(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)})^t\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}) \\ &+ (\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}))^t\Sigma_{22.1}^{-1}(\mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}))\}. \end{aligned}$$

Taking the quotient given above we obtain

$$f_{\mathbf{X}_2|\mathbf{X}_{(1)}}(\mathbf{x}_{(2)}|\mathbf{x}_{(1)}) = \mathit{const}\exp -\frac{1}{2}\boldsymbol{\zeta}^t\Sigma_{22.1}^{-1}\boldsymbol{\zeta},$$

where

$$\boldsymbol{\zeta} = \mathbf{x}_{(2)} - \boldsymbol{\mu}_{(2)} - \sum_{21}\Sigma_{11}^{-1}(\mathbf{x}_{(1)} - \boldsymbol{\mu}_{(1)}).$$

Q.E.D.

COROLLARY 1. *If* $\binom{X}{Y}$ *is* $\mathcal{N}_2\left(\binom{\mu}{\nu}, \Sigma\right)$*, then the conditional density of* Y *given* $X = x$ *is the same as the density of a random variable which is* $\mathcal{N}(EY - \frac{s.d.Y}{s.d.X}\rho_{x,y}(x - \mu), \sigma^2)$, *where s.d.$X$ means standard deviation of $X, \rho_{X,Y}$ is the correlation coefficient of X, Y and* $\sigma^2 = var(Y) - (Cov(X,Y))^2/varX$.

DEFINITION: If $X_1, \cdots, X_p$ are random variables with a joint absolutely continuous distribution, and if $E|X_1| < \infty$, then $E(X_1|X_2, \cdots, X_p)$ is called the regression of X_1 on $X_2, \cdots, X_p$. Also, $E(X_1|X_2 = x_2, \cdots, X_p = x_p)$ is called the regression of X_1 on $X_2 = x_2, \cdots, X_p = x_p$.

Re: Notation: If $\mathbf{X}$ and $\mathbf{Y}$ are $p-$ and $q-$dimensional random vectors with joint absolutely continuous distribution, then we denote

$$E(\mathbf{X}|\mathbf{Y}) = \begin{pmatrix} E(X_1|\mathbf{Y}) \\ \vdots \\ E(X_p|\mathbf{Y}) \end{pmatrix} \text{ and } E(\mathbf{X}|\mathbf{Y} = \mathbf{y}) = \begin{pmatrix} E(X_1|\mathbf{Y} = \mathbf{y}) \\ \vdots \\ E(X_p|\mathbf{Y} = \mathbf{y}) \end{pmatrix}.$$

THEOREM 3. *If* $\binom{\mathbf{Y}}{\mathbf{X}}$ *is* $\mathcal{N}_{p+q}(\binom{\nu}{\mu}, \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$ *where* $\dim Y = \dim \nu = rank \sum_{11} = p$, *then the regression of* $\mathbf{Y}$ *on* $\mathbf{X}$ *is linear, i.e.,* $E(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \nu + \sum_{12}\sum_{22}^{-1}(\mathbf{x} - \mu)$, *and* $E(\mathbf{Y}|\mathbf{X}) = \nu + \sum_{12}\sum_{22}^{-1}(\mathbf{X} - \mu)$.

Proof: This follows from the definition of $E(\mathbf{Y}|\mathbf{X})$ and Theorem 2. Q.E.D.

A special case of Theorem 3 is this: if $\binom{Y}{X}$ is $\mathcal{N}_2\left(\binom{EY}{EX}, \sum\right)$, then $\sum_{22} = Var(X), \sum_{12} = Cov(X,Y) = \sigma_X\sigma_Y\rho_{XY}$ and $\sum_{11} = var(Y)$. Hence

$$E(Y|X) = E(Y) + \frac{cov(X,Y)}{var(X)}(X - E(X)) = E(Y) + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(X - E(X)).$$

EXERCISES

1. Prove: If $\binom{X}{Y}$ is $\mathcal{N}_2\left(\binom{0}{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, then $|\rho| < 1$.

2. Determine a and $B = (b_{ij})$ in the following statement: if $\binom{X}{Y}$ is $\mathcal{N}_2(\mu, C)$, where $C = (c_{ij})$ and $c_{11}c_{22} - c_{12}c_{21} > 0$, then for every fixed value of y, the conditional density $f_{X|Y}(x|y)$ is (in x) the density of a random variable which is $\mathcal{N}(a, B)$.

3. If $\mathbf{Z}$ is an n-dimensional random vector with absolutely continuous joint distribution, and if its density is $f_{\mathbf{Z}}(\boldsymbol{\zeta}) = K \exp -\frac{1}{2}(\boldsymbol{\zeta} - \boldsymbol{\alpha})^t\Gamma(\boldsymbol{\zeta} - \boldsymbol{\alpha})$, all $\boldsymbol{\zeta} \in \mathbf{R}^n$, where Γ is a positive definite $n \times n$ matrix, prove that $K = \sqrt{det\Gamma}/(2\pi)^{n/2}$.

§4. **(Random) Orthogonal Matrices.** We explore independence further through the use of (random) orthogonal matrices.

LEMMA 1. *If $X_1, \cdots, X_n$ are independent $\mathcal{N}(0, \sigma^2)$ random variables, if A is an $n \times n$ orthogonal matrix, and if $\mathbf{Y} = A\mathbf{X}$, then $Y_1, \cdots, Y_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.*

Proof: Since A is non-singular, and since $\mathbf{X}$ is $\mathcal{N}_n(0, \sigma^2 I_n)$, then $A\mathbf{X}$ is $\mathcal{N}_n(A0, \sigma^2 A I_n A^t) = \mathcal{N}_n(0, \sigma^2 I_n)$. Q.E.D.

THEOREM 1. *Let $Z_1, \cdots, Z_n$ be i.i.d. $\mathcal{N}(O, \sigma^2)$ random variables, and let $\mathbf{P} = (P_{ij})$ be an $n \times n$ matrix of random variables which is an orthogonal matrix with probability 1, i.e., $P[\mathbf{P}^t\mathbf{P} = I_n] = 1$, and such that the joint distribution function of the P_{ij}'s is absolutely continuous. Assume that $\mathbf{Z}$ and $\mathbf{P}$ are independent, and let $\mathbf{W} = \mathbf{PZ}$. Then $W_1, \cdots, W_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.*

Proof: Let us denote $[\mathbf{W} \leq \mathbf{w}] = \cap_{j=1}^n [W_j \leq w_j]$. Then, letting $\mathcal{P}$ denote the set of all $n \times n$ orthogonal matrices, we have

$$\begin{aligned} P[\mathbf{W} \leq \mathbf{w}] &= P[\mathbf{PZ} \leq \mathbf{w}] = E(E(I_{[\mathbf{PZ} \leq \mathbf{w}]} | \mathbf{P})) \\ &= \int_{\mathcal{P}} E(I_{[\mathbf{PZ} \leq \mathbf{w}]} | \mathbf{P} = p) f_{\mathbf{P}}(p) dp. \end{aligned}$$

Now by Proposition 6 of Section 2, and by Lemma 1 above, it follows that

$$\begin{aligned} P[\mathbf{W} \leq \mathbf{w}] &= \int_{\mathcal{P}} E(I_{[p\mathbf{Z} \leq \mathbf{w}]}) f_{\mathbf{P}}(p) dp \\ &= \int_{\mathcal{P}} P[\mathbf{Z} \leq \mathbf{w}] f_{\mathbf{P}}(p) dp \end{aligned}$$

$$\begin{aligned} &= \int_{\mathcal{P}} \left(\frac{1}{(2\pi\sigma^2)^{n/2}} \int_{-\infty}^{w_1} \cdots \int_{-\infty}^{w_n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} t_i^2} dt_n \cdots dt_1 \right) f_{\mathbf{P}}(p)dp \\ &= (2\pi\sigma^2)^{-n/2} \int_{-\infty}^{w_n} \cdots \int_{-\infty}^{w_1} (\exp -\frac{1}{2\sigma^2} \sum_{j=1}^{n} t_j^2) dt_n \cdots dt_1. \end{aligned}$$

Q.E.D.

THEOREM 2. *If* $\mathbf{U}$ *is* $\mathcal{N}_n(0, \sigma^2 I_n)$*, if* P_0 *is a* $k \times n$ *matrix with* $k < n$ *and such that* $P_0 P_0^t = I_k$*, and if* $\mathbf{V} = P_0\mathbf{U}$*, then* $\mathbf{V}$ *and* $\frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$ *are independent, and* $\frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$ *has the chi-square distribution with* $n - k$ *degrees of freedom.*

Proof: Let P_1 be an $(n-k) \times n$ matrix such that $\binom{P_0}{P_1}$ is an orthogonal matrix. (Such a matrix always exists.) Let $\mathbf{T} = P_1\mathbf{U}$. Then $\begin{pmatrix} \mathbf{V} \\ \vdots \\ \mathbf{T} \end{pmatrix} = \begin{pmatrix} P_0\mathbf{U} \\ \vdots \\ P_1\mathbf{U} \end{pmatrix} = \begin{pmatrix} P_0 \\ \vdots \\ P_1 \end{pmatrix} \mathbf{U}$. Thus $\mathbf{V}$ and $\mathbf{T}$ are independent by Lemma 1. But

$$\mathbf{U}^t\mathbf{U} = \mathbf{U}^t(P_0^t \vdots P_1^t) \begin{pmatrix} P_0 \\ \vdots \\ P_1 \end{pmatrix} \mathbf{U} = (\mathbf{V}^t \vdots \mathbf{T}^t) \begin{pmatrix} \mathbf{V} \\ \mathbf{T} \end{pmatrix} = \mathbf{V}^t\mathbf{V} + \mathbf{T}^t\mathbf{T}.$$

Now $\frac{1}{\sigma^2}\mathbf{T}^t\mathbf{T}$ is the sum of $n-k$ squares of independent $\mathcal{N}(0,1)$ random variables and thus has the χ^2_{n-k}-distribution. But $\frac{1}{\sigma^2}\mathbf{T}^t\mathbf{T} = \frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$. Q.E.D.

COROLLARY TO THEOREM 2: *If* $X_1, \cdots, X_n$ *are independent* $\mathcal{N}(\mu, \sigma^2)$ *random variables, and if* $\bar{X} = (X_1 + \cdots + X_n)/n$ *and* $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2$*, then* $\bar{X}$ *and* s^2 *are independent, and* $\frac{n-1}{\sigma^2}s^2$ *has the* χ^2_{n-1}*-distribution.*

Proof: Without loss of generality we take $\mu = 0$. In Theorem 2, let $k = 1$, and let $P_0 = \left(\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}}\right)$. Then let $\mathbf{U} = \mathbf{X}$ and $V = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_j$. By Theorem 2, $\frac{1}{\sigma^2}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2) = \frac{n-1}{\sigma^2}s^2$ and $\bar{X}$ are independent, and $\frac{n-1}{\sigma^2}s^2$ has the χ^2_{n-k}-distribution. Q.E.D.

We next wish to extend Theorem 2 to the case where P_0 is a $k \times n$ matrix of random variables which satisfy $P_0(\omega)P_0(\omega)^t = I_k$ for all $\omega \in \Omega$.

LEMMA 2. *Let $\{U_{ij}, 1 \le i \le k, 1 \le j \le n\}$ be kn random variables $(1 \le k < n)$ such that the $k \times n$ random matrix $U = (U_{ij})$ satisfies $UU^t = I_k$. Then there exist $n(n-k)$ random variables $\{V_{ij} : k+1 \le i \le n, 1 \le j \le n\}$ which are Borel-measurable functions of the U_{ij}'s such that if $V = (V_{ij})$, then $\begin{pmatrix} U \\ \cdots \\ V \end{pmatrix}$ is a (random) orthogonal matrix.*

Proof: Let $\mathbf{e}_1, \cdots, \mathbf{e}_n$ be any basis of $\mathbf{R}^n$, and let $\mathcal{S}$ denote the set of all subsets of size $n-k$ of $\{1, 2, \cdots, n\}$. There are $\binom{n}{k}$ elements of $\mathcal{S}$, and we may assign to these an arbitrary fixed ordering; let us denote them by $s_1, \cdots, s_r$, where $r = \binom{n}{k}$. Let us denote $\mathbf{U}_i = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{in} \end{pmatrix}$, i.e., $\mathbf{U}_i$ is the transpose of the ith row of the matrix U. The hypothesis $UU^t = I_k$ implies that the vectors $\mathbf{U}_1(\omega), \cdots, \mathbf{U}_k(\omega)$ form an orthonormal set of vectors for every $\omega \in \Omega$, and hence are linearly independent. Hence, for every $\omega \in \Omega$ there exists $s_j = \{j_1, \cdots, j_{n-k}\} \in \mathcal{S}$ such that the vectors $\mathbf{U}_1(\omega), \cdots, \mathbf{U}_k(\omega), \mathbf{e}_{j_1} \cdots \mathbf{e}_{j_{n-k}}$ form a basis of $\mathbf{R}^n$. Thus, if we let $A_j = \{\omega \in \Omega : U_1(\omega), \cdots, U_k(\omega), \{\mathbf{e}_i : i \in s_j\}$ is a basis of $\mathbf{R}^n\}$, then $\Omega = \cup_{j=1}^{r} A_j$. We observe that A_j is an event in the sigma algebra Ω and is actually *measurable* with respect to the sub-sigma-algebra generated by U; indeed, if $s_j = \{j_1, \cdots, j_{n-k}\}$,then $A_j = \{\omega \in \Omega : \det(\mathbf{U}_1(\omega) \vdots \cdots \vdots \mathbf{U}_k(\omega) \vdots \mathbf{e}_{j_1} \vdots \cdots \vdots \mathbf{e}_{j_{n-k}}) \neq 0\}$. Next, define $B_1 = A_1, B_2 = A_2 \setminus A_1$ and, in general, $B_q = A_q \setminus \left(\cup_{i=1}^{q-1} A_i\right)$ for $2 \le q \le r$. Now each B_j is also in the sub-sigma algebra generated by U, i.e., I_{B_i} is a Borel-measurable function of U. It remains for us to define V over each B_q. If $s_q = \{j_1, \cdots, j_{n-k}\}$, consider over B_q the vector functions

$\{\mathbf{U}_1, \cdots, \mathbf{U}_k, \mathbf{e}_{j_1} . \cdots, \mathbf{e}_{j_{n-k}}\}$: For every $\omega \in B_q$, the first k vectors are orthonormal. One may define $\mathbf{V}_1^{(q)}, \cdots, \mathbf{V}_{n-k}^{(q)}$ by the Gram-Schmidt process, and thus, every $\mathbf{V}_j^{(q)}$ is a Borel-measurable function of $\mathbf{U}_1, \cdots, \mathbf{U}_k$ *over* B_q. Now define $\mathbf{V}_j = \sum_{q=1}^{r} \mathbf{V}_j^{(q)} I_{B_q}$ and then $V = (\mathbf{V}_1 \vdots \cdots \vdots \mathbf{V}_{n-k})^t$. Q.E.D.

THEOREM 3. *If, in Theorem 2,* P_0 *is, in addition, a random matrix, and if* P_0 *and* $\mathbf{U}$ *are independent, then* $\mathbf{V} = P_0\mathbf{U}$ *and* $\frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$ *are independent, and* $\frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$ *has the* χ^2_{n-k}*-distribution.*

Proof: By Lemma 2 there exists an $(n-k) \times n$ matrix P_1 whose elements are Borel-measurable functions of those in P_0 and such that $\begin{pmatrix} P_0 \\ \cdots \\ P_1 \end{pmatrix} (P_0^t \vdots P_1^t) = I_n$, i.e., $\begin{pmatrix} P_0 \\ \cdots \\ P_1 \end{pmatrix}$ is a random orthogonal matrix. Thus if we define $\mathbf{V} = P_0\mathbf{U}$ and $\mathbf{T} = P_1\mathbf{U}$ and recall Theorem 1, we obtain that the coordinates of $\begin{pmatrix} \mathbf{V} \\ \mathbf{T} \end{pmatrix}$ are independent and $\mathcal{N}(0, \sigma^2)$. Denote $\mathbf{Z} = \begin{pmatrix} \mathbf{V} \\ \cdots \\ \mathbf{T} \end{pmatrix}$ and $Q = \begin{pmatrix} P_0 \\ \cdots \\ P_1 \end{pmatrix}$. Then $\mathbf{Z} = Q\mathbf{U}$. Now $\mathbf{V}$ and $\mathbf{T}$ being independent implies $\mathbf{V}$ and $\mathbf{T}^t\mathbf{T}$ are independent. But

$$\begin{aligned} \mathbf{T}^t\mathbf{T} &= \mathbf{Z}^t\mathbf{Z} - \mathbf{V}^t\mathbf{V} = \mathbf{U}^t Q^t Q \mathbf{U} - \mathbf{V}^t\mathbf{V} \\ &= \mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V}, \end{aligned}$$

and thus $\mathbf{V}$ and $\frac{1}{\sigma^2}(\mathbf{U}^t\mathbf{U} - \mathbf{V}^t\mathbf{V})$ are independent. Q.E.D.

EXERCISES

1. Prove: if $\mathbf{x}_1, \cdots, \mathbf{x}_k$ are a set of orthonormal vectors in $\mathbf{R}^n$, then they are linearly independent.

2. Prove: If M is an $n \times n$ orthogonal matrix, and if the last row of M is $(1/\sqrt{n}, \cdots, 1/\sqrt{n})$, then the sum of numbers in each row is zero.

3. Prove: If $\mathbf{X}$ is $\mathcal{N}_n(\mu, \sum)$, then there is an orthogonal $n \times n$ matrix P such that the coordinates of $P\mathbf{X}$ are independent random variables.

CHAPTER 2. THE WISHART DISTRIBUTION.

§1. Samples on a Normal Population. Before we begin our discussion of the Wishart distribution, we shall find maximum likelihood estimates of μ and $\sum$ based on a sample of size n on a $\mathcal{N}_p(\mu, \sum)$ population. More precisely, let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d. $\mathcal{N}_p(\mu, \sum)$, where $n > p$. For $1 \leq r \leq n$, we denote

$$\mathbf{X}_r = \begin{pmatrix} \mathbf{X}_{1r} \\ \vdots \\ X_{pr} \end{pmatrix},$$

and $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n) = (X_{ij})$, a $p \times n$ matrix. We further denote

$$\begin{aligned} \bar{\mathbf{X}}_{i\cdot} &= \frac{1}{n} \sum_{j=1}^{n} X_{ij}, \\ s_{ij} &= \sum_{r=1}^{n} (X_{ir} - \bar{X}_{i\cdot})(X_{jr} - \bar{X}_{j\cdot}), \end{aligned}$$

and $S = (s_{ij})$. Let $\mathbf{1}_n$ denote an n-dimensional vector consisting of all 1's. One can easily show that $\frac{1}{n-1} S$ is an unbiased estimate of $\sum$, i.e., $E\left(\frac{1}{n-1} S\right) = \sum$. Let us define $\bar{\mathbf{X}} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}_j$; then $\bar{\mathbf{X}} = \frac{1}{n} X \mathbf{1}_n = \begin{pmatrix} \bar{X}_{1\cdot} \\ \vdots \\ \bar{X}_{p\cdot} \end{pmatrix}$.

REMARK 1: $S = X(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t) X^t$.

Proof: We first note that $X I_n X^t = X X^t = (\sum_{r=1}^{n} X_{ir} X_{jr})$. Also

$$\begin{aligned} \frac{1}{n} X \mathbf{1}_n \mathbf{1}_n^t X^t &= n \left(\frac{1}{n} X \mathbf{1}_n \right) \left(\frac{1}{n} X \mathbf{1}_n \right)^t = n \bar{\mathbf{X}} \bar{\mathbf{X}}^t \\ &= (n \bar{\mathbf{X}}_{i\cdot} \bar{\mathbf{X}}_{j\cdot}^t), \end{aligned}$$

and thus $X(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^t) X^t = (s_{ij}) = S$. Q.E.D.

Recall that the trace of a square $n \times n$ matrix $A = (a_{ij})$ is defined by $tr(A) = \sum_{i=1}^{n} a_{ii}$. One property of trace is: if A and B are $n \times n$ matrices, then $tr(A+B) = tr(A) + tr(B)$. Another property is: If A is an $m \times n$ matrix, and if B is an $n \times m$ matrix, then $tr(AB) = tr(BA)$.

REMARK 2: *If X is as above, and if $x = (\mathbf{x}_1 \vdots \cdots \vdots \mathbf{x}_n)$ is a $p \times n$ matrix of real numbers, i.e., $x \in \mathbf{R}^{p \times n}$, then*

$$f_X(x) = \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} \sum_{\alpha=1}^{n} tr\{\Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t\}.$$

Proof: For $1 \leq i \leq n$,

$$f_{\mathbf{X}_i}(\mathbf{x}_i) = \frac{\sqrt{\det \Sigma^{-1}}}{(2\pi)^{n/2}} \exp -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}).$$

and thus

$$\begin{aligned} f_X(x) &= \prod_{\alpha=1}^{n} f_{\mathbf{X}_\alpha}(\mathbf{x}_\alpha) \\ &= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}). \end{aligned}$$

But by a property of trace,

$$\begin{aligned} (\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu}) &= tr\{(\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})\} \\ &= tr\{\Sigma^{-1}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t\} \end{aligned}$$

for $1 \leq \alpha \leq n$. Substituting this into the formula for $f_X(x)$ above we obtain the result. Q.E.D.

Now, for every $x = (x_{ij}) \in \mathbf{R}^{p \times n}$, let us denote $\bar{x}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$, $\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_{i\cdot}$ and $A = (a_{ij}) = \sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^t$. One can easily verify that $a_{ij} = \sum_{\alpha=1}^{n}(x_{i\alpha} - \bar{x}_{i\cdot})(x_{j\alpha} - \bar{x}_{j\cdot})$.

LEMMA 1. *For every* $\mathbf{b} \in \mathbf{R}^p$,

$$\sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \mathbf{b})(\mathbf{x}_\alpha - \mathbf{b})^t = A + n(\bar{\mathbf{x}} - \mathbf{b})(\bar{\mathbf{x}} - \mathbf{b})^t.$$

Proof: We observe that

$$\begin{aligned}\sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \mathbf{b})(\mathbf{x}_\alpha - \mathbf{b})^t &= \sum_{\alpha=1}^{n} ((\mathbf{x}_\alpha - \bar{\mathbf{x}}) - (\mathbf{b} - \bar{\mathbf{x}}))((\mathbf{x}_\alpha - \bar{\mathbf{x}}) - (\mathbf{b} - \bar{\mathbf{x}}))^t \\ &= \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^t + n(\bar{\mathbf{x}} - \mathbf{b})(\bar{\mathbf{x}} - \mathbf{b})^t \\ &= A + n(\bar{\mathbf{x}} - \mathbf{b})(\bar{\mathbf{x}} - \mathbf{b})^t.\end{aligned}$$

Q.E.D.

LEMMA 2. *The following identity holds:*

$$\sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_\alpha - \boldsymbol{\mu}) = tr\{\Sigma^{-1} A\} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})^t \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}).$$

Proof: Using the properties of trace, we obtain

$$\begin{aligned}\sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_\alpha - \boldsymbol{\mu}) &= \sum_{\alpha=1}^{n} tr\{(\mathbf{x}_\alpha - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_\alpha - \boldsymbol{\mu})\} \\ &= \sum_{\alpha=1}^{n} tr\{\Sigma^{-1} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t\} \\ &= tr\{\Sigma^{-1} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t\}.\end{aligned}$$

Applying Lemma 1, the above becomes

$$\begin{aligned}&= tr\{\Sigma^{-1} A + n\Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^t\} \\ &= tr\{\Sigma^{-1} A\} + ntr\{\Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})^t\} \\ &= tr\{\Sigma^{-1} A\} + n(\bar{\mathbf{x}} - \boldsymbol{\mu})^t \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}).\end{aligned}$$

Q.E.D.

LEMMA 3. *Whatever the value of $\sum$, the density of X is maximized when $\mu = \bar{\mathbf{x}}$.*

Proof: Applying Lemma 2 to the formula for the density of X given in Remark 2, noting that $tr\{\sum^{-1} A\}$ does not depend on μ, and making use of the fact that $\sum$ being positive definite implies that $\sum^{-1}$ is positive definite, it follows that $n(\bar{\mathbf{x}} - \mu)^t \sum^{-1}(\bar{\mathbf{x}} - \mu) \geq 0$ for all μ. It is zero only when $\mu = \bar{\mathbf{x}}$. Q.E.D.

LEMMA 4. *If A is a symmetric $n \times n$ matrix, and if all its eigenvalues are equal to K, then $A = KI_n$.*

Proof: A being symmetric implies that there is an $n \times n$ orthogonal matrix P such that $P^t AP$ is a diagonal matrix. All the diagonal elements are eigenvalues, and hence $P^t A\mathrm{P} = KI_n$. Thus $A = KI_n$. Q.E.D.

LEMMA 5. *If Q is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$, then $tr\, Q = \sum_{i=1}^n \lambda_i$.*

Proof: There exists an orthogonal $n \times n$ matrix P such that $P^t QP = diag\{\lambda_1, \cdots, \lambda_n\}$. But $\sum_{i=1}^n \lambda_i = tr\{P^t QP\} = tr\{QPP^t\} = tr\, Q$. Q.E.D.

LEMMA 6. *If $\mathcal{Q}$ denotes the set of all positive-definite $n \times n$ matrices, and if $f : \mathcal{Q} \to \mathbf{R}^1$ is a function defined by*

$$f(Q) = N \log \det Q - trQ, \;\; all\, Q \in \mathcal{Q},$$

where $N > 0$, then f achieves its unique maximum over $\mathcal{Q}$ at $Q = NI_n$.

Proof: Let $Q \in \mathcal{Q}$, and let $\lambda_1, \cdots, \lambda_n$ denote the (necessarily real) eigenvalues of Q. Since Q is positive-definite, then all $\lambda_i > 0$, and by Lemma 5, $tr\, Q = \sum_{i=1}^{n} \lambda_i$. Also $\det Q = \prod_{j=1}^{n} \lambda_j$. Thus

$$f(Q) = N \log \left(\prod_{i=1}^{n} \lambda_i \right) - \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} (N \log \lambda_i - \lambda_i).$$

Now $N \log \lambda_i - \lambda_i$ is maximized at λ_i for which $(N/\lambda_i) - 1 = 0$, i.e., when $\lambda_i = N$ for all i. Hence f is maximized at any $Q \in \mathcal{Q}$ for which all its eigenvalues are equal to N. By Lemma 4, there is only one such $Q \in \mathcal{Q}$, namely, $Q = NI_n$. Q.E.D.

LEMMA 7. *Let $\mathcal{Q}$ be as in Lemma 6, and let $D \in \mathcal{Q}$ be fixed. If, for each $C \in \mathcal{Q}, f(C)$ is defined by*

$$f(C) = \frac{1}{2} N \log \det C - \frac{1}{2} tr(CD),$$

then f achieves exactly one maximum value over $\mathcal{Q}$, namely at $C = ND^{-1}$, and this maximum is $f(ND^{-1}) = \frac{1}{2} Nn \log N - \frac{1}{2} N \log \det D - \frac{1}{2} nN$.

Proof: There exists a positive-definite symmetric matrix $D^{1/2}$ such that $D = D^{1/2} D^{1/2}$. By the properties of determinants and trace, we have

$$\begin{aligned} f(C) &= \frac{1}{2} N \log \frac{\det(D^{1/2} C D^{1/2})}{\det D} - \frac{1}{2} tr(D^{1/2} C D^{1/2}) \\ &= -\frac{1}{2} N \log \det D + \frac{1}{2} N \log \det(D^{1/2} C D^{1/2}) - \frac{1}{2} tr(D^{1/2} C D^{1/2}). \end{aligned}$$

Now $-\frac{1}{2} N \log \det D$ does not depend on C, so by Lemma 6, f achieves its unique maximum when $D^{1/2}CD^{1/2} = NI_n$, or $C = ND^{-1}$. Moreover,

$$\begin{aligned} f(ND^{-1}) &= \frac{1}{2}N\log\det(ND^{-1}) - \frac{1}{2}tr(ND^{-1}D) \\ &= \frac{1}{2}N\log(N^n\det D^{-1}) - \frac{1}{2}tr(NI) \\ &= \frac{1}{2}Nn\log N - \frac{1}{2}N\log\det D - \frac{1}{2}nN. \end{aligned}$$

Q.E.D.

LEMMA 8. *Let $Z = (Z_{ij})$ be an $n \times n$ matrix of independent random variables, each having an absolutely continuous distribution function. Then* $P[\det Z \neq 0] = 1$.

Proof: We prove this by induction on n. The lemma is clearly true for $n = 1$. Assuming it to be true for $n - 1$ where $n \geq 2$, we shall prove it true for n. Let Q_{ij} denote the cofactor of Z_{ij} in Z, and consider the following change of variable:

$$\begin{aligned} W_{11} &= Z_{11}Q_{11} + Z_{12}Q_{12} + \cdots + Z_{1n}Q_{1n} \\ W_{12} &= Z_{12} \\ &\vdots \\ W_{nn} &= Z_{nn} \end{aligned}$$

By induction hypothesis, $P[Q_{11} = 0] = 0$. One should note that $Q_{11}, \cdots, Q_{1n}$ do not depend on Z_{11}. Hence we can solve for $Z_{11}, \cdots, Z_{nn}$ in terms of $W_{11}, \cdots, W_{nn}$ and obtain the absolute value of the Jacobian of the W's with respect to the Z's to obtain

$$|\det \begin{pmatrix} \frac{1}{Q_{11}} & \cdots & Q_{1n} & 0 & \cdots & 0 \\ & 1 \ddots & & & 0's & \\ & & & & & \\ & & 1 \ddots & & & \\ O's & & & & & \ddots 1 \end{pmatrix}| = 1/|Q_{11}| \neq 0$$

with probability one. Hence we can find the joint density of the W's since we do know that the Z's do have a joint density. We integrate out $w_{12}, \cdots, w_{nn}$ to find the marginal density of W_{11}. But $W_{11} = \det Z$, and since W_{11} has a density and hence a continuous distribution function, it follows that $P[W_{11} = 0] = 0$. Q.E.D.

LEMMA 9. *If* $\mathbf{X}_1, \cdots, \mathbf{X}_n$ *are all* $\mathcal{N}_p(\mu, \sum)$ *and independent with* $1 \leq p < n$, *if* $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$ *and if* $S = X(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)X^t$, *then* S *is non-singular with probability one.*

Proof: Let $\sum^{1/2}$ be a positive definite $p \times p$ matrix which satisfies $\sum = \sum^{1/2}\sum^{1/2}$. Define $\sum^{-1/2} = (\sum^{1/2})^{-1}$. Thus $rank(\sum^{-1/2} S \sum^{-1/2}) = rank(S)$. Note that $\sum^{-1/2} X = (\sum^{-1/2}\mathbf{X}_1 \vdots \cdots \vdots \sum^{1/2}\mathbf{X}_n)$ and that $\sum^{-1/2}\mathbf{X}_j$ is $\mathcal{N}_p(\sum^{-1/2}\mu, I_p)$ for $1 \leq j \leq n$. Hence all entries in the $p \times n$ matrix $\sum^{-1/2} X$ are independent normally distributed random variables. We next observe that the matrix $I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$ has rank $n-1$, is symmetric and is idempotent. Let R be an $n \times n$ orthogonal matrix such that $R(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)R^t$ is diagonal. Idempotency of $I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$ implies that of $R(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)R^t$; this matrix also has rank $n-1$ and hence has $n-1$ 1's along the main diagonal and one zero. We may assume without loss of generality that

$$R(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)R^t = diag(1, \cdots, 1, 0).$$

Since all pn random variables in the matrix $\sum^{-1/2} X$ are independent, normally distributed random variables with unit variances, and since R^t is an orthogonal matrix, it follows that all pn entries in $\sum^{-1/2} XR^t$ are normally distributed with unit variances. Hence (since $p < n$), $P[rank(\sum^{-1/2} XR^t) = p] = 1$ by Lemma 8. Now

$$\begin{aligned}\sum^{-1/2} S \sum^{-1/2} &= \sum^{-1/2} X R^t R(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)R^t R X^t \sum^{-1/2} \\ &= (\sum^{-1/2} X R^t diag(1, \cdots, 1, 0))(\sum^{-1/2} X R^t diag(1, \cdots, 1, 0))^t \\ &= ZZ^t,\end{aligned}$$

where Z is a $p \times (n-1)$ matrix of independent normally distributed random variables. By Lemma 8, $P[\det ZZ^t = 0] = 0$, and thus $P[\det S \neq 0] = 1$. Q.E.D.

THEOREM 1. *If* $\mathbf{X}_1, \cdots, \mathbf{X}_n$ *are i.i.d.* $\mathcal{N}_p(\boldsymbol{\mu}, \sum)$ *where* $1 \leq p < n$, *then the maximum likelihood estimates of* $\boldsymbol{\mu}$ *and* $\sum$ *are* $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ *and* $\hat{\sum} = \frac{1}{n} S$.

Proof: By Lemma 3, the maximum likelihood estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, whatever the value of $\sum$. Thus we can separately seek $\sum$ that maximizes the joint density, or, what amounts to the same thing, the logarithm of the joint density. By Remark 2,

$$\log f_X(x) = -\frac{np}{2}\log(2\pi) + \frac{n}{2}\log\det\sum^{-1} - \frac{1}{2}tr(\sum^{-1}(\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t)).$$

Hence we must find $\sum$ which maximizes the function

$$\varphi(\sum^{-1}) = \frac{n}{2}\log\det\sum^{-1} - \frac{1}{2}tr(\sum^{-1}(\sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t)),$$

are (replacing $\boldsymbol{\mu}$ by $\bar{\mathbf{x}}$),

$$\varphi(\sum^{-1}) = \frac{n}{2}\log\det\sum^{-1} - \frac{1}{2}tr(\sum^{-1}A),$$

where, as before, $A = \sum_{\alpha=1}^{n}(\mathbf{x}_\alpha - \bar{\mathbf{x}})(\mathbf{x}_\alpha - \bar{\mathbf{x}})^t$. Applying Lemma 7, φ is maximized when

$$\hat{\sum}^{-1} = nS^{-1}$$

provided S^{-1} exists with probability one. But by Lemma 9, S is non-singular with probability one. Thus $\hat{\sum} = \frac{1}{n}S$ is the maximum likelihood estimate of $\sum$. Q.E.D.

EXERCISES

1. Prove that $\frac{1}{n-1}S$ is an unbiased estimate of $\sum$.

2. Prove: If A is an $m \times n$ matrix, and if B is an $n \times m$ matrix, then $tr(AB) = tr(BA)$.

3. Prove: If C is an $n \times n$ positive-definite symmetric matrix, then so is C^{-1}.

4. Prove: If A is a $n \times n$ symmetric matrix, and if P is an $n \times n$ orthogonal matrix such that $P^tAP = diag(\lambda_1, \cdots, \lambda_n)$, then $\lambda_1, \cdots, \lambda_n$ are eigenvalues of A, and the ith column of P is an eigenvector of A corresponding to $\lambda_i, 1 \leq i \leq n$.

5. Prove: If A is a symmetric $n \times n$ matrix with eigenvalues $\lambda_1, \cdots, \lambda_n$, then $\det A = \prod_{j=1}^{n} \lambda_j$.

6. Prove: If D is a symmetric positive-definite $n \times n$ matrix, then there exists a symmetric, positive-definite matrix $D^{1/2}$ such that $D = D^{1/2}D^{1/2}$.

7. Prove: If $\mathbf{X}$ is $\mathcal{N}_p(\mu, \sum)$, if $\sum^{1/2}$ is a symmetric, positive-definite matrix satisfying $\sum = \sum^{1/2} \sum^{1/2}$, and if we define $\sum^{-1/2} = (\sum^{1/2})^{-1}$, then $\sum^{-1} = \sum^{-1/2} \sum^{-1/2}$, and the coordinates of $\sum^{-1/2} \mathbf{X}$ are independent.

8. Prove: If $\mathbf{X}$ and $\sum^{-1/2}$ are as in problem 7, and if P is a $p \times p$ orthogonal matrix, then the coordinates of $\mathbf{Y} = P \sum^{-1/2} \mathbf{X}$ are independent.

9. Prove that $I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t$ is idempotent.

10. Prove that $rank(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t) = n - 1$.

11. Prove: If A is a square matrix and is diagonal, and if A is idempotent, then its diagonal elements can only be 0's and 1's.

12. Prove that there exists an $n \times n$ orthogonal matrix R which satisfies

$$R(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)R^t = diag(1, 1, \cdots, 1, 0).$$

13. In the proof of Lemma 9 there is the statement that "all pn entries in $\sum^{-1/2} XR^t$ are normally distributed with unit variances." Prove this.

§2. **Lemmas for the Wishart Matrix.** This is a technical section devoted to the proofs of fifteen lemmas and one theorem. An overview is as follows. Let X be a $p \times n$ matrix of i.i.d. $\mathcal{N}(0,1)$ random variables. We first show that there exists a lower triangular matrix B^{-1} such that the rows of $B^{-1}X$ form an orthonormal system of p vectors. The matrix B^{-1} is not necessarily unique but can be constructed so that each diagonal element is a positive random variable. Thus B^{-1} is non-singular with probability one, and we denote its inverse by B. It will turn out that $B = (b_{ij})$ is lower triangular. The culmination is a theorem which states that all $p(p+1)/2$ random variables on or below the main diagonal are independent, that b_{ij} has the $\mathcal{N}(0,1)$ distribution if $i > j$ and b_{ii}^2 has the χ^2_{n-i+1}-distribution.

Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d. $\mathcal{N}_p(0, I_p)$, where $1 \leq p < n$, let $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$ and denote $A = XX^t$. If $Z = (z_{ij})$ is a $k \times k$ symmetric matrix, then we define $\mathbf{Z}_\Delta$ by $Z = (z_{11}\, z_{21}\, z_{22}\, z_{31}\, z_{32}\, z_{33} \cdots z_{k1} \cdots z_{kk})^t$. Note that Z_Δ is a vertical vector formed by the rows of the lower triangle of Z. Further, if $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are vectors in $\mathbf{R}^p$, then we denote $x = (\mathbf{x}_1 \vdots \cdots \vdots \mathbf{x}_n)$ and $a = xx^t$.

LEMMA 1. *Let* $\mathbf{x}_1, \cdots, \mathbf{x}_p$ *be linearly independent vectors in* $\mathbf{R}^p$. *Then there exist constants* $c_{ij}, 1 \leq i \leq p, 1 \leq j \leq i$, *such that* $c_{11}\mathbf{x}_1, c_{21}\mathbf{x}_1 + c_{22}\mathbf{x}_2, \cdots, c_{p1}\mathbf{x}_1 + \cdots + c_{pp}\mathbf{x}_p$ *form an orthonormal basis of* $\mathbf{R}^p$.

Proof: This is accomplished by the Gram-Schmidt process. Q.E.D.

Lemma 1 may be rewritten as follows:

LEMMA 1': *If D is a $p \times p$ nonsingular matrix, then there exists a lower triangular $p \times p$ matrix C_1 such that $C_1 D$ is an orthogonal matrix. Also, there exists an upper triangular matrix C_2 such that $C_2 D$ is an orthogonal matrix.*

LEMMA 2. *If C is a $k \times k$ lower triangular matrix, and if all diagonal elements of C are positive, then C is non-singular, and C^{-1} is lower triangular with all its diagonal elements positive.*

Proof: If $C = (c_{ij})$ and if $c_{ij} = 0$ for all $j > i$ (as in the hypothesis), then $\det C = \prod_{i=1}^{k} c_{ii}$. Since $c_{ii} > 0$ for all i, then $\det C > 0$, and C is non-singular. Note that for each i the cofactor of c_{ii} is the determinant of a lower triangular matrix, all of whose diagonal elements are positive. Hence all the diagonal elements of C^{-1} are positive. If $j > i$, then the cofactor of c_{ji} is $(-1)^{i+j}$ times the determinant of a lower triangular matrix with $j - i$ diagonal elements being zeros. Thus if $C^{-1} = (b_{ij})$, then $b_{ij} = 0$ if $j > i$. Q.E.D.

LEMMA 3. *If B is a positive-definite symmetric matrix, then there exists a lower triangular matrix D which satisfies $DD^t = B$.*

Proof: Let $B^{-1/2}$ be a positive-definite matrix satisfying $B^{-1/2}B^{-1/2} = B^{-1}$, and let $B^{1/2} = (B^{-1/2})^{-1}$. Note that $B^{1/2}$ is positive definite, and $B = B^{1/2}B^{1/2}$. By Lemma 1 there exists a lower triangular matrix C such that $CB^{1/2}$ is an orthogonal matrix. Then $CB^{1/2}B^{1/2}C^t = I$, or $B = C^{-1}(C^t)^{-1} = C^{-1}(C^{-1})^t$. By Lemma 2, C^{-1} is lower triangular. Thus we may take $D = C^{-1}$. Q.E.D.

LEMMA 4. *The matrix $A = XX^t$ is positive definite with probability one.*

Proof: Since $n > p, X$ has rank p with probability one. Hence for all $\zeta \in \mathbf{R}^p \backslash \{0\}, \zeta^t X \neq 0^t$ with probability one, which implies $\zeta^t XX^t \zeta > 0$ with probability one. Q.E.D.

Notation: Let X_i^t denote the ith row of X, i.e., $X^t = (X_1 \vdots \cdots \vdots X_p)$. Note that by Lemma 8 of Section 1 the vectors $X_1, \cdots, X_p$ are linearly independent with probability one.

LEMMA 5. *There exists a $p \times p$ lower triangular matrix of random variables denoted by $B^{-1} = (b^{ij})$ such that b^{kl} is a measurable function of $X_1, \cdots, X_k$ only, and such that if*

$$\begin{aligned} Y_1 &= b^{11} X_1 \\ Y_2 &= b^{21} X_1 + b^{22} X_2 \\ &\vdots \\ Y_p &= b^{p1} X_1 + \cdots + b^{pp} X_p, \end{aligned}$$

then $Y_1, \cdots, Y_p$ form an orthonormal system in $\mathbf{R}^n$ with probability 1. In addition $b^{ii} > 0$ with probability one for $1 \leq i \leq p$.

Proof: This follows from Lemma 1. Q.E.D.

Notation: $Y = \begin{pmatrix} Y_1^t \\ \cdots \\ \vdots \\ \cdots \\ Y_p^t \end{pmatrix}$. We observe that $Y = B^{-1}X$. By Lemma 5, $\det(B^{-1}) = \prod_{j=1}^p b^{ii} > 0$

with probability one. Thus the inverse of B^{-1} exists with probability one, and we may define B by $B = (B^{-1})^{-1}$. We denote $B = (b_{ij})$, and observe that, by Lemma 2, B is lower triangular. We collect all this in the following lemma.

LEMMA 6. *The following hold:*

(1) $Y = B^{-1}X$,

(2) $YY^t = I_p$ *with probability one,*

(3) $X = BY$,

(4) $B = XY^t$,

(5) $A = BB^t$, *and*

(6) $b_{ij} = X_i^t Y_j$.

Proof: (1) follows from Lemma 5. (2) is true because $Y_1, \cdots, Y_p$ are orthonormal with probability one. (3) follows from (1). (4) follows from (2) and (3). (5) follows from (2) and (4) and the definition of A. (6) follows from (4).

LEMMA 7. *For $2 \leq i \leq p$, the sets of random vectors $\{Y_1, \cdots, Y_{i-1}\}$ and $\{X_i, \cdots, X_p\}$ are independent.*

Proof: Recall that $Y_1, \cdots, Y_p$ form an orthonormal system, where Y_j (by the Gram-Schmidt process that occurred to make Y out of X) is a function of $X_1, \cdots, X_j$ only. Since all np random variables in X are independent, then $\{X_i, \cdots, X_p\}$ and $\{Y_1, \cdots, Y_{i-1}\}$ are independent.

Q.E.D.

LEMMA 8. *For $2 \leq i \leq p$, the matrix*

$$\begin{pmatrix} Y_1^t \\ \cdots \\ \vdots \\ \cdots \\ Y_{i-1}^t \end{pmatrix}$$

constitutes the first $i-1$ rows of a random orthogonal matrix, all of whose entries are random variables that are independent of $X_i, \cdots, X_p$.

Proof: This follows from Lemma 2 in Section 4 of Chapter 1 and from Lemma 7 above.

Q.E.D.

LEMMA 9. *For $2 \leq i \leq p$,*

$$\begin{pmatrix} Y_1^t \\ \vdots \\ Y_{i-1}^t \end{pmatrix} X_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{i,i-1} \end{pmatrix}.$$

Proof: By Lemma 6(4), $B = XY^t$ or $B^t = YX^t$. The first $i-1$ terms of the ith column of both sides of this equation yield our result. Q.E.D.

LEMMA 10. *For $2 \leq i \leq p$, the random variables $b_{i1}, \cdots, b_{i,i-1}$ are independent, and each is $\mathcal{N}(0,1)$.*

Proof: First observe that $Y_1, \cdots, Y_{i-1}$ are independent of X_i; this follows because Y_j is a

Borel-measurable function of $X_1, \cdots, X_j, 1 \leq j \leq i-1$. By Lemma 2 of Section 4 of Chapter 1, $Y_1, \cdots, Y_{i-1}$ are the first $i-1$ columns of an $n \times n$ random orthogonal matrix Q, and Q and X_i are independent. Since the random variables in X_i are independent and each $\mathcal{N}(0,1)$, then by Theorem 1 of Section 4 of Chapter 1, the coordinates of QX_i are independent and $\mathcal{N}(0,1)$. Hence the first $i-1$ coordinates of QX_i, namely, $b_{i1}, \cdots, b_{i,i-1}$ are independent, and each is $\mathcal{N}(0,1)$. Q.E.D.

LEMMA 11. *For* $1 \leq i \leq p, \quad X_i^t X_i = \sum_{j=1}^{i} b_{ij}^2$.

Proof: By Lemma 6, $X = BY$, and hence $XX^t = BYY^tB^t = BB^t$. The conclusion easily follows. Q.E.D.

LEMMA 12. *For* $1 \leq i \leq p, X_i^t X_i$ *has the chi-square distribution with* n *degrees of freedom.*

Proof: This follows from the fact that all coordinates of X_i are $\mathcal{N}(0,1)$. Q.E.D.

LEMMA 13. *The random vector* $\begin{pmatrix} b_{i1} \\ \vdots \\ b_{i,i-1} \end{pmatrix}$ *and the random variable* $b_{ii}^2 = X_i^t X_i - \sum_{j=1}^{i-1} b_{ij}^2$ *are independent, and* b_{ii}^2 *has the* χ_{n-i+1}^2*-distribution.*

Proof: We apply Theorem 3 of Section 4 of Chapter 1 which states in our case: if X_i is $\mathcal{N}_n(0, I_n)$ (which it is) for $2 \leq i \leq n$, if $P_0 = \begin{pmatrix} Y_1^t \\ \cdots \\ \vdots \\ \cdots \\ Y_{i-1}^t \end{pmatrix}$ is an $(i-1) \times n$ random matrix whose rows form an orthonormal system (which it does), if $\begin{pmatrix} b_{i1} \\ \vdots \\ b_{i,i-1} \end{pmatrix} = P_0X_i$ (which is true by Lemma 9), and if P_0 and X_i are independent (which they are, by Lemma 7), then $X_i^tX_i - (P_0X_i)^tP_0X_i$ and P_0X_i are independent and $X_i^tX_i - (P_0X_i)^tP_0X_i$ has the $\chi^2_{n-(i-1)}$-distribution. But, by Lemma 9, $(P_0X_i)^tP_0X_i = \sum_{j=1}^{i-1} b_{ij}^2$; thus by our application of the theorem above, we obtain the lemma. Q.E.D.

LEMMA 14. *For $2 \leq i \leq p$, the random variables $b_{i1}, \cdots, b_{i,i-1}, b_{ii}$ are independent.*

Proof: By Lemmas 10 and 13, $b_{i1}, \cdots, b_{i,i-1}, b_{ii}^2$ are independent. Recall that $P[b^{ii} > 0] = 1$. Also one easily proves that $b_{ii} = 1/b^{ii}$. Hence $b_{i1}, \cdots, b_{i,i-1}, b_{ii}$ are independent. Q.E.D.

LEMMA 15. *The rows of B are stochastically independent.*

Proof: Let us denote

$$P_{i-1} = \begin{pmatrix} X_1^t \\ \cdots \\ \vdots \\ \cdots \\ X_{i-1}^t \end{pmatrix} \text{ and } Q_{i-1} = \begin{pmatrix} Y_1^t \\ \cdots \\ \vdots \\ \cdots \\ Y_{i-1}^t \end{pmatrix}$$

and let B_{i-1}^{-1} denote the first $i-1$ rows and first $i-1$ columns of B^{-1}. Because of the fact that B^{-1} is lower triangular, it follows that $Q_{i-1} = B_{i-1}^{-1}P_{i-1}$. But note that all of the random variables in B_{i-1}^{-1} are Borel-measurable functions of those in P_{i-1}, the first $i-1$ rows of X, and thus we may write $Q_{i-1} = \psi(P_{i-1})$, where ψ is a Borel-measurable function. Let us agree on the following notation: if $\mathbf{V} = \begin{pmatrix} V_1 \\ \vdots \\ V_r \end{pmatrix}$ is a random vector, and if $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_r \end{pmatrix} \in \mathbf{R}^r$, then we shall denote $[\mathbf{V} \leq \mathbf{v}] = \bigcap_{i=1}^{r}[V_i \leq v_i]$. Let p_{i-1} be an $(i-1) \times n$ matrix of linearly independent rows of real numbers. Then for $\mathbf{x} \in \mathbf{R}^{i-1}$, we have, for $y \geq 0$,

$$\begin{aligned} & P([b_{i1} \leq x_1, \cdots, b_{i,i-1} \leq x_{i-1}, b_{ii}^2 \leq y] | P_{i-1} = p_{i-1}) \\ & = P([Q_{i-1}X_i \leq \mathbf{x}][X_i^t X_i - X_i^t Q_{i-1}^t Q_{i-1} X_i \leq y] | P_{i-1} = p_{i-1}) \\ & = P([\psi(p_{i-1})X_i \leq \mathbf{x}][X_i^t X_i - X_i^t \psi(p_{i-1})^t \psi(p_{i-1}) X_i \leq y]) \\ & = \int_{-\infty}^{x_1} \varphi(t)dt \cdots \int_{-\infty}^{x_{i-1}} \varphi(t)dt \int_0^y \chi_{n-i+1}^2(t)dt, \end{aligned}$$

where φ is the density of the $\mathcal{N}(0,1)$ distribution and $\chi_{n-i+1}^2(t)$ is the density of the χ_{n-i+1}^2-distribution. Thus we see that the conditional joint distribution of the ith row of B namely, $b_{i1}, \cdots, b_{ii}$, given the values of the first $i-1$ rows of X, is the same as that of the unconditional distribution. We use this just-proved *fact* to prove that the rows of B are independent. Let $R_i = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{ii} \end{pmatrix}$, and let C_i be any Borel-set in $\mathbf{R}^i, 1 \leq i \leq p$. Then

$$\begin{aligned} P(\cap_{i=1}^{p}[R_i \in C_i]) &= E\left(\prod_{i=1}^{p} I_{[R_i \in C_i]}\right) \\ &= E\left(E\left(\prod_{i=1}^{p} I_{[R_i \in C_i]} | X_1, \cdots, X_{p-1}\right)\right) \\ &= E\left(\prod_{i=1}^{p-1} I_{[R_i \in C_i]} E\left(I_{[R_p \in C_p]} | X_1, \cdots, X_{p-1}\right)\right). \end{aligned}$$

But by the *fact* proved above, if $2 \leq i \leq p$,

$$E(I_{[R_i \in C_i]} | X_1, \cdots, X_{i-1}) = E(I_{[R_i \in C_i]}) = P[R_i \in C_i].$$

Hence

$$P\left(\bigcap_{i=1}^{p}[R_i \in C_i]\right) = P[R_p \in C_p] E\left(\prod_{i=1}^{p-1} I_{[R_i \in C_i]}\right).$$

Repeating this argument $p-2$ more times we obtain

$$P\left(\bigcap_{i=1}^{p}[R_i \in C_i]\right) = \prod_{i=1}^{p} P[R_i \in C_i].$$

Q.E.D.

THEOREM 1. *All the random variables in B are independent.*

Proof: This follows from Lemma 14 and 15.

Thus we have obtained the joint distribution of B_Δ.

EXERCISES

1. Supply a proof that $B_{ii} = 1/b^{ii}$ in the proof of Lemma 14.

2. Prove: If D is an $n \times n$ lower triangular matrix each of whose diagonal elements is unequal to zero, then D is non-singular and D^{-1} is lower triangular.

§3 . **The Wishart Distribution.** In this section we define the Wishart distribution and obtain a number of its important properties. We then define and obtain the distribution of the Hotelling T^2-statistic.

DEFINITION. *Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \sum)$, where $1 \leq p < n$. If X and D are defined by $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$ and $D = (X - \boldsymbol{\mu}\mathbf{1}_n^t)(X - \boldsymbol{\mu}\mathbf{1}_n^t)^t$, then D is said to have the $W_p(n, \sum)$-distribution, known as the Wishart distribution determined by p, n and $\sum$.*

It should be noticed that the distribution of D does not depend on $\boldsymbol{\mu}$; indeed, D only depends on $\mathbf{X}_i - \boldsymbol{\mu}, 1 \leq i \leq n$. Strictly speaking, the distribution of D really refers to the joint distribution of the $p(p+1)/2$ random variables that are on or below the main diagonal of D; this is because D is a symmetric matrix. In reference to these $p(p+1)/2$ lower triangular elements of D we use the notation D_Δ. But sometimes we shall only speak of the distribution of D and shall refer to it as having the $W_p(n, \sum)$-distribution.

LEMMA 1. *If D is $W_p(n, \sum)$, and if H is an $m \times p$ matrix of constants of rank m, where $m \leq p$, then HDH^t has the $W_m(n, H \sum H^t)$ distribution.*

Proof: Let X be a sample of size n on a $\mathcal{N}_p(\boldsymbol{\mu}, \sum)$ distribution. Then $Y = HX$ is a sample of size n on a $\mathcal{N}_m(H\boldsymbol{\mu}, H \sum H^t)$ distribution. We may write

$$D = (X - \boldsymbol{\mu}\mathbf{1}_n^t)(X - \boldsymbol{\mu}\mathbf{1}_n^t)^t.$$

Then, clearly,

$$HDH^t = (Y - H\boldsymbol{\mu}\mathbf{1}_n^t)(Y - H\boldsymbol{\mu}\mathbf{1}_n^t)^t.$$

Thus HDH^t has the $W_m(n, H\sum H^t)$ distribution. Q.E.D.

COROLLARY 1 TO LEMMA 1. *If D is $W_p(n,\sum)$, if*

$$D = \left(\begin{array}{c|c} D_{11} & D_{12} \\ \hdashline D_{21} & D_{22} \end{array}\right)$$

is a partition of D where D_{11} is an $m \times m$ submatrix, and if

$$\sum = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

is a partition of $\sum$ where $\sum_{11}$ is an $m \times m$ submatrix, then D_{11} has the $W_m(n, \sum_{11})$ distribution.

Proof: Let $H = (I_m|0)$ be an $m \times p$ matrix. Then Lemma 1 yields the result. Q.E.D.

COROLLARY 2 TO LEMMA 1. *If $\mathrm{h} \in \mathbf{R}^p$ is a constant vector, if $\mathrm{h} \neq 0$, and if D is $W_p(n,\sum)$, then $\frac{\mathrm{h}^t D \mathrm{h}}{\mathrm{h}^t \sum \mathrm{h}}$ has the χ^2_n-distribution.*

Proof: By Corollary 1, $\mathrm{h}^t D\mathrm{h}$ has the $W_1(n, \mathrm{h}^t \sum \mathrm{h})$ distribution, i.e., $\mathrm{h}^t D\mathrm{h}$ has the same distribution as does $\sum_{i=1}^n Y_i^2$, where $Y_1, \cdots, Y_n$ are independent, each being $\mathcal{N}(0, \mathrm{h}^t \sum \mathrm{h})$. Hence $\mathrm{h}^t D\mathrm{h}/\mathrm{h}^t \sum \mathrm{h}$ has the χ^2_n-distribution. Q.E.D.

LEMMA 2. *If h is a p-dimensional random vector such that $P[\mathrm{h} \neq 0] = 1$, if D is $W_p(n,\sum)$, and if h and D are independent, then*

(i) $\mathrm{h}^t D\mathrm{h}/\mathrm{h}^t \sum \mathrm{h}$ has the χ^2_n-distribution, and

(ii) $\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\sum\mathbf{h}$ *and* $\mathbf{h}$ *are independent.*

Proof: We shall prove this lemma only in the case where $\mathbf{h}$ has a joint absolutely continuous distribution. By Corollary 2 to Lemma 1 and by Propositions 4 and 6 in Chapter 1 we have, for $x > 0$,

$$\begin{aligned} P[\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\textstyle\sum\mathbf{h} \le x] &= E(I_{[\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\sum\mathbf{h}\le x])}) \\ &= \int_{\mathbf{R}^p} E(I_{[\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\sum\mathbf{h}\le x]}|\mathbf{h} = \mathbf{r})f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\ &= \int_{\mathbf{R}^p} P[\mathbf{r}^t D\mathbf{r}/\mathbf{r}^t\textstyle\sum\mathbf{r} \le x]f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\ &= \int_{\mathbf{R}^p}\left(\int_0^x \chi_n^2(t)dt\right)f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} = \int_0^x \chi_n^2(t)dt, \end{aligned}$$

where $\chi_n^2(t)$ denotes the density of the χ_n^2-distribution. This proves (i). In order to prove(ii), we first note that in our proof of part (i) we showed that

$$P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\textstyle\sum\mathbf{h} \le x]|\mathbf{h} = \mathbf{r}) = \int_0^x \chi_n^2(t)dt$$

for all $x > 0$ and all $\mathbf{r} \in \mathbf{R}^p$. Hence, for all Borel sets A in $\mathbf{R}^p$,

$$\begin{aligned} &P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\textstyle\sum\mathbf{h} \le x][\mathbf{h} \in A]) \\ &= \int_{\mathbf{R}^p} P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\textstyle\sum\mathbf{h} \le x][\mathbf{h} \in A]|\mathbf{h} = \mathbf{r})f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\ &= \int_{\mathbf{R}^p} P([\mathbf{r}^t D\mathbf{r}/\mathbf{r}^t\textstyle\sum\mathbf{r} \le x][\mathbf{r} \in A])f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\ &= \int_A P([\mathbf{r}^t D\mathbf{r}/\mathbf{r}^t\textstyle\sum\mathbf{r} \le x])f_{\mathbf{h}}(\mathbf{r})d\mathbf{r}. \end{aligned}$$

Now by the remark above, by the fact that $P[\mathbf{h} \ne \mathbf{0}] = 1$, we see that by Corollary 2 to Lemma 1 and for all $\mathbf{r} \ne \mathbf{0}$, and by (i),

$$P([\mathbf{r}^t D\mathbf{r}/\mathbf{r}^t\textstyle\sum\mathbf{r} \le x]) = P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t\textstyle\sum\mathbf{h} \le x]).$$

Hence

$$\begin{aligned}
&P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t \textstyle\sum \mathbf{h} \le x][\mathbf{h} \in A]) \\
&\quad = \int_A P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t \textstyle\sum \mathbf{h} \le x]) f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\
&\quad = P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t \textstyle\sum \mathbf{h} \le x]) \int_A f_{\mathbf{h}}(\mathbf{r})d\mathbf{r} \\
&\quad = P([\mathbf{h}^t D\mathbf{h}/\mathbf{h}^t \textstyle\sum \mathbf{h} \le x]) P([\mathbf{h} \in A]).
\end{aligned}$$

Q.E.D.

We shall henceforth use the following notation for elements of the matrices $D, D^{-1}, \sum$ and $\sum^{-1}$:

$$D = (d_{ij}), D^{-1} = (d^{ij}), \textstyle\sum = (\sigma_{ij}) \text{ and } \textstyle\sum^{-1} = (\sigma^{ij}).$$

LEMMA 3. *If D is $W_p(n, \sum)$, then $\det D / \det \sum$ has the same distribution as the product of p independent random variables whose distributions are $\chi^2_n, \chi^2_{n-1}, \cdots, \chi^2_{n-p+1}$, and σ^{pp}/d^{pp} has the χ^2_{n-p+1}-distribution.*

Proof: Let C be a lower triangular matrix of constants such that $\sum = CC^t$, and define $A = C^{-1}D(C^{-1})^t$. Then, by Lemma 1, A is $W_p(n, C^{-1}\sum(C^{-1})^t)$, i.e., A is $W_p(n, I_p)$. Thus A has the same joint distribution as does XX^t, where X is a $p \times n$ matrix of independent $\mathcal{N}(0,1)$ random variables. For distributional purposes we may define $A = XX^t$. Now with respect to X let B be as defined in Section 2, i.e., $B = (b_{ij})$ is a $p \times p$ lower triangular matrix, $\{b_{ij}, 1 \le j \le i, 1 \le i \le p\}$ are independent random variables, b_{ij} is $\mathcal{N}(0,1)$ if $j < i$, b^2_{ii} is χ^2_{n-i+1} and $A = BB^t$. Let us define $T = CB$. Clearly, T is lower triangular. We next note that

$$TT^t = CBB^tC^t = CAC^t = CC^{-1}D(C^{-1})^tC^t = D,$$

i.e., $TT^t = D$. Thus

$$\begin{aligned} \det D &= \det(TT^t) = \det(CBB^tC^t) = \det(CC^t)\det(BB^t) \\ &= \det(\sum)\prod_{i=1}^{p} b_{ii}^2. \end{aligned}$$

By Lemma 13 in Section 2 of Chapter 2, we obtain the first conclusion of our lemma. In order to obtain the second conclusion, let $D_{p-1}, T_{p-1}, A_{p-1}, B_{p-1} \sum_{p-1}$ and C_{p-1} be $(p-1)\times(p-1)$ matrices obtained from the first $p-1$ rows and the first $p-1$ columns of $D, T, A, B, \sum$ and C respectively. From relations established above it easily follows that $D_{p-1} = T_{p-1}T_{p-1}^t, A_{p-1} = B_{p-1}B_{p-1}^t, \sum_{p-1} = C_{p-1}C_{p-1}^t, T_{p-1} = C_{p-1}B_{p-1}$ and $D_{p-1} = C_{p-1}A_{p-1}C_{p-1}^t$. From these and the fact that B is lower triangular we get

$$\frac{\det A}{\det A_{p-1}} = \frac{\det(BB^t)}{\det(B_{p-1}B_{p-1}^t)} = \frac{(\det B)^2}{(\det B_{p-1})^2} = b_{pp}^2.$$

Hence

$$\begin{aligned} b_{pp}^2 &= \frac{\det A}{\det A_{p-1}} = \frac{\det A \det C \det C^t \det C_{p-1} \det C_{p-1}^t}{\det A_{p-1} \det C \det C^t \det C_{p-1} \det C_{p-1}^t} \\ &= \frac{\det(CAC^t)}{\det(C_{p-1}A_{p-1}C_{p-1}^t)} \frac{\det(C_{p-1}C_{p-1}^t)}{\det(CC^t)} \\ &= \frac{\det D}{\det D_{p-1}} \frac{\det \sum_{p-1}}{\det \sum}. \end{aligned}$$

Note that $D^{-1} = (d^{ij})$ where $d^{ij} = (\text{cofactor of } d_{ji})/\det D$. Since $\det D_{p-1}$ is the cofactor of d_{pp}, it follows that $d^{pp} = (\det D_{p-1})/\det D$. Similarly, $\sigma^{pp} = (\det \sum_{p-1})/\det \sum$. Hence $b_{pp}^2 = \sigma^{pp}/d^{pp}$ which has the χ_{n-p+1}^2-distribution. Q.E.D.

LEMMA 4. *Let D be $W_p(n,\sum)$, let* $\mathbf{h}$ *be a p-dimensional random vector with a joint absolutely continuous distribution function, and assume that D and* $\mathbf{h}$ *are independent. Then*

(i) $\frac{\mathbf{h}^t\sum^{-1}\mathbf{h}}{\mathbf{h}^t D^{-1}\mathbf{h}}$ has the χ^2_{n-p+1}*-distribution, and*

(ii) $\frac{\mathbf{h}^t\sum^{-1}\mathbf{h}}{\mathbf{h}^t D^{-1}\mathbf{h}}$ *and* $\mathbf{h}$ *are independent.*

Proof: Let H be a $p \times p$ orthogonal matrix of random variables whose pth row is $(1/\|\mathbf{h}\|)\mathbf{h}^t$ and such that all entries in H are measurable functions of $\mathbf{h}$. We may write $H = H(\mathbf{h})$. Let $\mathbf{h}_0 \in$ range $(\mathbf{h})$, and let $H_0 = H(\mathbf{h}_0)$. Now H_0 is an orthogonal $p \times p$ matrix of numbers. Note that as a consequence of our hypotheses, H and D are independent. Define $D^* = H_0DH_0^t$ and $\sum^* = H_0\sum H_0^t$. By Lemma 1 and the fact that H and D are independent, the conditional distribution of HDH^t given $H = H_0$ is the unconditional distribution of $H_0DH_0^t$, which is $W_p(n, H_0\sum H_0^t)$. Let us denote

$$D^* = (d^*_{ij}), D^{*-1} = (d^{*ij}), \sum^* = (\sigma^*_{ij})$$

and $\sum^{*-1} = (\sigma^{*ij})$, where D^* and $\sum^*$ are as denoted above. By Lemma 3, σ^{*pp}/d^{*pp} has the χ^2_{n-p+1}-distribution. Note that $H_0D^{-1}H_0^t = D^{*-1}$. Now, for appropriate matrices H_1, H_2, H_3 and H_4 we have

$$\begin{aligned} H_0D^{-1}H_0^t &= \begin{pmatrix} H_1 \\ \cdots \\ \frac{1}{\|\mathbf{h}_0\|}\mathbf{h}_0^t \end{pmatrix} D^{-1}\left(H_1^t \vdots \frac{1}{\|\mathbf{h}_0\|}\mathbf{h}_0\right) \\ &= \left(\begin{array}{c|c} H_2 & H_3 \\ \hline H_4 & \frac{1}{\|\mathbf{h}_0\|^2}\mathbf{h}_0^tD^{-1}\mathbf{h}_0 \end{array}\right) \end{aligned}$$

So $d^{*pp} = \frac{1}{\|\mathbf{h}_0\|^2}\mathbf{h}_0^t D^{-1}\mathbf{h}_0$. Similarly $\sigma^{*pp} = \frac{1}{\|\mathbf{h}_0\|^2}\mathbf{h}_0^t \sum^{-1}\mathbf{h}_0$. As noted above, σ^{*pp}/d^{*pp} has the χ^2_{n-p+1}-distribution. Thus $\frac{\mathbf{h}_0^t\sum^{-1}\mathbf{h}_0}{\mathbf{h}_0^t D^{-1}\mathbf{h}_0}$ has the same distribution function. Now let $F(x)$ be the distribution function for the χ^2_{n-p+1}-distribution, and let $g(\cdot)$ be the joint density of $\mathbf{h}$. Then, since $\mathbf{h}$ and D are independent we have

$$\begin{aligned} P\left[\frac{\mathbf{h}^t\sum^{-1}\mathbf{h}}{\mathbf{h}^t D^{-1}\mathbf{h}} \le x\right] &= \int_{\mathbf{R}^p} P\left(\left[\frac{\mathbf{h}^t\sum^{-1}\mathbf{h}}{\mathbf{h}^t D^{-1}\mathbf{h}} \le x\right] |\mathbf{h} = \mathbf{h}_0\right) g(\mathbf{h}_0)d\mathbf{h}_0 \\ &= \int_{\mathbf{R}^p} P\left[\frac{\mathbf{h}_0^t\sum^{-1}\mathbf{h}_0}{\mathbf{h}_0^t D^{-1}\mathbf{h}_0} \le x\right] g(\mathbf{h}_0)d\mathbf{h}_0 \\ &= \int_{\mathbf{R}^p} F(x)g(\mathbf{h}_0)d\mathbf{h}_0 = F(x), \end{aligned}$$

which proves (i). The proof of (ii) follows the same steps for the proof of conclusion in (ii) in Lemma 2. Q.E.D.

LEMMA 5. *If* $\mathbf{U}$ *is* $\mathcal{N}_p(\mu, \sum)$, *then* $(\mathbf{U}-\mu)^t\sum^{-1}(\mathbf{U}-\mu)$ *has the* χ^2_p*-distribution.*

Proof: Since $\sum$ and therefore $\sum^{-1}$ are positive definite, there exists a positive definite matrix $\sum^{-1/2}$ such that $\sum^{-1/2}\sum^{-1/2} = \sum^{-1}$. One easily verifies that $\sum^{-1/2}(\mathbf{U}-\mu)$ is $\mathcal{N}_p(0, I_p)$. Hence $(\mathbf{U}-\mu)^t\sum^{-1}(\mathbf{U}-\mu)$ is the sum of squares of p independent $\mathcal{N}(0,1)$ random variables. Q.E.D.

THEOREM 1. *If* $\mathbf{U}$ *is* $\mathcal{N}_p(\mu, \sum)$, *if* D *is* $W_p(r, \sum)$, *if* $p < r$, *and if* $\mathbf{U}$ *and* D *are independent, then*

$$\frac{r-p+1}{p}(\mathbf{U}-\mu)^t D^{-1}(\mathbf{U}-\mu)$$

has the $F_{p,r-p+1}$*-distribution.*

Proof: By Lemma 4, if we denote $A = \frac{(\mathbf{U}-\mu)^t\sum^{-1}(\mathbf{U}-\mu)}{(\mathbf{U}-\mu)^t D^{-1}(\mathbf{U}-\mu)}$, then A has the χ^2_{r-p+1}-distribution, and A and $\mathbf{U}$ are independent. Let $B = (\mathbf{U} - \mu)^t \sum^{-1} (\mathbf{U} - \mu)$. Then A and B are independent. By Lemma 5,

$$\frac{B/p}{A/(r-p+1)}$$

has the $F_{p,r-p+1}$-distribution, i.e., $\frac{r-p+1}{p}(\mathbf{U} - \mu)^t D^{-1}(\mathbf{U} - \mu)$ has the $F_{p,r-p+1}$-distribution. Q.E.D.

EXERCISES

1. Prove: If $\mathbf{h}$ is a p-dimensional random vector such that $P[\mathbf{h} \neq \mathbf{0}] = 1$, then there exists a $p \times p$ orthogonal random matrix H whose pth row is $(1/\|\mathbf{h}\|)\mathbf{h}^t$ and such that $H = (h_{ij})$ is a Borel measurable function of $\mathbf{h}$.

2. Prove: if W has the $W_1(n, I_1)$ distribution, then it has the χ_n^2-distribution.

3. Prove that the Wishart distribution as defined in this chapter does not depend on $\boldsymbol{\mu}$.

CHAPTER 3. HOTELLING'S T^2-STATISTIC.

§1. Relationships between $\bar{\mathbf{X}}$ and S. Our main aim in this course is the extension of results from one-dimensional normal populations to p-dimensional normal populations. This means at the beginning to obtain a p-dimensional analogue of the one- and two-sample t-tests and the F-test for linear models. In the one- and two-sample t-tests we were concerned with $\bar{X}$ and s^2, obtained from a sample of size $n, X_1, \cdots, X_n$, on and $\mathcal{N}(\mu, \sigma^2)$-population. We recall that $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X})^2$. The properties needed about $\bar{X}$ and s^2 were these:

(i) $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$,

(ii) $(n-1)s^2/\sigma^2$ has the χ^2_{n-1}-distribution, and

(iii) $\bar{X}$ and s^2 are independent random variables.

This section is devoted to a generalization of (i), (ii) and (iii) given in Theorem 1.

Again we deal with $\mathbf{X}_1, \cdots, \mathbf{X}_n$, a sample of (independent random vectors of) size n from a population which is $\mathcal{N}_p(\boldsymbol{\mu}, \sum)$, where $\boldsymbol{\mu}$ and $\sum$ are unknown. The notations $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i, X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$ and $S = X(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)X^t$ are as before.

THEOREM 1. *If $\bar{\mathbf{X}}$ and S are as above, then*

(i) $\sqrt{n}\bar{\mathbf{X}}$ *is* $\mathcal{N}_p(\sqrt{n}\boldsymbol{\mu}, \sum)$,

(ii) S *has the* $W_p(n-1, \sum)$ *distribution, and*

(iii) $\bar{\mathbf{X}}$ *and* S *are independent.*

Proof: Let $x = (\mathbf{x}_1 \vdots \cdots \vdots \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbf{R}^p$ for each i. Recall from Section 1 in Chapter 2 that the joint density of X is

$$\begin{aligned} f_X(x) &= (2\pi)^{-\frac{np}{2}} (\det \textstyle\sum^{-1})^{\frac{n}{2}} \exp -\frac{1}{2} tr\{\textstyle\sum^{-1} \sum_{\alpha=1}^{n} (\mathbf{x}_\alpha - \boldsymbol{\mu})(\mathbf{x}_\alpha - \boldsymbol{\mu})^t\} \\ &= (2\pi)^{-\frac{np}{2}} (\det \textstyle\sum^{-1})^{\frac{n}{2}} \exp -\frac{1}{2} tr\{\textstyle\sum^{-1} (x - \boldsymbol{\mu}\mathbf{1}_n^t)(x - \boldsymbol{\mu}\mathbf{1}_n^t)^t\}. \end{aligned}$$

Let P be an $n \times n$ orthogonal matrix whose last column is $(\frac{1}{\sqrt{n}} \frac{1}{\sqrt{n}} \cdots \frac{1}{\sqrt{n}})^t$, i.e., $P = (Q \vdots \frac{1}{\sqrt{n}} \mathbf{1}_n) = (p_{ij})$. Let us define

$$Y = XP \text{ and } y = xP,$$

where $y = (y_{ij})$ and $x = (x_{ij})$ are $p \times n$ matrices of real numbers. Note that the mapping between x and y is one-to-one and continuously differentiable, and

$$y_{ij} = \sum_{r=1}^{n} x_{ir} p_{rj}.$$

Thus

$$\frac{\partial y_{ij}}{\partial x_{ir}} = p_{rj} \text{ and } \frac{\partial y_{ij}}{\partial x_{kr}} = 0 \text{ if } i \neq k.$$

We wish to compute the Jacobian of

$$y_{11}, \cdots, y_{1n}, y_{21}, \cdots, y_{2n}, \cdots, y_{p1}, \cdots, y_{pn}$$

(which determine the columns) with respect to $x_{11}, \cdots, x_{1n}, x_{21}, \cdots, x_{2n}, \cdots, x_{p1}, \cdots, x_{pn}$

(which determine the rows). Using the above partial derivative formulas we obtain

$$|J\left(\frac{y}{x}\right)| = \left|\begin{array}{c|c|c|c} \mathrm{P} & 0 & 0\text{'s} & 0 \\ \hline 0 & \mathrm{P} & 0\text{'s} & 0 \\ \hline 0\text{'s} & 0\text{'s} & & 0\text{'s} \\ \hline 0 & 0 & 0\text{'s} & \mathrm{P} \end{array}\right| = |P|^p = \pm 1.$$

Since the last column of $\mathbf{P}$ is $\frac{1}{\sqrt{n}}\mathbf{1}_n$, we may write

$$Y = (Z\vdots\sqrt{n}\bar{\mathrm{X}}) \text{ or } y = (z\vdots\sqrt{n}\bar{\mathrm{x}})$$

for some $p \times (n-1)$ matrix Z (or z). We observe that $\mathbf{1}_n^t P = (0 \cdots 0\sqrt{n})$, since the first $n-1$ columns of $\mathbf{P}$ are orthogonal to the last column which is equal to $\frac{1}{\sqrt{n}}\mathbf{1}_n$. Using this, we obtain

$$\begin{aligned} (x - \mu\mathbf{1}_n^t)(x - \mu\mathbf{1}_n^t)^t &= (x - \mu\mathbf{1}_n^t)PP^t(x - \mu\mathbf{1}_n^t)^t \\ &= (y - \mu\mathbf{1}_n^t P)(y - \mu\mathbf{1}_n^t \mathrm{P})^t \\ &= (z\vdots\sqrt{n}(\bar{\mathrm{x}} - \mu))(z\vdots\sqrt{n}(\bar{\mathrm{x}} - \mu))^t \\ &= zz^t + n(\bar{\mathrm{x}} - \mu)(\bar{\mathrm{x}} - \mu)^t. \end{aligned}$$

Now $x = yP^t$, so

$$\begin{aligned} f_Y(y) &= f_X(yP^t)|J\left(\frac{x}{y}\right)| \\ &= (2\pi)^{-\frac{np}{2}}(\det\Sigma^{-1})^{\frac{n}{2}} \exp -\frac{1}{2}tr\{\Sigma^{-1}(zz^t + n(\bar{\mathrm{x}} - \mu)(\bar{\mathrm{x}} - \mu)^t\} \\ &= (2\pi)^{-\frac{np}{2}}(\det\Sigma^{-1})^{\frac{n}{2}} \exp -\frac{1}{2}\{tr(\Sigma^{-1}zz^t) + \sqrt{n}(\bar{\mathrm{x}} - \mu)^t\Sigma^{-1}\sqrt{n}(\bar{\mathrm{x}} - \mu)\}. \end{aligned}$$

Hence

$$\begin{aligned} f_Y(y) &= f_{(Z\vdots\sqrt{n}\bar{\mathbf{X}})}(z\vdots\sqrt{n}\bar{\mathrm{x}}) \\ &= (2\pi)^{-p/2}(\det{\textstyle\sum}^{-1})^{1/2}\exp -\frac{1}{2}(\sqrt{n}(\bar{\mathrm{x}}-\mu)^t{\textstyle\sum}^{-1}\sqrt{n}(\bar{\mathrm{x}}-\mu)) \\ &\quad \cdot(2\pi)^{-\frac{(n-1)p}{2}}(\det{\textstyle\sum}^{-1})^{\frac{n-1}{2}}\exp -\frac{1}{2}tr({\textstyle\sum}^{-1}zz^t). \end{aligned}$$

Since the joint density of Z and $\sqrt{n}\bar{\mathbf{X}}$ factors, it follows that Z and $\bar{\mathbf{X}}$ are independent. Clearly, $\sqrt{n}\bar{\mathbf{X}}$ is $\mathcal{N}_p(\sqrt{n}\mu, \sum)$. Comparing the density of Z with the density of X given at the beginning of this proof, we see that the distribution of Z is the same as that of a sample of size $n-1$ on a $\mathcal{N}_p(0,\sum)$ population. Thus, ZZ^t has the $W_p(n-1,\sum)$-distribution, and $\sqrt{n}\bar{\mathbf{X}}$ and ZZ^t are independent. Recalling that $Y=(Z\vdots\sqrt{n}\bar{X})$, we obtain $YY^t = ZZ^t + n\bar{\mathbf{X}}\bar{\mathbf{X}}^t$. Since $Y = X\mathbf{P}$, we have

$$\begin{aligned} ZZ^t &= YY^t - n\bar{\mathbf{X}}\bar{\mathbf{X}}^t = XPP^tX - n\bar{\mathbf{X}}\bar{\mathbf{X}}^t \\ &= X(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)X^t = S. \end{aligned}$$

Hence $\bar{\mathbf{X}}$ and S are independent, and S has the $W_p(n-1,\sum)$-distribution. Q.E.D.

Henceforth we shall say that X is a sample of size n from a $\mathcal{N}_p(\mu,\sum)$ distribution or population if X is a $p \times n$ matrix and if $X = (\mathbf{X}_1\vdots\cdots\vdots\mathbf{X}_n)$, where the random vectors $\mathbf{X}_1,\cdots,\mathbf{X}_n$ are independent and $\mathcal{N}_p(\mu,\sum)$.

THEOREM 2. *If X is a sample of size $n (n > p)$ on $\mathcal{N}_p(\mu,\sum)$, and if T^2 is defined by*

$$T^2 = \frac{n-p}{p}n(\bar{\mathbf{X}}-\mu)^tS^{-1}(\bar{\mathbf{X}}-\mu),$$

then T^2 has the $F_{p,n-p}$-distribution.

Proof: By Theorem 1 in Section 1 of this Chapter, $\sqrt{n}\bar{\mathbf{X}}$ is $\mathcal{N}_p(\sqrt{n}\mu, \Sigma), S$ is $W_p(n-1, \Sigma)$ and $\bar{\mathbf{X}}$ and S are independent. Thus by Theorem 1 in Section 3 of Chapter 2, T^2 has the $F_{p,n-p}$-distribution. Q.E.D.

The statistic T^2 in Theorem 2 is referred to as Hotelling's T^2-statistic. In the next chapter we shall explore a few of its applications.

EXERCISES

1. Prove: if P is an $n \times n$ orthogonal matrix whose nth column is $n^{-1/2}\mathbf{1}_n$, then $\mathbf{1}_n^t P = (0 \cdots 0 \sqrt{n})$.

§2. **Simultaneous Confidence Intervals.** Most of our applications of Hotelling's T^2-statistic will be in hypothesis testing. However, many times in connection with hypothesis testing we shall wish to obtain simultaneous confidence intervals for a number of linear functionals of the mean of a multivariate normal distribution. This is especially the case when we reject a null hypothesis. We shall first obtain Scheffé-type simultaneous confidence intervals for *all* linear functionals. Then in a more practical mood we shall obtain Bonferroni-type intervals for a finite number of functionals.

LEMMA 1. *If* $\mathbf{c} \in \mathbf{R}^n$*, then*

$$\max\{\mathbf{c}^t\mathbf{x} : \mathbf{x} \in \mathbf{R}^n, \|\mathbf{x}\| = 1\} = \sqrt{\mathbf{c}^t\mathbf{c}} = \|\mathbf{c}\|.$$

Proof: If $\mathbf{c} = \mathbf{0}$, then the conclusion is obviously true. If $\mathbf{c} \neq \mathbf{0}$, denote $\mathbf{d} = \frac{1}{\|\mathbf{c}\|}\mathbf{c}$. Thus $\|\mathbf{d}\| = 1$. Hence

$$\max\{\mathbf{c}^t\mathbf{x} : \mathbf{x} \in \mathbf{R}^n, \|\mathbf{x}\| = 1\} = \sqrt{\mathbf{c}^t\mathbf{c}}\max\{\mathbf{d}^t\mathbf{x} : \mathbf{x} \in \mathbf{R}^n, \|\mathbf{x}\| = 1\}.$$

By the Cauchy-Schwarz inequality, $|\mathbf{d}^t\mathbf{x}| \leq \|\mathbf{d}\| \cdot \|\mathbf{x}\| = 1$, so $|\mathbf{d}^t\mathbf{x}| \leq 1$, and $|\mathbf{d}^t\mathbf{x}|$ achieves its upper bound when $\mathbf{x} = \mathbf{d}$. Hence the conclusion. Q.E.D.

LEMMA 2. *If* $\mathbf{b} \in \mathbf{R}^n$*, and if* M *is a positive-definite* $n \times n$ *matrix, then*

$$\max\left\{\frac{(\mathbf{b}^t\mathbf{x})^2}{\mathbf{x}^tM\mathbf{x}} : \mathbf{x} \in \mathbf{R}^n \setminus \{0\}\right\} = \mathbf{b}^tM^{-1}\mathbf{b}.$$

Proof: Since M is positive-definite, then so is M^{-1}. Thus there exists a positive-definite matrix $M^{-1/2}$ such that $M^{-1/2}M^{-1/2} = M^{-1}$. Applying Lemma 1,

$$\begin{aligned}
\max\left\{\frac{(\mathbf{b}^t\mathbf{x})^2}{\mathbf{x}^tM\mathbf{x}} : \mathbf{x} \neq \mathbf{0}\right\} &= \max\left\{\frac{(\mathbf{b}^tM^{-1/2}\mathbf{y})^2}{\mathbf{y}^tM^{-1/2}MM^{-1/2}\mathbf{y}} : \mathbf{y} \neq \mathbf{0}\right\} \\
&= \max\left\{\frac{(\mathbf{b}^tM^{-1/2}\mathbf{y})^2}{\|\mathbf{y}\|^2} : \mathbf{y} \neq \mathbf{0}\right\} \\
&= \max\left\{(\mathbf{b}^tM^{-1/2}\zeta)^2 : \|\zeta\| = 1\right\} \\
&= \mathbf{b}^tM^{-1/2}M^{-1/2}\mathbf{b} = \mathbf{b}^tM^{-1}\mathbf{b}.
\end{aligned}$$

Q.E.D.

THEOREM 1. *If X is a sample of size $n > p$ on $\mathcal{N}_p(\mu, \sum)$, and if K is defined by*

$$K = \sqrt{\frac{pF_{\alpha;p,n-p}}{n(n-p)}},$$

then

$$P[\alpha^t\bar{\mathbf{X}} - K\sqrt{\alpha^tS\alpha} \leq \alpha^t\mu \leq \alpha^t\bar{\mathbf{X}} + K\sqrt{\alpha^tS\alpha}, \text{ all } \alpha \in \mathbf{R}^p] = 1 - \alpha.$$

Proof: By Lemma 2,

$$\max\left\{n\frac{n-p}{p}\frac{(\alpha^t(\bar{\mathbf{X}} - \mu))^2}{\alpha^tS\alpha} : \alpha \neq 0\right\} = n\frac{n-p}{p}(\bar{\mathbf{X}} - \mu)^tS^{-1}(\bar{\mathbf{X}} - \mu).$$

Now the right side of this equation is Hotelling's T^2-statistic found in Theorem 2 of Section 3 of Chapter 2. Hence by that theorem,

$$P\left[\max_{\alpha \neq 0} n\frac{n-p}{p}\frac{(\alpha^t(\bar{\mathbf{X}} - \mu))^2}{\alpha^tS\alpha} \leq F_{\alpha;p,n-p}\right] = 1 - \alpha.$$

This yields the conclusion of the theorem. Q.E.D.

If the number of linear functionals of μ is large, then the simultaneous confidence intervals supplied by Theorem 1 may be used. However, if there are but few such functionals, one might be able to obtain shorter simultaneous confidence intervals by means of Bonferroni-type intervals which we now derive.

If X is a sample of size n on $\mathcal{N}(\boldsymbol{\mu},\sum)$, where $1 \leq p < n$, then $\bar{\mathbf{X}}$ is $\mathcal{N}_p(\boldsymbol{\mu},\frac{1}{n}\sum)$ and S is $W_p(n-1,\sum)$. For fixed $\boldsymbol{\alpha} \in \mathbf{R}^p \setminus \{0\}$, it follows that $\boldsymbol{\alpha}^t\bar{\mathbf{X}}$ is $\mathcal{N}(\boldsymbol{\alpha}^t\boldsymbol{\mu},\frac{1}{n}\boldsymbol{\alpha}^t\sum\boldsymbol{\alpha})$, or $\sqrt{n}(\boldsymbol{\alpha}^t\bar{\mathbf{X}}-\boldsymbol{\alpha}^t\boldsymbol{\mu})$ is $\mathcal{N}(0,\boldsymbol{\alpha}^t\sum\boldsymbol{\alpha})$, or $\sqrt{n}\frac{\boldsymbol{\alpha}^t\bar{\mathbf{X}}-\boldsymbol{\alpha}^t\boldsymbol{\mu}}{\sqrt{\boldsymbol{\alpha}^t\sum\boldsymbol{\alpha}}}$ is $\mathcal{N}(0,1)$. By Lemma 2 in Section 3 of Chapter 2, $\frac{1}{\boldsymbol{\alpha}^t\sum\boldsymbol{\alpha}}\boldsymbol{\alpha}^tS\boldsymbol{\alpha}$ has the χ^2_{n-1}-distribution and is independent of $\bar{\mathbf{X}}$. Thus $\frac{\sqrt{n}(\boldsymbol{\alpha}^t\bar{\mathbf{X}}-\boldsymbol{\alpha}^t\boldsymbol{\mu})}{\sqrt{\frac{\boldsymbol{\alpha}^t s\boldsymbol{\alpha}}{(n-1)}}}$ has the t_{n-1}-distribution. Now suppose we have r linear functionals of $\boldsymbol{\mu}$ determined by vectors $\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_r$. Thus if $t_0 > 0$ is a number satisfying $P[Z \leq t_0] = 1-\frac{\alpha}{2r}$, where Z is a random variable with the t_{n-1}-distribution, then our r simultaneous confidence intervals for $\boldsymbol{\alpha}_1^t\boldsymbol{\mu},\cdots,\boldsymbol{\alpha}_r^t\boldsymbol{\mu}$ are

$$\boldsymbol{\alpha}_i^t\bar{\mathbf{X}} \pm t_0\sqrt{\frac{\boldsymbol{\alpha}_i^tS\boldsymbol{\alpha}_i}{n(n-1)}}, 1 \leq i \leq r,$$

respectively. As in the text *Linear Regression Analysis*, one can show that

$$P\left(\bigcap_{j=1}^{r}\left[\boldsymbol{\alpha}_j^t\bar{\mathbf{X}} - t_0\sqrt{\frac{\boldsymbol{\alpha}_j^tS\boldsymbol{\alpha}_j}{n(n-1)}} \leq \boldsymbol{\alpha}_j^t\boldsymbol{\mu} \leq \boldsymbol{\alpha}_j^t\bar{\mathbf{X}} + t_0\sqrt{\frac{\boldsymbol{\alpha}_j^tS\boldsymbol{\alpha}_j}{n(n-1)}}\right]\right) \geq 1-\alpha.$$

Simultaneous confidence intervals obtained in this way are referred to here as Bonferroni-type intervals.

EXERCISES

1. Prove: If $A_1, \cdots, A_r$ are events, if $0 < \alpha < 1$, and if $P(A_i) = 1 - \alpha/r, 1 \leq i \leq r$, then

$$P\left(\bigcap_{j=1}^{r} A_j\right) \geq 1 - \alpha.$$

§3. **Application #1: Test of Hypothesis for the Mean Vector.** Let X be a sample of size n on a $\mathcal{N}_p(\mu, \sum)$ population, i.e., $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$, where the columns of X are independent and each is $\mathcal{N}_p(\mu, \sum)$. It is assumed that $1 \leq p < n$. The mean vector μ and the covariance matrix $\sum$ are unknown. For some given vector $\mu_0 \in \mathbf{R}^p$ we wish to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_{alt} : \mu \neq \mu_0$ with level of significance α. In order to do this we construct the test statistic

$$T^2 = n\frac{n-p}{p}(\bar{\mathbf{X}} - \mu_0)^t S^{-1}(\bar{\mathbf{X}} - \mu_0).$$

When H_0 is true, then by Theorem 2 of Section 3 of Chapter 2, T^2 has the $F_{p,n-p}$-distribution. Clearly T^2 is non-negative. If T^2 were too small, it would be because $\bar{\mathbf{X}}$ is close to μ_0 which would strengthen one's belief that $\mu = \mu_0$. Thus we reject H_0 if $T^2 \geq F_{\alpha;p,n-p}$, and $P_{H_0}[Reject\, H_0] = \alpha$. This is a test of utmost practical importance, and we give two examples of it.

Our first example is a *test of significance on contrasts.* An experimental setting for such a test might be as follows. In biomedical experimental or clinical trials one might take the same measurement on the same subject at p different times $t_1 < t_2 < \cdots < t_p$, and then one might inquire whether the increases that one witnesses are genuine or are just due to chance. Thus one might take n subjects and take measurements on all of them at those very same times. The p measurements on the ith subject form the ith vector observation $\mathbf{X}_i$ which is assumed to be $\mathcal{N}_p(\mu, \sum)$ for unknown μ and $\sum$. The problem then becomes that of testing whether the coordinates of μ are equal or not. Formally, $\mathbf{X}_1, \cdots, \mathbf{X}_n$ will denote a sample of size n on $\mathcal{N}_p(\mu, \sum)$ where μ and $\sum$ are unknown. We wish to test the null hypothesis

$H_0 : \mu_1 = \cdots = \mu_p$ against the alternative H_1: not all μ_i's are equal, with level of significance α. If we denote

$$\mathbf{X}_i = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{pi} \end{pmatrix} \text{ and } \mathbf{Y}_i = \begin{pmatrix} X_{1i} - X_{2i} \\ \vdots \\ X_{1i} - X_{pi} \end{pmatrix},$$

then we note that $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are an (observable) sample of size n on $\mathcal{N}_{p-1}(C\boldsymbol{\mu}, C \sum C^t)$, where C is a $(p-1) \times p$ matrix defined by

$$C = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ \vdots & & & & & \\ 1 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}.$$

Since

$$C\boldsymbol{\mu} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_p \end{pmatrix},$$

we observe that H_0 is true if and only if $C\boldsymbol{\mu} = \mathbf{0}$. Thus if we denote $Y = (\mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n), S_Y = Y(I_n - \frac{1}{n}1_n 1_n^t)Y^t$ and $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$, we shall reject H_0 if

$$T^2 = n\frac{n-(p-1)}{p-1}\bar{\mathbf{Y}}^t S_Y^{-1} \bar{\mathbf{Y}} \geq F_{\alpha;p-1,n-p+1}.$$

Since all differences $\mu_i - \mu_j, i \neq j$, can be obtained via linear functionals of $C\boldsymbol{\mu}$, one can use simultaneous confidence intervals to detect which pairs of coordinates are unequal and which coordinate in each pair is larger.

Our second example deals with the classical one-way analysis of variance problem with equal numbers of observations in the cells but with unequal cell variances. More precisely, let $X_{1i}, \cdots, X_{ri}$ be r independence observations on $\mathcal{N}(\mu_i, \sigma_i^2), 1 \leq i \leq s$, and assume all rs random variables $\{X_{ki}, 1 \leq k \leq r, 1 \leq i \leq s\}$ are independent. We wish to test the null hypothesis $H_0 : \mu_1 = \cdots = \mu_s$ against the alternative, H_{alt} : not all μ_i's are equal, with level of significance α. Since not all σ_i^2 are assumed to be equal we cannot use the classical F-test given in *Linear Regression Analysis.* We must assume now that $r \geq s$. Let $Y_{11} = X_{11} - X_{12}, Y_{21} = X_{11} - X_{13}, \cdots, Y_{s-1,1} = X_{11} - X_{1s}$, and, in general, $Y_{ij} = X_{j1} - X_{j,i+1}$. Now let

$$\mathbf{Y}_1 = \begin{pmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{s-1,1} \end{pmatrix}, \cdots, \mathbf{Y}_r = \begin{pmatrix} Y_{1r} \\ Y_{2r} \\ \vdots \\ Y_{s-1,r} \end{pmatrix}$$

It follows that $\mathbf{Y}_1, \cdots, \mathbf{Y}_r$ are independent random vectors with common distribution $\mathcal{N}_{s-1}(\boldsymbol{\nu}, \sum)$, where

$$\boldsymbol{\nu} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \vdots \\ \mu_1 - \mu_s \end{pmatrix}$$

and $\sum$ is some covariance matrix. Thus our desired test is to consider $\mathbf{Y}_1, \cdots, \mathbf{Y}_r$ as a sample of size r on a $\mathcal{N}_{s-1}(\boldsymbol{\nu}, \sum)$ population and to test $H_0 : \boldsymbol{\nu} = \mathbf{0}$ against the alternative, $H_{alt} : \boldsymbol{\nu} \neq \mathbf{0}$, with level of significance α. Letting Y, S_Y and $\bar{\mathbf{Y}}$ be as defined earlier in this section, we shall reject the null hypothesis if

$$r\frac{r-(s-1)}{s-1}\bar{\mathbf{Y}}^t S_{\bar{Y}}^{-1}\bar{\mathbf{Y}} \geq F_{\alpha;s-1,r-s+1}.$$

If we reject the null hypothesis we may rank $\mu_1, \mu_2, \cdots, \mu_s$ by means of simultaneous confidence intervals discussed in Section 1.

EXERCISES

1. Determine $\sum$ in the second example of Application #1.

2. Prove that $\mathbf{Y}_1$ in the first example of Application #1 is $\mathcal{N}_{p-1}(C\boldsymbol{\mu}, C\sum C^t)$.

§4. **Application #2: Multivariate Paired Comparisons.** Let $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_n)$ be a sample of size n on a $\mathcal{N}_p(\boldsymbol{\mu}, \sum_1)$ population, and let $Y = (\mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n)$ be a sample *of the same size*, n, an a $\mathcal{N}_p(\boldsymbol{\nu}, \sum_2)$ population, where $\boldsymbol{\mu}, \boldsymbol{\nu}, \sum_1$ and $\sum_2$ are unknown and where $n > p \geq 1$. We wish to test the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\nu}$ against the alternative $H_{alt} : \boldsymbol{\mu} \neq \boldsymbol{\nu}$ with level of significance α. We first denote $\mathbf{Z}_i = \mathbf{X}_i - \mathbf{Y}_i, 1 \leq i \leq n$. Then one may easily show that $\mathbf{Z}_1, \cdots, \mathbf{Z}_n$ are independent, all having the $\mathcal{N}_p(\boldsymbol{\mu} - \boldsymbol{\nu}, \sum_1 + \sum_2)$-distribution. If $H_0 : \boldsymbol{\mu} = \boldsymbol{\nu}$ is true, then each $\mathbf{Z}_i$ is $\mathcal{N}_p(0, \sum_1 + \sum_2)$. Hence, denoting

$$T^2 = n\frac{n-p}{p}\bar{\mathbf{Z}}^t S_Z^{-1} \bar{\mathbf{Z}},$$

where $\bar{\mathbf{Z}} = (\mathbf{Z}_1 + \cdots + \mathbf{Z}_n)/n$ and

$$S = (\mathbf{Z}_1 \vdots \cdots \vdots \mathbf{Z}_n)(I_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^t)(\mathbf{Z}_1 \vdots \cdots \vdots \mathbf{Z}_n),$$

we reject H_0 when $T^2 \geq F_{\alpha;p,n-p}$.

If we reject H_0, we may use the Bonferroni-type simultaneous confidence intervals developed in Section 1 to detect which pairs μ_i, ν_i are unequal and, if so, which of μ_i and ν_i is large.

It should be noticed that the same test applies when $\mathbf{X}_i$ and $\mathbf{Y}_i$ are not necessarily independent. Indeed, we may relax the above assumptions to the following: $\begin{pmatrix} \mathbf{X}_1 \\ \cdots \\ \mathbf{Y}_1 \end{pmatrix}, \cdots, \begin{pmatrix} \mathbf{X}_n \\ \cdots \\ \mathbf{Y}_n \end{pmatrix}$ are i.i.d. $\mathcal{N}_{2p}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \cdots \\ \boldsymbol{\nu} \end{pmatrix}, \sum\right)$ where

$$\Sigma = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right),$$

Σ_{11} being a $p \times p$ submatrix, and $\boldsymbol{\mu}, \boldsymbol{\nu}$ and Σ are unknown.

EXERCISES

§5. Application #3. The Multivariate Two-Sample T^2-test. Application #2 was a two-sample test where we were fortunate enough to have equal sample sizes. Such a circumstance is not the usual case. When the two sample sizes are unequal, we must assume and hope that both population covariance matrices are the same. In order to be able to obtain a test in this situation we need the following theorem.

THEOREM 1. *If D_1 is $W_p(m,\sum)$, if D_2 is $W_p(n,\sum)$, and if D_1 and D_2 are independent, then $D_1 + D_2$ is $W_p(m+n,\sum)$.*

Proof: Let $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_m)$ be a sample of size m on a $\mathcal{N}_p(\mathbf{0},\sum)$ population, and let $Y = (\mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n)$ be a sample of size n on the same population, with X and Y independent. Then XX^t is $W_p(m,\sum), YY^t$ is $W_p(n,\sum)$ and XX^t and YY^t are independent. However

$$\begin{aligned} XX^t + YY^t &= \sum_{i=1}^{m} \mathbf{X}_i\mathbf{X}_i^t + \sum_{j=1}^{n} \mathbf{Y}_j\mathbf{Y}_j^t \\ &= (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_m \vdots \mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n)(\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_m \vdots \mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n)^t \end{aligned}$$

which is $W_p(m+n,\sum)$. Q.E.D.

We now state our hypothesis testing problem. Let $\mathbf{X}_1, \cdots, \mathbf{X}_m$ be a sample of size m on $\mathcal{N}_p(\boldsymbol{\mu},\sum)$, let $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ be a sample of size n on $\mathcal{N}_p(\boldsymbol{\nu},\sum)$, and assume that $\boldsymbol{\mu}, \boldsymbol{\nu}$ and $\sum$ are unknown. We wish to test the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\nu}$ against the alternative $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\nu}$ with level of significance α. Let us denote $\bar{\mathbf{X}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{X}_i, S_X = X(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^t)X^t, \bar{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i$ and $S_Y = Y(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^t)Y^t$, where $X = (\mathbf{X}_1 \vdots \cdots \vdots \mathbf{X}_m)$ and $Y = (\mathbf{Y}_1 \vdots \cdots \vdots \mathbf{Y}_n)$. By Theorem 1 of §1 and the fact that X and Y are independent, it follows that $\bar{\mathbf{X}}, \bar{\mathbf{Y}}, S_X$ and

S_Y are independent. By Theorem 1 above, $S_X + S_Y$ is $W_p(m+n-2, \sum)$, and $S_X + S_Y$ and $\bar{\mathbf{X}} - \bar{\mathbf{Y}}$ are independent. We easily verify that

$$Cov(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) = Cov(\bar{\mathbf{X}}) + Cov(\bar{\mathbf{Y}}) = \left(\frac{1}{m} + \frac{1}{n}\right)\sum.$$

Hence

$$\sqrt{\frac{mn}{m+n}}(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - (\mu - \nu)) \text{ is } \mathcal{N}_p(0, \sum)$$

and is also independent of $S_X + S_Y$, If we define

$$T^2(\mu - \nu) = \frac{mn}{m+n}\frac{m+n-p-1}{p}(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \mu + \nu)^t(S_X + S_Y)^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \mu + \nu),$$

then $T^2(\mu - \nu)$ has the $F_{p,m+n-p-1}$-distribution. If $H_0 : \mu = \nu$ is true, then $T^2(0)$ has the $F_{p,m+n-p-1}$-distribution. Thus, a test at level α of H_0 against H_1 would be to reject H_0 when

$$\frac{mn}{m+n}\frac{m+n-p-1}{p}(\bar{\mathbf{X}} - \bar{\mathbf{Y}})^t(S_X + S_Y)^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) \geq F_{\alpha;p,m+n-p-1}.$$

In case we reject the null hypothesis, we shall wish to use simultaneous $100(1-\alpha)\%$ confidence intervals. Referring to Lemma 2 and the proof of Theorem 1 in Section 2, we observe that

$$\begin{aligned}
&\max\left\{\frac{mn}{m+n}\frac{m+n-p-1}{p}\frac{(\alpha^t(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - (\mu - \nu)))^2}{\alpha^t(S_X + S_Y)\alpha} : \alpha \neq 0\right\} \\
&\quad = \frac{mn}{m+n}\frac{m+n-p-1}{p}(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - (\mu - \nu))^t(S_X + S_Y)^{-1}(\bar{\mathbf{X}} - \bar{\mathbf{Y}} - (\mu - \nu)),
\end{aligned}$$

the right hand side having the $F_{p,m+n-p-1}$-distribution. Thus, simultaneous $100(1-\alpha)\%$ Scheffé-type confidence intervals of $\mu_i - \nu_i, 1 \leq i \leq p$, are of the form

$$\bar{X}_{i\cdot} - \bar{Y}_{i\cdot} \pm \sqrt{s_{ii}\frac{p(m+n)}{mn(m+n-p-1)}F_{\alpha/2;p,m+n-p-1}},$$

where s_{ii} denotes the ith diagonal element in $S_X + S_Y, \bar{X}_{i\cdot}$ is the arithmetic mean of the ith coordinates of $\bar{X}_1, \cdots, \bar{X}_n$ and $\bar{Y}_{i\cdot}$ is the corresponding function of $\bar{Y}_1, \cdots, \bar{Y}_n$.

In order to obtain Bonferroni-type intervals for $\mu_i - \nu_i, 1 \leq i \leq p$, which should be shorter in this case than Scheffé-type intervals, we first observe that $\sqrt{\frac{mn}{m+n}}((\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - (\boldsymbol{\mu} - \boldsymbol{\nu}))$ is $\mathcal{N}_p(0, \sum)$. Denoting $\sum = (\sigma_{ij})$, and using the notation from above, we have:

$$\sqrt{\frac{mn}{m+n}}((\bar{X}_{i\cdot} - \bar{Y}_{i\cdot}) - (\mu_i - \nu_i)) \text{ is } \mathcal{N}(0, \sigma_{ii}).$$

Let s_{Xi}^2 and s_{Yi}^2 denote the sample variances of the ith coordinates of $\mathbf{X}_1, \cdots, \mathbf{X}_m$ and of $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ respectively. Then it follows that

$$\frac{\sqrt{\frac{mn}{m+n}}((\bar{X}_{i\cdot} - \bar{Y}_{i\cdot}) - (\mu_i - \nu_i))}{\sqrt{\frac{(m-1)s_{Xi}^2 + (n-1)s_{Yi}^2}{m+n-2}}}$$

has the t_{m+n-2}-distribution. Let U be a random variable with this distribution, and let $t_0 > 0$ be such that $P[U \geq t_0] = \alpha/2p$. Then $100(1-\alpha)\%$ simultaneous Bonferroni-type intervals for $\mu_1 - \nu_1, \cdots, \mu_p - \nu_p$ are

$$\bar{X}_{i\cdot} - \bar{Y}_{i\cdot} \pm t_0 \sqrt{\frac{m+n}{mn}} \cdot \sqrt{\frac{(m-1)s_{Xi}^2 + (n-1)s_{Yi}^2}{m+n-2}}, 1 \leq i \leq p.$$

EXERCISES

§6. **Application #4. Linear Hypotheses.** We are given a sample of size n, call it $\mathbf{X}_1,\cdots,\mathbf{X}_n$, on $\mathcal{N}_p(\boldsymbol{\mu},\sum)$, where $\boldsymbol{\mu}$ and $\sum$ are unknown and where $1 \le p < n$. We are also given r known linearly independent vectors $\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_r$, where $1 \le r < p$. Our problem is twofold. We first wish to test the null hypothesis that $\boldsymbol{\mu}$ is a linear combination of $\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_r$, i.e., do there exist constants $\theta_1,\cdots,\theta_r$ such that $\boldsymbol{\mu} = \sum_{i=1}^r \theta_i\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_1\vdots\cdots,\vdots\boldsymbol{\alpha}_r)\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \begin{pmatrix}\theta_1\\ \vdots\\ \theta_r\end{pmatrix}$? Our second problem arises if we have not rejected the above null hypothesis; in this case we wish to find unbiased estimates and confidence intervals for $\theta_1,\cdots,\theta_r$.

An example of this that might arise is as follows. We have n observations $\mathbf{X}_1,\cdots,\mathbf{X}_n$ on a $\mathcal{N}_5(\boldsymbol{\mu},\sum)$ population where $n > 5$, and we wish to test $H_0 : \mu_1 = \mu_2 = \mu_3$ and $\mu_4 = \mu_5$ against the obvious alternative. In this case we let

$$\boldsymbol{\alpha}_1 = \begin{pmatrix}1\\1\\1\\0\\0\end{pmatrix} \text{ and } \boldsymbol{\alpha}_2 = \begin{pmatrix}0\\0\\0\\1\\1\end{pmatrix},$$

and thus we test the equivalent null hypothesis that $\boldsymbol{\mu} = \theta_1\boldsymbol{\alpha}_1 + \theta_2\boldsymbol{\alpha}_2$ for some θ_1,θ_2. This is merely an example, because no new theory is needed to solve this problem. Indeed, let

$$C = \begin{pmatrix}1 & -1 & 0 & 0 & 0\\ 1 & 0 & -1 & 0 & 0\\ 0 & 0 & 0 & 1 & -1\end{pmatrix},$$

and denote $\mathbf{Y}_i = C\mathbf{X}_i, 1 \leq i \leq n$. Then $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are i.i.d. $\mathcal{N}_3(\boldsymbol{\nu}, \sum_1)$, where

$$\boldsymbol{\nu} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \\ \mu_4 - \mu_5 \end{pmatrix}$$

and $\sum_1 = C \sum C^t$ is the unknown covariance matrix. The random vectors $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are observable and the test is to reject H_0 if

$$n\frac{n-3}{3}\bar{\mathbf{Y}}^t S_{\bar{Y}}^{-1} \bar{\mathbf{Y}} \geq F_{\alpha;3,n-3}.$$

A more sophisticated example of this will occur in Application #5 in Section 7. It will involve taking observations on a "stochastic process" $\{X(t), a \leq t \leq b\}$, determining if a polynomial of a certain degree is "close to" the expectation $\{E(X(t)), a \leq t \leq b\}$, and, if so, to determine unbiased estimates and shortest confidence intervals of the coefficients.

Now back to our general problem. Let us denote $A = (\boldsymbol{\alpha}_1 \vdots \cdots \vdots \boldsymbol{\alpha}_r)$. Since $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_r$ are linearly independent and $r < p$, it follows that $rank(A) = r$.

LEMMA 1. *Let A be as defined above. Then there exists a $(p-r) \times p$ matrix D of rank $p-r$ such that $DA = 0$, where 0 is the $(p-r) \times r$ matrix composed entirely of zeros.*

Proof: Let L be the r-dimensional linear subspace of $\mathbf{R}^p$ generated by $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_r$, i.e., $L = col.sp.A$. Clearly, $\dim L = r$. We know there exists an orthonormal basis $\mathbf{x}_1, \cdots, \mathbf{x}_p$ of $\mathbf{R}^p$ such that $\mathbf{x}_1, \cdots, \mathbf{x}_r$ is an orthonormal basis of L. Define

$$D = \begin{pmatrix} \mathbf{x}_{r+1}^t \\ \vdots \\ \mathbf{x}_p^t \end{pmatrix}.$$

For $1 \le i \le p-r, \mathbf{x}_{r+i}^t \alpha_j = 0$ for $1 \le j \le r$. This implies $DA = 0$. Q.E.D.

For D as defined in Lemma 1 define $\mathbf{Y}_i = D\mathbf{X}_i, 1 \le i \le n$. Every such $\mathbf{Y}$ has the $\mathcal{N}_{p-r}(D\mu, D\sum D^t)$-distribution, and $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are independent. If the null hypothesis $H_0 : \mu = A\theta$ for some $\theta \in \mathbf{R}^p$ is true, then each $\mathbf{Y}_i$ is $\mathcal{N}_{p-r}(0, D\sum D^t)$. Thus if H_0 is true, if we define T^2 by

$$T^2 = n\frac{n-(p-r)}{p-r}\bar{\mathbf{Y}}^t S_Y^{-1}\bar{\mathbf{Y}},$$

then T^2 has the $F_{p-r,n-(p-r)}$-distribution, and we reject H_0 if $T^2 \ge F_{\alpha;p-r,n-(p-r)}$.

REMARK 1. *The above test does not depend on D; i.e., the value of T^2 does not depend upon D.*

Proof: We first observe that $\bar{\mathbf{Y}} = D\bar{\mathbf{X}}$ and $S_Y = DS_XD^t$. Hence

$$T^2 = n\frac{n-p+r}{p-r}\bar{\mathbf{X}}^t D^t (DS_XD^t)^{-1} D\bar{\mathbf{X}}.$$

Now let D_0 be any other $(p-r) \times p$ matrix of full rank which satisfies $D_0A = 0$. Then D^t and D_0^t have the same column space, namely, the orthocomplement of the column space of A, i.e., $\{\mathbf{x} \in \mathbf{R}^p : \mathbf{x}^t A = 0\}$. Hence there exists a nonsingular $(p-r) \times (p-r)$ matrix C such that $D^t = D_0^t C$. Hence

$$\begin{aligned} T^2 &= n\frac{n-p+r}{p-r}\bar{\mathbf{X}}^t D_0^t C (C^t D_0 S_X D_0^t C)^{-1} C^t D_0 \bar{\mathbf{X}} \\ &= n\frac{n-p+r}{p-r}\bar{\mathbf{X}}^t D_0^t (D_0 S_X D_0^t)^{-1} D_0 \bar{\mathbf{X}}. \end{aligned}$$

Q.E.D.

It should be noted that, given A, D can be obtained by the Gram-Schmidt process. Exercise 2 at the end of this section provides a quick way of obtaining D.

Let us suppose now that we do not reject $H_0 : E\mathbf{X}_i = A\boldsymbol{\theta}, 1 \leq i \leq n$ for some $\boldsymbol{\theta} \in \mathbf{R}^p$. Our next problem is to obtain unbiased estimates and, in particular, shortest simultaneous confidence intervals for $\theta_1, \cdots, \theta_r$.

Recall that by Theorem 1 in Section 2 that

$$P[\mathbf{x}^t\bar{\mathbf{X}} - K\sqrt{\mathbf{x}^tS\mathbf{x}} \leq \mathbf{x}^tA\boldsymbol{\theta} \leq \mathbf{x}^t\bar{\mathbf{X}} + K\sqrt{\mathbf{x}^tS\mathbf{x}}, \text{ all } \mathbf{x} \in \mathbf{R}^p] = 1 - \alpha,$$

where $K = \sqrt{pF_{\alpha;p,n-p}/n(n-p)}$. Now $\mathbf{x}^t\bar{\mathbf{X}}$ is an unbiased estimate of $\mathbf{x}^tA\boldsymbol{\theta}$. Hence in order to find an unbiased estimate of θ_i we must find an $\mathbf{x} \in \mathbf{R}^p$ such that $\mathbf{x}^tA = (0 \cdots 010 \cdots 0)$, where 1 appears in the ith place. There are many such $\mathbf{x}'s$ and thus we might add another requirement, namely, to shorten the confidence interval as much as possible. In other words, we shall wish to find an $\mathbf{x}_i \in \mathbf{R}^p$ which minimizes $\mathbf{x}^tS\mathbf{x}$ and yet satisfies $\mathbf{x}_i^tA = (0 \cdots 010 \cdots 0)$ for all i. It will turn out that such an $\mathbf{x}$ is a random vector, in particular, a function of S. There are at least two solutions to this problem. A best one is due to John Reid which we present below.

Reid slightly generalizes the problem. Given any $\mathbf{v} \in \mathbf{R}^r$, the problem becomes: minimize $\mathbf{x}^tS\mathbf{x}$ subject to $\mathbf{x}^tA = \mathbf{v}^t$. Let us define $P = (A^tS^{-1}A)^{-1}$. Note that P is an $r \times r$ positive definite matrix. Then define $\mathbf{x}_0 = S^{-1}AP\mathbf{v}$. Note that $\mathbf{x}_0 \in \mathbf{R}^p$ and $\mathbf{x}_0$ is a random vector.

REMARK 2. $\mathbf{x}_0$ *satisfies the constraint* $\mathbf{x}^tA = \mathbf{v}^t$.

Proof: One needs only to observe that

$$\mathbf{x}_0^t A = \mathbf{v}^t P^t A^t S^{-1} S S^{-1} A = \mathbf{v}^t (A^t S^{-1} A)^{-1} A^t S^{-1} A = \mathbf{v}^t.$$

Q.E.D.

LEMMA 2. *If* $\mathbf{x}^t A = \mathbf{v}^t$, *then* $\mathbf{x}^t S \mathbf{x} = (\mathbf{x} - \mathbf{x}_0)^t S (\mathbf{x} - \mathbf{x}_0) + \mathbf{v}^t P \mathbf{v}$.

Proof: We first observe that $(\mathbf{x} - \mathbf{x}_0)^t S (\mathbf{x} - \mathbf{x}_0) = \mathbf{x}^t S \mathbf{x} - 2\mathbf{x}_0^t S \mathbf{x} + \mathbf{x}_0^t S \mathbf{x}_0$. But

$$\begin{aligned} \mathbf{x}_0^t S \mathbf{x} &= \mathbf{v}^t P^t A^t S^{-1} S \mathbf{x} = \mathbf{v}^t P^t A^t \mathbf{x} \\ &= \mathbf{v}^t P^t \mathbf{v} = \mathbf{v}^t P \mathbf{v}. \end{aligned}$$

Also, since P is symmetric

$$\begin{aligned} \mathbf{x}_0^t S \mathbf{x}_0 &= \mathbf{v}^t P^t (A^t S^{-1} S S^{-1} A) P \mathbf{v} \\ &= \mathbf{v}^t P P^{-1} P \mathbf{v} = \mathbf{v}^t P \mathbf{v}. \end{aligned}$$

Q.E.D.

THEOREM 1. *The quadratic form* $\mathbf{x}^t S \mathbf{x}$ *is minimized subject to the constraint* $\mathbf{x}^t A = \mathbf{v}^t$ *when* $\mathbf{x} = S^{-1} A (A^t S^{-1} A)^{-1} \mathbf{v}$, *in which case the value of* $\mathbf{x}^t S \mathbf{x}$ *is* $\mathbf{v}^t (A^t S^{-1} A)^{-1} \mathbf{v}$.

Proof: By Lemma 2, $\mathbf{x}^t S \mathbf{x}$ is minimized subject to the constraint when $\mathbf{x} = \mathbf{x}_0$. Q.E.D.

Now, for $1 \le i \le p$, let us denote $\mathbf{v}_i^t = (0 \cdots 010 \cdots 0); \mathbf{v}_i$ is a p-dimensional vector in $\mathbf{R}^r$ in which 1 appears as the ith coordinate, and all other coordinates are zeros. Thus, $\mathbf{x}_i$ defined by

$$\mathbf{x}_i = S^{-1} A (A^t S^{-1} A)^{-1} \mathbf{v}_i,$$

satisfies $\mathbf{x}_i^t A = \mathbf{v}_i^t$ and, subject to this constraint, minimizes $\mathbf{x}^t S\mathbf{x}$. Now $\mathbf{x}_i$ is a function of S and hence is a random vector. We now wish to show that $\mathbf{x}_i^t\bar{\mathbf{X}}$ is an unbiased estimate of θ_i and that the joint $100(1-\alpha)\%$ confidence intervals of $\theta_1,\cdots,\theta_r$ using $\mathbf{x}_1,\cdots,\mathbf{x}_r$ cover θ with probability $\geq 1-\alpha$.

THEOREM 2. *For $1 \leq i \leq r, \mathbf{x}_i^t\bar{\mathbf{X}}$ is an unbiased estimate of θ_i.*

Proof: Note that $\mathbf{x}_i$ is a Borel-measurable function of $S = S_X$, and hence, by the side condition, $P(\cap_{i=1}^r[\mathbf{x}_i(S)^tA\theta = \theta_i]) = 1$. Since $\bar{\mathbf{X}}$ and S are independent, we have

$$\begin{aligned}
E(\mathbf{x}_i(S)^t\bar{\mathbf{X}}) &= E(\mathbf{x}_i(S)^t\bar{\mathbf{X}}I_{[\mathbf{x}_i(S)^tA\theta=\theta_i]}) \\
&= \int_{\mathbf{R}^{p(p+1)/2}} E(\mathbf{x}_i(S)^t\bar{\mathbf{X}}I_{[\mathbf{x}_i(S)^tA\theta=\theta_i]}|S = s)dP \circ S^{-1}(s) \\
&= \int_M E(\mathbf{x}_i(s)^t\bar{\mathbf{X}}I_{[\mathbf{x}_i(s)^tA\theta=\theta_i]})dP \circ S^{-1}(s),
\end{aligned}$$

where M is the set of all positive-definite matrices. (We know that $P[S \in M] = 1$). Now $\mathbf{x}_i(s)$ is defined for all positive-definite $p \times p$ matrices s, and it is defined over M to satisfy $\mathbf{x}_i(s)^tA\theta = \theta_i$. Thus the indicator in the last integral is 1 , and

$$E(\mathbf{x}_i(S)^t\bar{\mathbf{X}}) = \int_M \mathbf{x}_i(s)^tA\theta dP \circ S^{-1}(s) = \int_M \theta_i dP \circ S^{-1}(s) = \theta_i.$$

Q.E.D.

Finally we prove:

THEOREM 3. *The following inequality is true: $P\left(\cap_{i=1}^r[L_i \leq \theta_i \leq U_i]\right) \geq 1-\alpha$, where $L_i = \mathbf{x}_i(S)^t\bar{\mathbf{X}} - K\sqrt{\mathbf{x}_i(S)^tS\mathbf{x}_i(S)}$, $U_i = \mathbf{x}_i(S)^t\bar{\mathbf{X}} + K\sqrt{\mathbf{x}_i(S)^tS\mathbf{x}_i(S)}$ and $K = \sqrt{pF_{\alpha;p,n-p}/n(n-p)}$.*

Proof: Let us denote

$$E = \bigcap_{\mathbf{x} \in \mathbf{R}^p} [\mathbf{x}^t \bar{\mathbf{X}} - K\sqrt{\mathbf{x}^t S \mathbf{x}} \leq \mathbf{x}^t A\theta \leq \mathbf{x}^t \bar{\mathbf{X}} + K\sqrt{\mathbf{x}^t S \mathbf{x}}].$$

By Theorem 1 in Section 2, $P(E) = 1 - \alpha$. Let $\omega \in E$. For this particular $\omega \in E$, we have

$$\mathbf{x}^t \bar{\mathbf{X}}(\omega) - K\sqrt{\mathbf{x}^t S(\omega) \mathbf{x}} \leq \mathbf{x}^t A\theta \leq \mathbf{x}^t \bar{\mathbf{X}}(\omega) + K\sqrt{\mathbf{x}^t S(\omega) \mathbf{x}}$$

holding for all $\mathbf{x} \in \mathbf{R}^p$. Thus it holds also for the vectors $\mathbf{x}_1(S(\omega)), \cdots, \mathbf{x}_r(S(\omega))$ as defined just before the statement of Theorem 2. Thus we have shown that $E \subset \cap_{j=1}^r [L_j \leq \theta_j \leq U_j]$, from which the theorem follows. Q.E.D.

EXERCISES

1. In the proof of Remark 1, prove the statement "...D^t and D_0^t have the same column space..."

2. Let $\alpha_{r+1}, \cdots, \alpha_p$ be vectors in $\mathbf{R}^p$ such that $\alpha_{r+1}, \cdots, \alpha_p$ are linearly independent, let $B = (\alpha_{r+1}|\cdots|\alpha_p)$, and define D by $D^t = B - A(A^tA)^{-1}A^tB$. Prove that $rank(D) = p - r$ and $DA = 0$.

3. Prove the statement made just before Remark 2 which states "Let us define $P = (A^tSA)^{-1}$. Note that P is an $r \times r$ positive-definite matrix."

§7. Application #5. Growth Curves. A stochastic process $\{X_t : t \in [0,T]\}$ is a collection of random variables. In the example just given, it is an uncountable set of random variables; for every real number t in the interval $[0,T], X_t$ is a random variable defined over a fixed probability space. The parameter t frequently refers to time. These random variables are usually not independent. We shall consider here a particular stochastic process called a *Gaussian process.* A Gaussian process $\{X_t : t \in [0,T]\}$ is a stochastic process such that for every integer n and every finite subset $\{t_1, \cdots, t_n\}$ of $[0,T]$ the random vector

$$\mathbf{X} = \begin{pmatrix} X_{t_1} \\ \vdots \\ X_{t_n} \end{pmatrix}$$

has a multivariate normal distribution. The experimental situation is that of selecting n individuals at random and taking a measurement on each individual at each of the same p times $t_1 < \cdots < t_p$. The p measurements taken on the ith individual constitute a p-dimensional random vector

$$\mathbf{X}_i = \begin{pmatrix} X_{t_1 i} \\ \vdots \\ X_{t_p i} \end{pmatrix}.$$

The n individuals are assumed to be independent of each other, and thus $\mathbf{X}_1, \cdots, \mathbf{X}_n$ are i.i.d. $\mathcal{N}_p(\boldsymbol{\mu}, \sum)$ for $\boldsymbol{\mu}, \sum$ unknown. Our problem is to determine the functional form of $E(X_t)$ as a function of $t \in [0,T]$. A good way of doing so is to try to fit a polynomial function of t to $E(X_t)$. This is the basic problem of this section.

Let $1 \leq r < p$. We wish to test the null hypothesis that there exist constants $c_0, c_1, \cdots, c_{r-1}$

such that

$$E(X_t) = c_0 + c_1 t + \cdots + c_{r-1} t^{r-1}, 0 \le t \le T,$$

based on the $n > p$ observations described in the previous paragraph. In terms of these observations, our null hypothesis becomes $H_0 : E(\mathbf{X}_i) = R\mathbf{c}$ for some $\mathbf{c} \in \mathbf{R}^r$, where

$$R = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{r-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{r-1} \\ & & \vdots & & \\ 1 & t_p & t_p^2 & \cdots & t_p^{r-1} \end{pmatrix} \text{ and } \mathbf{c} = \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{r-1} \end{pmatrix}.$$

Thus Application #4 in Section 6 may be used. The first requirement on us is to find a $p \times (p-r)$ matrix D of rank $p - r$ which satisfies $D^t R = 0$.

LEMMA 1. *Let*

$$B = \begin{pmatrix} t_1^r & \cdots & t_1^{p-1} \\ t_2^r & \cdots & t_2^{p-1} \\ & \vdots & \\ t_p^r & \cdots & t_p^{p-1} \end{pmatrix}, \textit{ and define}$$

$D = B - R(R^tR)^{-1}R^tB$ where R is as defined above. Then rank$(B) = p - r$ and $D^tR = \mathbf{O}$.

Proof: One may directly verify that $D^tR = 0$. In order to prove that D has full rank we observe that $(R \vdots B)$ is a $p \times p$ Vandermonde matrix and is therefore non-singular. This implies that R has full rank, and thus R^tR is non-singular, which gives meaning to $(R^tR)^{-1}$ in the definition of D. Now recall that a non-singular matrix remains non-singular if any column is replaced by itself plus a linear combination of other columns. Thus it follows that $(R \vdots B - R(R^tR)^{-1}R^tB)$ is non-singular, i.e., D has full rank. Q.E.D.

Now one may apply the procedures established in Section 6. Let $\mathbf{Y}_i = D^t\mathbf{X}_i, 1 \leq i \leq p$. If the null hypothesis is true, then $E\mathbf{Y}_i = 0$ for $1 \leq i \leq n$. Thus if T^2 is defined by $T^2 = n\frac{n-(p-r)}{p-r}\bar{\mathbf{Y}}^t S_{\bar{Y}}^{-1}\bar{\mathbf{Y}}$, one rejects the null hypothesis if $T^2 \geq F_{\alpha;p-r,n-p+r}$. If we do not reject the null hypothesis, we can estimate $c_0, c_r, \cdots, c_{r-1}$ using the procedure established in Section 6.

EXERCISES

1. Prove that $\det(R\vdots B) \neq 0$ using the fundamental theorem of algebra.

2. Using the fundamental theorem of algebra, prove that the determinant of the $n \times n$ Vandermonde matrix,

$$\begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{n-1} \end{bmatrix},$$

is not zero if and only if $t_1, \cdots, t_n$ are distinct.

CHAPTER 4. INFERENCE ON MULTIVARIATE LINEAR MODELS.

§1. The Multivariate Linear Model. In this chapter we shall define the multivariate linear model. Then we shall obtain the likelihood ratio test of a linear hypothesis followed by some applications.

DEFINITION. *The general multivariate linear model is defined by the equation*

$$Y = X\beta + Z.$$

Here

$$Y = \begin{pmatrix} \mathbf{Y}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Y}_n^t \end{pmatrix} \text{ and } Z = \begin{pmatrix} \mathbf{Z}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Z}_n^t \end{pmatrix}$$

where $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are observable p-dimensional random vectors, and $\mathbf{Z}_1, \cdots, \mathbf{Z}_n$ are unobservable i.i.d. random vectors with common distribution being $\mathcal{N}_p(\mathbf{0}, \sum), \sum$ being unknown. The matrix X is an $n \times k$ matrix of known numbers whose rank is $k < n$, and $\beta = (\beta_{ij})$ is a $k \times p$ matrix of unknown parameters.

In the above-defined linear model, $E(Y) = X\beta$. The model states that each column in EY is in the column space of X. A null hypothesis H_0 to be tested is that all of the column vectors of EY are in some known linear subspace of the column space of X, i.e., H_0 is true when

$$EY = X_0\gamma$$

where X_0 is an $n \times (k-q)$ matrix of known numbers with rank $X_0 = k - q$ and $col.spX_0 \subset col.sp.X$, and γ is a $(k-q) \times p$ matrix of unknown parameters. Naturally $q \geq 1$. We wish to obtain the likelihood ratio test of H_0.

Let Q^t be an $n \times n$ orthogonal matrix such that the first $k-q$ of its columns span $col.sp.X_0$ and the first k of its columns span $col.sp.X$, and let Q be its transpose. Let us define the $n \times p$ random matrix W by

$$W = QY = \begin{pmatrix} \mathbf{W}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{W}_n^t \end{pmatrix}$$

THEOREM 1. *The random vectors* $\mathbf{W}_1, \cdots, \mathbf{W}_n$ *are independent,* $\mathbf{W}_i$ *has the* $\mathcal{N}_p(\zeta_i, \Sigma)$*-distribution for some* $\zeta_i \in \mathrm{R}^p, 1 \leq i \leq n$*, and* $\zeta_i = 0$ *for all* $i \geq k+1$.

Proof: We first note that $EW = E(QY) = QX\beta$. Now $(QX)^t = X^tQ^t$, and since the first k columns of Q^t are orthonormal and span $col.sp.X$ while the last $n-k$ columns of Q^t are orthonormal and orthogonal to $col.sp.X$, it follows that the last $n-k$ columns of X^tQ^t are all null vectors, i.e. the last $n-k$ rows of

$$EW = QX\beta = \begin{pmatrix} \zeta_1^t \\ \vdots \\ \zeta_n^t \end{pmatrix}$$

contain all zeros. Thus $\zeta_i^t = 0$ for all $i \geq k+1$. We now wish to prove that $\mathbf{W}_1, \cdots, \mathbf{W}_n$ are independent with $\mathbf{W}_i$ being $\mathcal{N}_p(\zeta_i \Sigma)$. It is sufficient to prove that the transposes of the rows of QZ are i.i.d. $\mathcal{N}_p(0,\Sigma)$. Thus, without loss of generality we may now take $W = QZ$, where

$$W = \begin{pmatrix} \mathbf{W}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{W}_n^t \end{pmatrix} \text{ and } Z = \begin{pmatrix} \mathbf{Z}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Z}_n^t \end{pmatrix}.$$

Note that $Q^tW = Z$ and $Z^t = W^tQ$. Let us denote

$$w = \begin{pmatrix} \mathbf{w}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{w}_n^t \end{pmatrix} \text{ and } z = \begin{pmatrix} \mathbf{z}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{z}_n^t \end{pmatrix},$$

which we may take to be $n \times p$ matrices. The joint density of Z is

$$\begin{aligned}
f_Z(z) &= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2}\sum_{j=1}^{n} \mathbf{z}_j^t \Sigma^{-1} \mathbf{z}_j \\
&= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} tr \sum_{j=1}^{n} \mathbf{z}_j^t \Sigma^{-1} \mathbf{z}_j \\
&= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2}\sum_{j=1}^{n} tr(\Sigma^{-1} \mathbf{z}_j \mathbf{z}_j^t) \\
&= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} tr \left(\Sigma^{-1} \sum_{j=1}^{n} \mathbf{z}_j \mathbf{z}_j^t\right) \\
&= \frac{(\det \Sigma^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} tr(\Sigma^{-1} z^t z).
\end{aligned}$$

Now $z^t = w^t Q$ or $z = Q^t w$, so

$$|\det \frac{\partial z}{\partial w}| = |\det \begin{pmatrix} Q^t & \vdots & 0 & \vdots & \cdots & \vdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & \vdots & Q^t & \vdots & \cdots & \vdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ \vdots & \vdots & \vdots & & & \vdots & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ 0 & \vdots & 0 & \vdots & & \vdots & Q^t \end{pmatrix} | = 1.$$

Hence

$$f_W(w) = f_Z(Q^t w) = \frac{(\det \sum^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2} tr(\sum^{-1} w^t w).$$

This is the same density as for Z, and hence we may conclude that $\mathbf{W}_1, \cdots, \mathbf{W}_n$ are independent, p-variate normal with the same covariance matrix $\sum$. Q.E.D.

COROLLARY TO THEOREM 1. *If H_0 is true, then $\zeta_i = 0$ for all $i > k - q$.*

Proof: This is obtained in the same way as in the proof of $\zeta_i = 0$ for $i \geq k+1$ in Theorem 1. Q.E.D.

As in the case of the univariate linear regression model, the hypothesis $h_0 : EY = X_0\gamma$ is equivalent to $H_0 : C\beta = 0$ where C is a $q \times k$ matrix whose rank is $q, q < k$.

THEOREM 2. *The likelihood ratio test of $H_0 : C\beta = 0$ against $H_1 : C\beta \neq 0$ is to reject H_0 when*

$$T = \frac{\det(\sum_{i=k+1}^{n} \mathbf{W}_i \mathbf{W}_i^t)}{\det(\sum_{i=k-q+1}^{n} \mathbf{W}_i \mathbf{W}_i^t)} \leq c,$$

where c is some constant.

Proof: By Theorem 1 the joint density of W, under the assumption that Y satisfies the linear model, is

$$f_W(w) = \frac{(\det \sum^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2}\left\{\textstyle\sum_{i=1}^{k}(\mathbf{w}_i - \zeta_i)^t\sum^{-1}(\mathbf{w}_i - \zeta_i) + \sum_{i=k+1}^{n}\mathbf{w}_i^t\sum^{-1}\mathbf{w}_i\right\}.$$

Let us denote

$$D = \sum_{i=1}^{k}(\mathbf{w}_i - \zeta)(\mathbf{w}_i - \zeta_i)^t + \sum_{i=k+1}^{n} \mathbf{w}_i\mathbf{w}_i^t.$$

Then

$$f_W(w) = \frac{(\det \sum^{-1})^{n/2}}{(2\pi)^{np/2}} \exp -\frac{1}{2}tr(\sum^{-1}D).$$

Now recall Lemma 7 in Section 1 of Chapter 2: If C and D are positive definite $p \times p$ matrices, D being constant, and if

$$f(C) = \frac{1}{2}N \ln \det C - \frac{1}{2}tr\ CD,$$

then f is maximized at $C = ND^{-1}$, and this maximum value is

$$f(ND^{-1}) = \frac{1}{2}Np \ln N - \frac{1}{2}N \ln \det D - \frac{1}{2}Np.$$

Taking logarithms of the density of W we obtain

$$\ln f_W(w) = \frac{1}{2}n \ln \det(\sum^{-1}) - \frac{1}{2}tr(\sum^{-1}D) - \frac{1}{2}np \ln(2\pi).$$

Using the above-recalled lemma, $f_W(w)$ is maximized at the same value of w as in $\ln f_W(w)$. Considering the first two terms on the right hand side of this last equation as $f(\sum^{-1})$ and applying the lemma, maximization occurs when $\sum^{-1} = nD^{-1}$, in which case the maximum of $\ln f_W(w)$, whatever the values of $\zeta_1, \cdots, \zeta_k$ may be, is

$$\max_{\sum^{-1}} \ln f_W(w) = \frac{1}{2}np \ln n - \frac{1}{2}n \ln \det D - \frac{1}{2}np - \frac{1}{2}np \ln 2\pi,$$

or

$$\max_{\sum^{-1}} f_W(w) = \left(\frac{2\pi}{n}\right)^{-\frac{1}{2}np} (\det D)^{-n/2} e^{-np/2}.$$

Thus, whatever, $\zeta_1, \cdots, \zeta_k$ we select, $f_W(w)$ as a function of $\sum^{-1}$ is maximized by selecting $\sum^{-1} = nD^{-1}$. We now wish to select $\zeta_1, \cdots, \zeta_k$ to maximize $\ln f_W(w)$. So, before selecting $\sum^{-1} = nD^{-1}$ we note that $\zeta_1, \cdots, \zeta_k$ should be selected so as to minimize $tr(\sum^{-1} D)$. In order to do this it is sufficient to find $\zeta_1, \cdots, \zeta_k$ that minimizes $tr(\sum^{-1} \sum_{\alpha=1}^{k} (\mathbf{w}_\alpha - \zeta_\alpha)(\mathbf{w}_\alpha - \zeta_\alpha)^t)$. This in turn equals

$$tr\left(\sum_{\alpha=1}^{k} (\mathbf{w}_\alpha - \zeta_\alpha)^t \sum^{-1} (\mathbf{w}_\alpha - \zeta_\alpha)\right) = \sum_{\alpha=1}^{k} (\mathbf{w}_\alpha - \zeta_\alpha)^t \sum^{-1} (\mathbf{w}_\alpha - \zeta_\alpha).$$

This last expression is always equal to or greater than zero since $\sum^{-1}$ is positive definite. Hence minimization of $tr \sum^{-1} D$ occurs when $\zeta_i = \mathbf{w}_i, 1 \leq i \leq k$. Thus the overall maximum is

$$\max f_W(w) = \left(\frac{2\pi}{n}\right)^{-\frac{1}{2}np} \left(\det \sum_{i=k+1}^{n} \mathbf{w}_i \mathbf{w}_i^t\right)^{-n/2} e^{-np/2}.$$

Similarly, when $H_0 : C\beta = 0$ is true (and again using Theorem 1) we have

$$\max f_W(w) = \left(\frac{2\pi}{n}\right)^{-\frac{np}{2}} (\det \sum_{i=k-q+1}^{n} \mathbf{w}_i \mathbf{w}_i^t)^{-n/2} e^{-np/2}.$$

The likelihood ratio test states that H_0 is to be rejected if, for some constant c

$$\lambda = \frac{\max\{f_W(w) : H_0 \text{ is true}\}}{\max\{f_W(w) : \text{ linear model is true}\}} \leq c.$$

This is equivalent to: reject H_0 if

$$T = \frac{\det \sum_{i=k+1}^{n} \mathbf{W}_i \mathbf{W}_i^t}{\det \sum_{i=k-q+1}^{n} \mathbf{W}_i \mathbf{W}_i^t} \leq c.$$

Q.E.D.

NOTE: $\sum_{i=k+1}^{n} \mathbf{W}_i \mathbf{W}_i^t$ has the $W_p(n-k, \sum)$ distribution, and, when H_0 is true

$$\sum_{i=k-q+1}^{n} \mathbf{W}_i \mathbf{W}_i^t = \sum_{i=k-q+1}^{k} \mathbf{W}_i \mathbf{W}_i^t + \sum_{i=k+1}^{n} \mathbf{W}_i \mathbf{W}_i^t$$

is the sum of two independent random matrices, the first being $W_p(q, \sum)$ and the second being $W_p(n-k, \sum)$ and is the statistic that appears in the numerator of T as defined in Theorem 2. In other words,

$$T = \frac{\det(B)}{\det(A+B)},$$

where

$$A = \sum_{i=k-q+1}^{k} \mathbf{W}_i \mathbf{W}_i^t \text{ and } B = \sum_{i=k+1}^{n} \mathbf{W}_i \mathbf{W}_i^t.$$

Theorem 2 presents us with two problems:

(i) How can we compute the value of T from an observation on Y? We do not know the value of Q and do not even know whether the value of T depends on Q.

(ii) Assuming a solution can be found for (i), how does one determine the distribution of T under H_0?

In the next section we shall obtain a solution to problem (i) and shall display an approximation to the distribution of T when H_0 is true.

EXERCISES

§2. The Test Statistic for the Multivariate General Linear Hypothesis. We now turn our attention to solving problem (i) as stated at the end of Section 1. We begin by partitioning Y by columns as well as by rows. We define $\mathbf{Y}^i$ by

$$Y = \begin{pmatrix} \mathbf{Y}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Y}_n^t \end{pmatrix} = (\mathbf{Y}^1 \vdots \cdots \vdots \mathbf{Y}^p).$$

We then define $\mathbf{W}^1, \cdots, \mathbf{W}^p$ by $W = (\mathbf{W}^1 \vdots \cdots \vdots \mathbf{W}^p)$. We recall: $W = QY$. Hence

$$W = \begin{pmatrix} \mathbf{W}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{W}_n^t \end{pmatrix} = (\mathbf{W}^1 \vdots \cdots \vdots \mathbf{W}^p) = QY = (Q\mathbf{Y}^1 \vdots \cdots \vdots Q\mathbf{Y}^p)$$

so that $\mathbf{W}^i = Q\mathbf{Y}^i$ for $1 \leq i \leq p$. Let us also denote

$$\mathbf{W}^i = \begin{pmatrix} W_1^i \\ \vdots \\ W_n^i \end{pmatrix}, \mathbf{Y}^i = \begin{pmatrix} Y_1^i \\ \vdots \\ Y_n^i \end{pmatrix} \text{ and } Q^t = (\alpha_1 \vdots \cdots \vdots \alpha_n).$$

Finally, we denote $P_X : \mathbf{R}^n \rightarrow \mathbf{R}^n$ as the operator on $\mathbf{R}^n$ that projects orthogonally onto the column space of X, i.e., for every $\mathbf{x} \in \mathbf{R}^n$,

$$P_X \mathbf{x} = X(X^t X)^{-1} X^t \mathbf{x}.$$

LEMMA 1. $P_X \mathbf{Y}^i = \sum_{\ell=1}^{k} W_\ell^i \alpha_\ell, 1 \leq i \leq p.$

Proof: As observed above, $\mathbf{W}^i = Q\mathbf{Y}^i$, or $\mathbf{Y}^i = Q^t\mathbf{W}^i = (\alpha_1 \vdots \cdots \vdots \alpha_n)\mathbf{W}^i = \sum_{q=1}^n W_q^i \alpha_q$. Since Q^t is an orthogonal matrix, and since $\alpha_1, \cdots, \alpha_k$ are an orthonormal basis of $col.sp.(X)$, it follows that $P_X\mathbf{Y}^i = \sum_{u=1}^k W_u^i \alpha_u$. Q.E.D.

LEMMA 2. *If we denote*

$$S_{ij} = (\mathbf{Y}^i - P_X\mathbf{Y}^i)^t(\mathbf{Y}^j - P_X\mathbf{Y}^j),$$

then $\sum_{v=k+1}^n \mathbf{W}_v\mathbf{W}_v^t = (S_{ij})$.

Proof: Recall that $Q\mathbf{Y}^i = \mathbf{W}^i = \begin{pmatrix} W_1^i \\ \vdots \\ W_n^i \end{pmatrix}$ and $\mathbf{Y}^i = Q^t\mathbf{W}^i = \sum_{w=1}^n W_w^i \alpha_w$. By Lemma 1,

$$P_X\mathbf{Y}^i = \sum_{w=1}^k W_w^i \alpha_w.$$

Thus

$$QP_X\mathbf{Y}^i = \begin{pmatrix} \alpha_1^t \\ \cdots \\ \vdots \\ \cdots \\ \alpha_n^t \end{pmatrix} \left(\sum_{w=1}^k W_w^i \alpha_w\right) = \begin{pmatrix} W_1^i \\ \vdots \\ W_k^i \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Hence

$$QY^i - QP_XY^i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ W^i_{k+1} \\ \vdots \\ W^i_n \end{pmatrix},$$

and for every $i, j, (QY^i - QP_XY^i)^t(QY^j - QP_XY^j) = \sum_{w=k+1}^n W^i_w W^j_w$. Since Q is an orthogonal matrix, it follows that

$$(\mathbf{Y}^i - P_X\mathbf{Y}^i)^t(\mathbf{Y}^j - P_X\mathbf{Y}^j) = \sum_{w=k+1}^{n} W^i_w W^j_w,$$

which appears in the ith row and jth column of $\sum_{u=k+1}^n \mathbf{W}_u\mathbf{W}^t_u$. Q.E.D.

Similarly, if P_0 denotes the orthogonal projection operator on the column space of X_0, we have

$$\sum_{\ell=k-q+1}^{n} \mathbf{W}_\ell\mathbf{W}^t_\ell = (S^0_{ij}),$$

where

$$S^0_{ij} = (\mathbf{Y}^i - P_0\mathbf{Y}^i)^t(\mathbf{Y}^j - P_0\mathbf{Y}^j).$$

Thus we obtain the theorem:

THEOREM 1. *The likelihood ratio test of $H_0 : C\beta = 0$ against the alternative $C\beta \neq 0$ with level of significance α is to reject H_0 if $T \leq c$, where c is a constant satisfying*

$P_{H_0}[T \leq c] = \alpha$, where

$$\begin{aligned} T &= \frac{\det B}{\det(A+B)}, B = (b_{ij}), A + B = (c_{ij}), \\ b_{ij} &= (\mathbf{Y}^i - X(X^tX)^{-1}X^t\mathbf{Y}^i)^t(\mathbf{Y}^j - X(X^tX)^{-1}X^t\mathbf{Y}^j), \end{aligned}$$

and

$$c_{ij} = (\mathbf{Y}^i - X_0(X_0^tX_0)^{-1}X_0^t\mathbf{Y}^j)^t(\mathbf{Y}^j - X_0(X_0^tX_0)^{-1}X_0^t\mathbf{Y}^j).$$

Our attention is next drawn to determination of the null distribution of T, now that we know (by the above theorem) how to compute T.

DEFINITION. *If a matrix B is $W_p(n-q, \sum)$, if $\mathbf{Z}_1, \cdots, \mathbf{Z}_q$ are i.i.d., each $\mathcal{N}_p(\mathbf{0}, \sum)$, if $B, \mathbf{Z}_1, \cdots, \mathbf{Z}_q$ are independent, and if $A = \sum_{r=1}^q \mathbf{Z}_r\mathbf{Z}_r^t$, then Wilks' Lambda criterion $\Lambda(n,p,q)$ is defined by*

$$\Lambda(n,p,q) = \frac{\det(B)}{\det(A+B)}.$$

THEOREM 2. *If $H_0 : C\beta = 0$ is true and if $q > p$ and $n - k > p$, then the distribution of the test statistic T is the same as that of $\Lambda(n - (k-q), p, q)$.*

Proof: This is an immediate Corollary of Theorem 1.

The derivation of the distribution of Wilks' lambda criterion is beyond the scope of this course. However, we present without proof M.S. Bartlett's approximation.

Bartlett's Approximation:

$-(n - \frac{1}{2}(p+q+1)) \ln \Lambda(n,p,q)$ has the χ^2_{pq}-distribution.

Hence the null distribution of T can be approximated as follows: By Theorem 1 and Bartlett's results $-(n-k+q-\frac{1}{2}(p+q+1))\ln T$ has the χ^2_{pq} distribution (approximately).

§3. **One-Way MANOVA:** Testing whether s samples come from the same population. For $i = 1, 2, \cdots, s$, let $\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, \cdots, \mathbf{Y}_{in_i}$ be a sample of size n_i on $\mathcal{N}_p(\mu_i, \sum)$. We wish to test $H_0 : \mu_1 = \cdots = \mu_s$ against the alternative that not all μ_i's are equal. Set

$$Y = \begin{pmatrix} \mathbf{Y}^t_{11} \\ \cdots \\ \mathbf{Y}^t_{12} \\ \cdots \\ \vdots \\ \mathbf{Y}^t_{1n_1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Y}^t_{s1} \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Y}^t_{sn_s} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & & \cdots & 0 \\ \vdots & \vdots & & \vdots & \\ 1 & 0 & & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & \ddots & \vdots & & \\ 0 & \cdots & & 0 & 1 \\ \vdots & & & \vdots & \vdots \\ 0 & \cdots & & 0 & 1 \end{pmatrix} \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1\ \mathit{rows} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2\ \mathit{rows} \\ \quad\vdots \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_s\ \mathit{rows} \end{array}$$

$$\beta = \begin{pmatrix} \mu^t_1 \\ \cdots \\ \vdots \\ \cdots \\ \mu^t_s \end{pmatrix}_{s \times p}, \quad X_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1}, \gamma = (\mu^t)_{1 \times p}$$

where $n = n_1 + \cdots + n_s$. Define

$$T = \frac{\det B}{\det(A+B)}$$

where $A + B = (a_{ij}), B = (b_{ij})$

$$b_{ij} = (\mathbf{Y}^i - X(X^tX)^{-1}X^t\mathbf{Y}^i)^t(\mathbf{Y}^j - X(X^tX)^{-1}X^t\mathbf{Y}^j)$$

and

$$a_{ij} = (\mathbf{Y}^i - X_0(X_0^tX_0)^{-1}X_0^t\mathbf{Y}^i)^t(\mathbf{Y}^j - X_0(X_0^tX_0)^{-1}X_0^t\mathbf{Y}^j).$$

Recall that Wilks' lambda criterion is

(*) $\Lambda(n,p,q) = \frac{\det A}{\det(A+B)}$

where A is $W_p(n-q, \sum)$, and A and B are independent. By Bartlett's approximation, $-(n - \frac{1}{2}(p+q+1)) \ln \Lambda(n,p,q)$ has the χ^2_{pq}-distribution. Now in our situation we interchange the roles of A and B. In (*), $A = \sum_{i=k+1}^{n} \mathbf{W}_i\mathbf{W}_i^t$ where now $k = s$. Hence A is $W_p(n-k, \sum)$; and $A + B = \sum_{i=k-q+1}^{n} \mathbf{W}_i\mathbf{W}_i^t$. Hence, if $x_{1-\alpha}$ is the $(1-\alpha)100$ percentile point of χ^2_{pq}, then (here n is $n - k + q$)

$$P\left[-(n-k+q-\frac{1}{2}(p+q+1)) \ln T \leq x_{1-\alpha}\right] = 1 - \alpha$$

or

$$P\left[\ln T \geq -\frac{x_{1-\alpha}}{(n-k+q-\frac{1}{2}(p+q+1))}\right] = 1 - \alpha.$$

Hence reject H_0 if

$$T \leq \exp - \frac{x_{1-\alpha}}{(n-k+q-\frac{1}{2}(p+q+1))}.$$

§4. **Application of MANOVA:** Testing Independence of two blocks of variables.

Let $\binom{\mathbf{Y}_1}{\mathbf{X}_1}, \cdots, \binom{\mathbf{Y}_n}{\mathbf{X}_n}$ be a sample of size n on $\binom{\mathbf{Y}}{\mathbf{X}}$, which is assumed to have the $\mathcal{N}_{p+q}\left(\binom{\boldsymbol{\nu}}{\boldsymbol{\mu}}, \sum\right)$ distribution, where the dimension of $\mathbf{Y}$ and $\boldsymbol{\nu}$ is p, the dimension of $\mathbf{X}$ and $\boldsymbol{\mu}$ is q, and $p+q<n$. We partition the matrix $\sum$ as follows:

$$\sum = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right) \text{ where } \sum_{11} \text{ is } p \times p.$$

The covariance matrix $\sum$ and the vectors $\boldsymbol{\mu}, \boldsymbol{\nu}$ are assumed to be unknown. Based on the sample indicated above we wish to mest the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are independent random vectors against the alternative that they are not. As noted earlier, $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if $Cov(\mathbf{X}, \mathbf{Y}) = 0$; this is a property of joint normality. Thus we wish to test $H_0 : \sum_{12} = 0$ against $Alt : \sum_{12} \neq 0$, with some level of significance α. We shall show that this can be done within the MANOVA model.

The following theorem is a restatement of Theorem 1 in SEction 3 of Chapter 1.

THEOREM 1. *If* $\binom{\mathbf{Y}}{\mathbf{X}}$ *is as above, and if* $\mathbf{Z}$ *is defined by* $\mathbf{Z} = \mathbf{Y} - E\mathbf{Y} - \sum_{12} \sum_{22}^{-1}(\mathbf{X} - E\mathbf{X})$*, then* $\mathbf{Z}$ *and* $\mathbf{X}$ *are independent random vectors.*

Thus we may write

$$\mathbf{Y} = E\mathbf{Y} + \sum_{12}\sum_{22}^{-1}(\mathbf{X} - E\mathbf{X}) + \mathbf{Z},$$

where $\mathbf{Z}, \mathbf{X}$ are independent.

COROLLARY TO THEOREM 1: *The random vectors* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent if and only if* $\sum_{21} \sum_{22} = \sum_{12}^{t} \sum_{22}^{-1} = 0$.

Proof: If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then $\sum_{21}^t = \sum_{12} = Cov(\mathbf{X},\mathbf{Y}) = 0$, and thus $\sum_{21}\sum_{22}^{-1} = 0$. Conversely, if $\sum_{12}\sum_{22}^{-1} = 0$, then $\mathbf{Y} = E\mathbf{Y} + \mathbf{Z}$, and by Theorem 1, $\mathbf{Z}$ and $\mathbf{X}$ are independent, which now implies $\mathbf{Y}$ and $\mathbf{X}$ are independent. Q.E.D.

We now wish to obtain a test of independence of $\mathbf{X}$ and $\mathbf{Y}$ by testing the null hypothesis $H_0 : \sum_{21}^t \sum_{22}^{-1} = 0$ against the alternative that $\sum_{21}^t \sum_{22}^{-1} \neq 0$ at some level of significance α. We may write

$$Y = X\beta + Z,$$

where

$$Y = \begin{pmatrix} \mathbf{Y}_1^t \\ \mathbf{Y}_2^t \\ \vdots \\ \mathbf{Y}_n^t \end{pmatrix}, X = \begin{pmatrix} 1\ \mathbf{X}_1^t \\ 1\ \mathbf{X}_2^t \\ \vdots\,\vdots \\ 1\ \mathbf{X}_n^t \end{pmatrix}, \beta = \begin{pmatrix} \lambda_1 \cdots \lambda_p \\ \cdots \\ \sum_{21}^t \sum_{22}^{-1} \end{pmatrix} \text{ and } Z = \begin{pmatrix} \mathbf{Z}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{Z}_n^t \end{pmatrix}.$$

By the corollary to Theorem 1 the model under the null hypothesis is $Y = X_0\gamma + Z$, where $X_0 = \mathbf{1}_n, \gamma = (\lambda_1 \cdots \lambda_p)$.

Let us denote $x = \begin{pmatrix} \mathbf{x}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{x}_n^t \end{pmatrix} \in \mathbf{R}^{n\times q}$ and $\mathcal{X} = \begin{pmatrix} \mathbf{X}_1^t \\ \cdots \\ \vdots \\ \cdots \\ \mathbf{X}_n^t \end{pmatrix}$, so that $X = (\mathbf{1}_n \vdots \mathcal{X})$. Note that under the null hypothesis, X and Z are independent random matrices. Now let us define

$Y(x)$ by $Y(x) = X(x)\beta + Z$, where $X(x) = (\mathbf{1}_n \vdots x)$. Then $Y = Y(\mathcal{X})$. Then define $T(x)$ by

$$T(x) = \frac{detB(x)}{det(A(x) + B(x))},$$

where $B(x) = (b_{ij}(x)), A(x){+}B(x) = (c_{ij}(x))$, where if $u_i(x) = \mathbf{Y}^i(x) - X(x)(X(x)^tX(x))^{-1}X(x)^t\mathbf{Y}^i$ and $v_i(x) = \mathbf{Y}^i(x) - X_0(X^tX_0)^{-1}X_0^t\mathbf{Y}^i$ and $Y = (\mathbf{Y}^1 \vdots \cdots \vdots \mathbf{Y}^p)$, then $b_{ij}(x) = u_i(x)^tu_j(x)$ and $c_{ij}(x) = v_i(x)^tv_j(x)$. The test is to reject H_0 if the observed valued of $T(\mathcal{X})$ is "too small". We shall make this more precise. First note that there is a constant c such that whatever value of x that $\mathcal{X}$ takes, then $P_{H_0}[T(x) \leq c] = \alpha$, where c does not depend on x; actually $c = c(n, p, q)$ depends only on Wilks' lambda-criterion. Note that

$$\begin{aligned} P_{H_0}[T(\mathcal{X}) \leq c] &= \int \overset{\cdots}{}_{\mathbf{R}^{n\times q}} \int P_{H_0}([T(\mathcal{X}) \leq c] | \mathcal{X} = \mathrm{x}) f_X(x) dx \\ &= \int \overset{\cdots}{}_{\mathbf{R}^{n\times q}} \int P_{H_0}([T(x) \leq c] f_X(x) dx = \alpha. \end{aligned}$$

Thus, the test is: reject H_0 if $T(\mathcal{X}) \leq c$.

Exercises

1. Prove: If A is any $p \times q$ matrix, then there exists a $(p+q) \times (p+q)$ positive definite matrix

$$\sum = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

with Σ_{11} being $p \times p$, such that $A = \sum_{21}^{t} \Sigma_{22}$.

2. Prove: If $\mathbf{X}$ and $\mathbf{Y}$ are random vectors, if $\binom{\mathbf{X}}{\mathbf{Y}}$ is $\mathcal{N}_{p+q}\left(\binom{\mu}{\nu}, \sum\right)$, where $\sum = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$, and where $\mathbf{X}$ is $\mathcal{N}_p(\mu, \sum_{11})$, then $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if $\sum_{12} = 0$.

CHAPTER 5. DISCRIMINANT ANALYSIS.

§1. **The Fisher Linear Discriminant Function.** Let $\mathbf{X}$ be one observable random vector which one observes <u>once</u>. We assume that $\mathbf{X}$ is either $\mathcal{N}_p(\mu, \sum)$ or $\mathcal{N}_p(\nu, \sum)$, where $\sum, \mu$ and $\nu \neq \mu$ are <u>known</u>. We wish to decide, based upon one observation on $\mathbf{X}$, whether $E\mathbf{X} = \mu$ or whether $E\mathbf{X} = \nu$. We wish to do this with the smallest possible misclassification probabilities.

We shall do this via the Neyman-Pearson fundamental lemma. Consider the problem to test $H_0 : E\mathbf{X} = \mu$ against the alternative $H_1 : E\mathbf{X} = \nu$. Accordingly the test is: Reject H_0 in favor of H_1 if, upon observing $\mathbf{X} = \mathrm{x}$,

$$\frac{\frac{\sqrt{det \sum^{-1}}}{(2\pi)^{p/2}} exp - \frac{1}{2}(\mathrm{x} - \nu)^t \sum^{-1}(\mathrm{x} - \nu)}{\frac{\sqrt{det \sum^{-1}}}{(2\pi)^{p/2}} exp - \frac{1}{2}(\mathrm{x} - \mu)^t \sum^{-1}(\mathrm{x} - \mu)} \geq \text{(some) } c,$$

or, equivalently, when

$$(\mathrm{x} - \mu)^t \sum^{-1}(\mathrm{x} - \mu) - (\mathrm{x} - \nu)^t \sum^{-1}(\mathrm{x} - \nu) \geq \text{(some) } c'.$$

In other words, we wish to reject H_0 if the inequality

$$(\nu - \mu)^t \sum^{-1}\mathbf{X} \geq \text{(some) } C$$

occurs. We wish to determine the value of C for which both errors of misclassification are equal and minimized. Letting P_μ and P_ν denote probability when $\mathbf{X}$ is $\mathcal{N}_p(\mu, \sum)$ and $\mathcal{N}_p(\nu, \sum)$ respectively, we must have

$$P_\mu[Reject\, H_0] = P_\nu[Accept\, H_0].$$

This means that C must satisfy the basic equation:

$$P_{\mu}[(\nu-\mu)^t\Sigma^{-1}\mathbf{X} \geq C] = P_{\nu}[(\nu-\mu)^t\Sigma^{-1}\mathbf{X} \leq C].$$

Now

$$(\nu-\mu)^t\Sigma^{-1}\mathbf{X} = (\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\mu) + (\nu-\mu)^t\Sigma^{-1}\mu$$

and

$$(\nu-\mu)^t\Sigma^{-1}\mathbf{X} = (\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\nu) + (\nu-\mu)^t\Sigma^{-1}\nu.$$

Thus the basic equation above becomes;

$$P_{\mu}[(\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\mu) \geq C-(\nu-\mu)^t\Sigma^{-1}\mu] = P_{\nu}[(\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\nu) \leq C-(\nu-\mu)^t\Sigma^{-1}\nu].$$

Now $(\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\mu)$ has the same distribution function under P_{μ} as does $(\nu-\mu)^t\Sigma^{-1}(\mathbf{X}-\nu)$ under P_{ν}. Hence

$$C-(\nu-\mu)^t\Sigma^{-1}\mu = -(C-(\nu-\mu)^t\Sigma^{-1}\nu).$$

Solving for C we obtain:

$$C = \frac{1}{2}(\nu-\mu)^t\Sigma^{-1}(\mu+\nu).$$

Thus our classification rule must be:

(i) choose $E\mathbf{X} = \mu$ if

$$(\nu-\mu)^t\Sigma^{-1}\mathbf{X} \leq \frac{1}{2}(\nu-\mu)^t\Sigma^{-1}(\nu+\mu),$$

and

(ii) choose $E\mathbf{X} = \boldsymbol{\nu}$ if

$$(\boldsymbol{\nu} - \boldsymbol{\mu})^t \Sigma^{-1}\mathbf{X} \geq \frac{1}{2}(\boldsymbol{\nu} - \boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{\nu} + \boldsymbol{\mu}).$$

The random variable $(\boldsymbol{\nu} - \boldsymbol{\mu})^t \Sigma^{-1} \mathbf{X}$ is called Fisher's linear discriminant function.

In practice, one sometimes has a large number of observations $\mathbf{X}_1, \cdots, \mathbf{X}_m$ on $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ and a large number $Y_1, \cdots, \mathbf{X}_n$ on $\mathcal{N}_p(\boldsymbol{\nu}, \Sigma)$. From these one estimates $\boldsymbol{\mu}, \boldsymbol{\nu}$ and Σ from formulae obtained earlier in these lectures and substitutes these estimates in (i) and (ii) above.

§2. **Discriminant Analysis.** We now consider the situation where we have two fixed samples at our disposal. One sample consists of $\mathbf{X}_1, \cdots, \mathbf{X}_m$ which are assumed to be independent and $\mathcal{N}_p(\mu, \sum_1)$, and the second sample consists of $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$, also independent and $\mathcal{N}_p(\nu, \sum_2)$. We assume that $\mu \neq \nu, p < \min\{m, n\}$, and the parameters $\mu, \nu, \sum_1$ and $\sum_2$ are unknown, and all $m+n$ observations are independent. Now let $\mathbf{Z}$ be another observation that is independent of all $m+n$ observations and is known to be either $\mathcal{N}_p(\mu, \sum_1)$ or $\mathcal{N}_p(\nu, \sum_2)$. The problem here is to decide whether $\mathbf{Z}$ is $\mathcal{N}_0(\nu, \sum_1)$ or $\mathcal{N}_p(\nu, \sum_2)$. Let us use the following standard notation:

$$\begin{aligned} X &= (\mathbf{X}_1 | \cdots | \mathbf{X}_m), \quad Y = (\mathbf{Y}_1 | \cdots | \mathbf{Y}_n), \\ \bar{\mathbf{X}} &= \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i, \quad S_x = X(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m^t) X^t, \\ \bar{\mathbf{Y}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \text{ and } S_y = Y(I_n - \tfrac{1}{n} \mathbf{1}_n \mathbf{1}_n^t) Y^t, \end{aligned}$$

where I_m is the $m \times m$ identity matrix. It is known that $\bar{\mathbf{X}}$ and S_x are independent, and the same holds true for $\mathbf{Y}$ and S_y. Since the two samples are independent, it follows that $\bar{\mathbf{X}}, S_x, \bar{\mathbf{Y}}$ and S_y are four independent sets of random variables.

In order to accomplish the classification of $\mathbf{Z}$, let us define

$$T_x^2 = \frac{m-p}{p} \frac{m}{m+1} (\mathbf{Z} - \bar{\mathbf{X}})^t S_x^{-1} (\mathbf{Z} - \bar{\mathbf{X}})$$

and

$$T_y^2 = \frac{n-p}{p} \frac{n}{n+1} (\mathbf{Z} - \bar{\mathbf{Y}})^t S_y^{-1} (\mathbf{Z} - \bar{\mathbf{Y}}).$$

PROPOSITION 1. *If* $\mathbf{Z}$ *is* $\mathcal{N}_p(\mu, \sum_1)$*, then* T_x^2 *has the F-distribution with* $(p, m-p)$ *degrees of freedom, and if* $\mathbf{Z}$ *is* $\mathcal{N}_p(\nu, \sum_2)$*, then* T_y^2 *has the F-distribution with* $(p, n-p)$ *degrees of freedom.*

Proof: If $\mathbf{Z}$ is $\mathcal{N}_p(\mu, \sum_1)$, then since $\bar{\mathbf{X}}$ is $\mathcal{N}_p(0, \frac{1}{m}\sum_1)$ and $\mathbf{Z}$ and $\bar{\mathbf{X}}$ are independent, it follows that $\mathbf{Z} - \bar{\mathbf{X}}$ is $\mathcal{N}_p(0, \frac{m+1}{m}\sum_1)$. Also, S_x has the Wishart $W_p(m-1, \sum_1)$-distribution and is independent of $\mathbf{Z} - \bar{\mathbf{X}}$, and thus by the result recalled in Section 1, it follows that T_x^2 has the F-distribution with $(p, m-p)$ degrees of freedom. The proof of the second statement is similar. Q.E.D.

Our crude discriminant analysis is as follows. Let $F_{m,p}$ be the distribution function of the F-distribution with $(p, m-p)$ degrees of freedom, and let $F_{n,p}$ be the same for $(p, n-p)$ degrees of freedom. Then observe $\alpha_x = 1 - F_{m,p}(T_x^2)$ and $\alpha_y = 1 - F_{n,p}(T_y^2)$. The rule proposed is this: if $\alpha_x > \alpha_y$, then classify $\mathbf{Z}$ as being $\mathcal{N}_p(\mu, \sum_1)$, while if $\alpha_y > \alpha_x$, then classify $\mathbf{Z}$ as being $\mathcal{N}_p(\nu, \sum_2)$. It should be noticed that α_x and α_y are random variables; in Section 3 we shall prove that $P[\alpha_x \neq \alpha_y] = 1$ no matter whether $\mathbf{Z}$ is $\mathcal{N}_p(\mu, \sum_1)$ or $\mathcal{N}_p(\nu, \sum_2)$, thus always affording us (theoretically) the ability to classify $\mathbf{Z}$. The important thing to notice here is the method, which is essentially testing whether $\mathbf{Z}$ is an outlier. We shall show in Section 4 that this method yields the procedure for the Fisher linear discriminant function in the situations where that method is theoretically applicable.

We wish to offer the following supplementary remarks. After having performed the crude discriminant analysis, one might require a more refined analysis. It might occur that both significance probabilities α_x and α_y are too large. In such a case one decides ahead of time on a maximum significance probability α_0. If both $\alpha_x > \alpha_0$ and $\alpha_y > \alpha_0$, then we agree not to classify without further information, while if at least one of the inequalities $\alpha_x \leq \alpha_0, \alpha_y \leq \alpha_0$ is true, then one should classify $\mathbf{X}$ according to which of α_x, α_y is larger. This more refined analysis could be used in situations where one has a cheap but uncertain procedure for classifying, with which the crude discriminant analysis can be applied, and where one also has an expensive and time-consuming procedure for classification which is always or almost always correct. Use of the more refined analysis could then yield a considerable savings, in time and/or expense. Another advantage of the crude discriminant analysis is that it easily lends itself in an obvious manner to classification among $k \geq 3$ populations.

§3. Inequality of P-values. As stated in Section 2, this section is devoted to a proof of the following theorem.

THEOREM 1. *No matter whether* $\mathbf{Z}$ *is* $\mathcal{N}_p(\mu, \sum_1)$ *or* $\mathcal{N}_p(\nu, \sum_2)$*, the random variables* α_x *and* α_y *satisfy:* $P[\alpha_x \neq \alpha_y] = 1$.

For greater simplicity in notation and ease in reading, we shall prove this prove this proposition only in the case $p = 1$; its proof for arbitrary p is essentially the same. In this case we have independent random variables $X_1, \cdots, X_m, Y_1, \cdots, Y_n, Z$, where each X_i is $\mathcal{N}(\mu, \sigma^2)$, each Y_j is $\mathcal{N}(\nu, \tau^2)$ and Z can have either of these two distributions. Let us denote

$$K(\mathbf{X}, Z) = \frac{m}{m+1} \frac{(Z - \bar{X}_m)^2}{s_x^2}$$

and

$$L(\mathbf{Y}, Z) = \frac{n}{N+1} \frac{(Z - \bar{Z}_n)^2}{s_y^2},$$

where $\bar{X}_m$ and s_x^2 are the sample mean and sample variance for $X_1, \cdots, X_m$ and $\bar{Y}_n$ and s_y^2 are the corresponding functions for $Y_1, \cdots, Y_n$. Suppose one observes that $K(\mathbf{X}, Z) = \kappa_x$ and $L(\mathbf{Y}, Z) = \kappa_y$ and computes α_x and α_y which are defined by:

$$\alpha_x = P[K(\mathbf{X}, Z) \geq \kappa_x]$$

when Z is $\mathcal{N}(\mu, \sigma^2)$, and

$$\alpha_y = P[L\mathbf{Y}, Z) \geq \kappa_y]$$

when Z is $\mathcal{N}(\nu, \tau^2)$. One could observe that α_x and α_y are random variables; indeed, if $F(x)$ denotes the distribution function of $K(\mathbf{X}, Z)$ when Z is $\mathcal{N}(\mu, \sigma^2)$, then $\alpha_x = 1 - F(K(\mathbf{X}, Z))$, and if G is the distribution function of $L(\mathbf{Y}, Z)$ when Z is $\mathcal{N}(\nu, \tau^2)$, then $\alpha_y = 1 - G(L(\mathbf{X}, Z))$. We shall have proved Theorem 1 by proving $P[F(K(\mathbf{X}, \mathbf{Z})) = G(L(\mathbf{Y}, Z))] = 0$ when Z is $\mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(\nu, \tau^2)$. In order to do this we need two lemmas.

LEMMA 1. *If U and V are positive random variables whose joint density is positive a.e. over $(0, \infty) \times (0, \infty)$, then U/V has a density which is positive a.e. over $(0, \infty)$.*

Proof: Let $f_U(\cdot), f_V(\cdot)$ and $f_{U,V}(\cdot, \cdot)$ be densities of U, of V and of (U, V), and denote $X = U/V, Y = V$. The Jacobian of the transformation is $J\left(\frac{U,V}{X,Y}\right) = Y$, and thus the joint density of X, Y is $f_{X,Y}(x, y) = f_{U,V}(xy, y)y$. Thus the density of X is $f_X(x) = \int_0^\infty f_{X,Y}(x, y)dy = \int_0^\infty y f_{U,V}(xy, y)dy$. Let us assume that the conclusion of the lemma is not true. Then there exists a Borel set $A \subset (0, \infty)$ of positive Lebesgue measure such that $\int_A f_X(x)dx = 0$. Hence $\int\int_{A\times(0,\infty)} y f_{U,V}(xy, y)dydx = 0$. Since $f_{U,V}$ is a density, this implies $f_{U,V}(xy, y) = 0$ a.e. over $A \times (0, \infty)$. But by Tonelli's theorem,

$$\begin{aligned} \int\int_{A\times(0,\infty)} h f_{U,V}(xy, y)dydx &= \int_0^\infty \left(\int_A f_{U,V}(xy, y)dx\right) dy \\ = \int_0^\infty \left(\int_{\frac{1}{y}A} f_{U,V}(x, y)dx\right) dy &= \int\int_B f_{U,V}(x, y)dxdy \end{aligned}$$

where $B = \{(x, y) : y > 0, x \epsilon \frac{1}{y} A\}$. For fixed $y > 0$, the Lebesgue measure of $\frac{1}{y}A$ is easily shown to be $\frac{1}{y}$ times the Lebesgue measure of A, i.e., it is positive. Thus by Tonelli's theorem the two-dimensional Lebesgue measure of B is positive (finite or infinite). Hente $\int\int_B f_{U,V}(x, y)dxdy > 0$, which by the last string of equalities contradicts the fact that $f_{U,V}(xy, y) = 0$ a.e. over $A \times (0, \infty)$. Q.E.D.

LEMMA 2. *Let F be a distribution function which satisfies (i) $F(0) = 0$, (ii) F is absolutely continuous and (iii) the density, f, of F is positive a.e. over $(0, \infty)$. Let U be a positive random variable with an absolutely continuous distribution. Then the distribution function of $F(U)$ is absolutely continuous.*

Proof: It suffices to show that there is a Borel-Measurable function g such that $P[F(U) \leq z] = \int_0^z g(t)dt$ for $0 \leq z \leq 1$. By (ii) and (iii) F has a unique inverse since $P[F(U) \leq z] = P[U \leq F^{-1}(z)] = \int_0^{F^{-1}(z)} f_U(t)dt$, where f_U is the density of U. Now we consider the change of variable $t = F^{-1}(v)$. From (i), (ii), the limits of integration are from 0 to z. Since $v = F(t)$, we have $dv = f(t)dt$, or $dt = \frac{1}{f(F^{-1}(v))}dv$. Thus

$$P[F(U) \leq z] = \int_0^z f_U(F^{-1}(v))\frac{1}{f(F^{-1}(v))}dv,$$

and

$$g(v) = f_U(F^{-1}(v))\frac{1}{f(F^{-1}(v))}.$$

Q.E.D.

Now we prove the theorem. Let F, G be as defined before Lemma 1, and for each z let us define $F_z(\cdot)$ and $G_z(\dot{})$ as the distribution functions of $K(\mathbf{X}, z)$ and $L(\mathbf{Y}, z)$. By Lemma 1, both $F_z(\cdot)$ and $G_z(\cdot)$ have densities which are zero a.e. over $(-\infty, 0)$ and are positive a.e. over $(0, \infty)$. Also, indpendence of K(**X**,z) and $L(\mathbf{Y}, z)$ imply that their joint density exists and is positive a.e.. Now let $A = [F(K(\mathbf{X}, Z)) = G(L(\mathbf{Y}, Z))]$, and assume that Z is either $\mathcal{N}(\mu, \sigma^2)$ or $\mathcal{N}(\nu, \tau^2)$. Then $P(A) = \int_{-\infty}^{\infty} E(I_A|Z = z)P \circ Z^{-1}(dz)$. Since $Z, \mathbf{X}$ and $\mathbf{Y}$ are independent it follows that

$$E(I_A|Z = z) = P[F(K(\mathbf{X}, z)) = G(L(\mathbf{Y}, z))].$$

By Lemma 2, $F(K(\mathbf{X}, z))$ and $G(L(\mathbf{Y}, z))$ have continuous distibution functions F_z and G_z respectively for each fixed z, and since they are, in addition, independent, we obtrain $E(I_A|Z = z) = 0$ for all z. Hence $P(A) = 0$. Q.E.D.

§4. **Re: The Fisher Linear Discriminant Function.** We now show how our method of analysis, when applied to a situation that Fisher linear discriminant function, actually coincides with it. Our method consists of testing whether $\mathbf{Z}$ is an outlier with respect to each population and then classifying $\mathbf{Z}$ according to the larger P-value. The situation covered by the Fisher linear discriminant function is one where it is assumed that $\mu \neq \nu, \sum_1$ and $\sum_2$ are known and, in addition, that $\sum_1 = \sum_2 = \sum$. (In practice, these parameters are estimated and the estimates are then substituted into the algorithm as if they were the actual parameters.) Thus we do not test whether $\mathbf{Z}$ is an outlier of each of two samples; instead we test whether $\mathbf{Z}$ is an outlier of each of two distributions, $\mathcal{N}_p(\mu, \sum)$ and $\mathcal{N}_p(\nu, \sum)$, where $\mu \neq \nu$ and $\sum$ are known. We recall that the random variable $(\nu - \mu)^t \sum^{-1} \mathbf{Z}$ is called the Fisher linear discriminant function; one decides that $\mathbf{Z}$ is $\mathcal{N}_p(\nu, \sum)$ if

$$(\nu - \mu)^t \textstyle\sum^{-1} \mathbf{Z} > \frac{1}{2}(\nu - \mu)^t \textstyle\sum^{-1}(\nu + \mu),$$

and one decides that $\mathbf{Z}$ is $\mathcal{N}_p(\mu, \sum)$ whenever $(\nu - \mu)^t \sum^{-1} \mathbf{Z} < \frac{1}{2}(\nu - \mu)^t \sum^{-1}(\nu + \mu)$. According to the principle followed by our <u>crude discriminant analysis</u> we would consider τ_μ^2 and τ_ν^2 defined by

$$\tau_\mu^2 = (\mathbf{Z} - \mu)^t \textstyle\sum^{-1}(\mathbf{Z} - \mu)$$

and

$$\tau_\nu^2 = (\mathbf{Z} - \nu)^t \textstyle\sum^{-1}(\mathbf{Z} - \nu).$$

If $\mathbf{Z}$ is $\mathcal{N}_p(\mu, \sum)$ then τ_μ^2 has the chi-square distribution with p degrees of freedom, while if $\mathbf{Z}$ is $\mathcal{N}_p(\nu, \sum)$, then τ_ν^2 has this same chi-squre distribution. Now suppose we observe $\mathbf{Z} = \mathbf{z}$.

According to the crude discriminant analysis method, we would compute

$$\beta_\mu = P[W \geq (\mathbf{z}-\mu)^t\Sigma^{-1}(\mathbf{z}-\mu)]$$

and

$$\beta_\nu = P[W \geq (\mathbf{z}-\nu)^t\Sigma^{-1}(\mathbf{z}-\nu)],$$

where W is a random variable whose distribution is chi-square with p degrees of freedom. Again note that β_μ are β_ν are random variables. If we observe $\beta_\mu > \beta_\nu$, then we would decide that $\mathbf{Z}$ is $\mathcal{N}_p(\mu,\Sigma)$, but if $\beta_\mu < \beta_\nu$ we would decide $\mathbf{Z}$ is $\mathcal{N}_p(\nu,\Sigma)$. Our purpose here is to prove that in this case <u>both procedures coincide</u>.

It is clear that $\beta_\mu > \beta_\nu$ if and only if

$$(\mathbf{z}-\mu)^t\Sigma^{-1}(\mathbf{z}-\mu) < (\mathbf{z}-\nu)^t\Sigma^{-1}(\mathbf{z}-\nu).$$

From the easily established fact that

$$(\mathbf{z}-\mu)^t\Sigma^{-1}(\mathbf{z}-\mu) = (\mathbf{z}-\nu)^t\Sigma^{-1}(\mathbf{z}-\nu) + 2(\nu-\mu)^t\Sigma^{-1}(\mathbf{z}-\nu) + (\nu-\mu)^t\Sigma^{-1}(\nu-\mu),$$

the above inequality becomes

$$2(\nu-\mu)^t\Sigma^{-1}(\mathbf{z}-\nu) + (\nu+\mu)^t\Sigma^{-1}(\nu-\mu) < 0.$$

After some elementary algebra, this inequality becomes $2(\nu-\mu)^t\Sigma^{-1}\mathbf{z} < ((\nu-\mu)^t\Sigma^{-1}(\nu+\mu)$. This proves the equivalence of the two methods.

A comment is in order here. In Morrison (1990) the case of unequal covariance matrices is treated as follows. If $\mathbf{Z}$ is known to be $\mathcal{N}_p(\mu,\Sigma_1)$ or $\mathcal{N}_p(\nu,\Sigma_2)$, where $\mu \neq \nu, \Sigma_1$ and Σ_2

are assumed to be known (but in practice are estimated), one considers the ratio,

$$\lambda(\mathbf{z}) = \varphi(\mathbf{z}) = \varphi(\mathbf{z}|\mu, \Sigma_1)/\varphi(\mathbf{z}|\nu, \Sigma_2),$$

of densities of the two distributions. If one observes that $\lambda(\mathbf{Z}) > 1$, then $\mathbf{Z}$ is classified as belonging to the $\mathcal{N}_p(\mu, \Sigma_1)$-population, and if one observes that $\lambda(\mathbf{Z}) < 1$, then one decides that $\mathbf{Z}$ is $\mathcal{N}_p(\nu, \Sigma_2)$. This is equivalent to the following: classify $\mathbf{Z}$ as being $\mathcal{N}_p(\mu, \Sigma_1)$ if

$$(\mathbf{Z} - \mu)^t \Sigma_1^{-1} (\mathbf{Z} - \mu) < (\mathbf{Z} - \nu)^t \Sigma_2^{-1} (\mathbf{Z} - \nu) - \ln(|\sum_1|/|\sum_2|)$$

and as being $\mathcal{N}_p(\nu, \Sigma_2)$ if the reverse inequality is true. Note that if one uses the method of testing for outliers as suggested in this paper, in this situation one would classify $\mathbf{Z}$ as being $\mathcal{N}_p(\mu, \Sigma_1)$ if

$$(\mathbf{Z} - \mu)^t \Sigma_1^{-1} (\mathbf{Z} - \mu) < (\mathbf{Z} - \nu)^t \Sigma_2^{-1} (\mathbf{Z} - \nu)$$

and would classify $\mathbf{Z}$ as being $\mathcal{N}_p(\nu, \Sigma_2)$ if the reverse inequality is true. Thus the method of classification by testing for outliers differs from the method of classification just mentioned by the term $\ln(|\Sigma_1|/|\Sigma_2|)$.