LECTURE NOTES FOR SURV 616/686

Statistical Methods II

Multivariate Analysis

CLASS #03

OVERVIEW: In this class we will review the properties of the Multivariate Normal distribution, and discuss basic estimation of the mean vector and covariance matrix of the distribution. We will present the Hotelling T^2 statistic, and demonstrate how to use it to make inference about the mean vector. We will also learn how to compare mean vectors from two different samples, with a *pooled* T^2 *statistic* similar to the *univariate pooled t-statistic*.

References

Johnson, Richard A., and Wichern, Dean W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York John Wiley.

Morrison, D.F. (1976). *Multivariate Statistical Methods*. New York McGraw Hill.

Data Set

NLSY: We will make frequent use of the National Longitudinal Survey of Youth, also known as the NLSY. The National Longitudinal Surveys are sponsored by the Bureau of Labor Statistics, and conducted by the Bureau of the Census and NORC (University of Chicago) for the Center for Human Resource Research (Ohio State University). The following general description is taken from the NLS Handbook.

The NLSY is a nationally representative sample of 12,686 young men and young women who were 14 to 22 years of age when they were first surveyed in 1979. During the years since that first interview, these young people have finished their schooling, moved out of their parental homes, made decisions on continuing education and training, entered the labor market, served in the military, married and started families of their own. Data collected during the yearly surveys of the NLSY chronicle and

provide a unique opportunity to study in detail the life course experiences of a group of young adults who can be considered representative of all men and women born in the late 50's and early 60's

The NLSY sampling design enables researchers to study in detail the longitudinal experiences of not only this particular age group of young Americans but to analyze the disparate life course experiences of such groups as women, Hispanics, blacks, and the economically disadvantaged. The NLSY is comprised of three subsamples: (1) a cross-sectional sample of 6,111 youth designed to be representative of the noninstitutional civilian segment of young people living in the U.S. in 1979 and born between January 1, 1957 and December 31, 1964; (2) a supplemental sample of 5,295 youth designed to oversample civilian Hispanic, black, and economically disadvantaged white youth living in the U.S. in 1979 and born between January 1, 1957 and December 31, 1964; and (3) a sample of 1,280 youth designed to represent the population born between January 1, 1957 and December 31, 1961 (who were ages 17-21 as of January 1, 1979) and who were enlisted in the four branches of the military as of September 30, 1978.

1. Multivariate Normal Distribution.

The density of the univariate normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \qquad x \in \Re \tag{1.1}$$

We commonly write $x \in \Re^p \; X \sim N(\mu, \sigma^2)$ to refer to a random variable having such a density function. We reviewed the properties of the univariate normal distribution in SURV 615. The *p-dimensional Multivariate Normal distribution* has density function

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right] \qquad \mathbf{x} \in \Re^p \tag{1.2}$$

We will write a p-dimensional multivariate normal random vector as a column vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \tag{1.3}$$

and write $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to represent that a random variable has the density given in (1.2). Note that the expression

$$(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2 \tag{1.4}$$

taken as a function of $\mathbf{x} \in \Re^p$ gives the equation of an *ellipsoid in p-dimensions* centered at μ. This is often stated that *contours of constant probability* are given by ellipsoids for the multivariate normal distribution since the quadratic form in (1.4) appears in the definition of the multivariate normal density function in (1.2).

The moment properties of the multivariate normal distribution are given by

$$E\{\mathbf{X}\} = \begin{pmatrix} E\{X_1\} \\ \vdots \\ E\{X_p\} \end{pmatrix} = \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \tag{1.5}$$

$$V\{\mathbf{X}\} = E\left\{(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})'\right\} = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$$

where

$$E\left\{(X_i-\mu_i)(X_j-\mu_j)\right\} = \sigma_{ij} \quad \text{for } i,j=1,\ldots,p \tag{1.6}$$

The multivariate normal distribution has several useful properties.

Property 1. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then each component X_i for $i = 1, ..., p$ has a univariate normal distribution, $X_i \sim N(\mu_i, \sigma_{ii})$.

Property 2. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any distinct subset of the components (say *k*) has a *k*-dimensional multivariate normal distribution.

Example 1. Assume

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} \sim N_5 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_{55} \end{pmatrix} \right)$$

then

$$\begin{pmatrix} X_4 \\ X_1 \\ X_5 \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_4 \\ \mu_1 \\ \mu_5 \end{pmatrix}, \begin{pmatrix} \sigma_{44} & \sigma_{41} & \sigma_{45} \\ \sigma_{14} & \sigma_{11} & \sigma_{15} \\ \sigma_{54} & \sigma_{51} & \sigma_{55} \end{pmatrix} \right)$$

Property 3. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{c}' = (c_1, c_2, ..., c_p)$ be a set of fixed vector of constants, then

$$\mathbf{c}'\mathbf{X} \sim N_p(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}).$$

Property 4. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{A}$ be a $k \times p$ matrix of fixed constants where $k \le p$, then

$$\mathbf{A}\mathbf{X} \sim N_k(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}').$$

Property 5. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_p$$

Proofs of all of these properties can be found in the references given above. An extremely useful property of the multivariate normal distribution is that we can easily write down conditional distributions involving sub components of the p-dimensional vector. This is given in Property 6.

Property 6. Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where we partition $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

where $\mathbf{X}_1$ is $k \times 1$, and $\mathbf{X}_2$ is $(p-k) \times 1$, and write

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N_p\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right)$$

where $\boldsymbol{\Sigma}_{12}$ is $k \times (p-k)$. Then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$, written as $(\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2)$, is given by

$$(\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2) \sim N_k\left(\boldsymbol{\mu}_{1|2}(\mathbf{x}_2), \boldsymbol{\Sigma}_{1|2}\right)$$

$$\boldsymbol{\mu}_{1|2}(\mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

An important point to notice is that the conditional mean depends on the conditioning value, while the conditional variance does not.

Example 2. For this example we take an extract of the data taken from the NLSY. In particular we have taken a subsample of 917 respondents who had measurements on three variables: (1) Scholastic Aptitude Test Math Score, (2) Scholastic Aptitude Test Verbal Score, and (3) Armed Forces Qualification Test Percentile Score (AFQT). The AFQT percentage is based on an AFQT raw score which is an *intelligence score*. We define the variables as

X_1 = AFQT Percentile Score

X_2 = SAT Math Score

X_3 = SAT Verbal Score

and assume that

$$\begin{pmatrix} \text{AFQT} \\ \text{SAT Math} \\ \text{SAT Verbal} \end{pmatrix} \sim N_3\left(\begin{pmatrix} 67 \\ 448 \\ 409 \end{pmatrix}, \begin{pmatrix} 652 & 2485 & 2418 \\ 2485 & 14755 & 11031 \\ 2418 & 11031 & 14666 \end{pmatrix}\right)$$

Actually the mean and covariance matrix were estimated by the methods described in the next section, and were not weighted to reflect the differential probabilities of selection. Therefore the estimates can not be considered representative of the population. By Property 2 the distribution of the SAT Math and Verbal scores is given by

$$\begin{pmatrix} \text{SAT Math} \\ \text{SAT Verbal} \end{pmatrix} \sim N_2\left(\begin{pmatrix} 448 \\ 409 \end{pmatrix}, \begin{pmatrix} 14755 & 11031 \\ 11031 & 14666 \end{pmatrix}\right)$$

Note that the correlation between the SAT Math and SAT Verbal is given by

$$Corr\{\text{SAT Math}, \text{SAT Verbal}\} = \frac{11031}{\sqrt{(14755)(14666)}} \approx 0.7499.$$

We can look at the distribution of the SAT Math and Verbal scores conditional upon various percentile scores for AFQT by using Property 6.

$$\left(\begin{pmatrix} \text{SAT Math} \\ \text{SAT Verbal} \end{pmatrix} \middle| \text{AFQT} = 30\right) \sim N_2\left(\begin{pmatrix} 308 \\ 273 \end{pmatrix}, \begin{pmatrix} 5277 & 1809 \\ 1809 & 5693 \end{pmatrix}\right)$$

$$\left(\begin{pmatrix} \text{SAT Math} \\ \text{SAT Verbal} \end{pmatrix} \middle| \text{AFQT} = 50\right) \sim N_2\left(\begin{pmatrix} 384 \\ 347 \end{pmatrix}, \begin{pmatrix} 5277 & 1809 \\ 1809 & 5693 \end{pmatrix}\right)$$

$$\left(\begin{pmatrix} \text{SAT Math} \\ \text{SAT Verbal} \end{pmatrix} \middle| \text{AFQT} = 70\right) \sim N_2\left(\begin{pmatrix} 461 \\ 421 \end{pmatrix}, \begin{pmatrix} 5277 & 1809 \\ 1809 & 5693 \end{pmatrix}\right)$$

Notice that both SAT scores increase as the AFQT scores get larger, which could be interpreted as saying that people with higher IQ scores, tend to get higher SAT scores. Note that we can also compute the correlation between the SAT scores conditional on the AFQT score, and we get

$$Corr\left\{\text{SAT Math},\text{SAT Verbal}\middle|\text{AFQT}\right\}=\frac{1809}{\sqrt{(5277)(5693)}}\approx 0.3300$$

this says that when you condition out the effect of AFQT score, the SAT Math and Verbal scores are less correlated.

2. Estimation of the Mean and Covariance Matrix.

Assume we have an independent random sample $\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_n$ each of which is distributed as a $N_p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ random variable. Define the sample mean vector, and the sample covariance matrix as

(2.1) $$\bar{\mathbf{X}}=n^{-1}\sum_{i=1}^{n}\mathbf{X}_i$$

$$\mathbf{S}=\left(n-1\right)^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'.$$

Compare this to the univariate results, where $\bar{x}=\sum_{i=1}^{n}x_i\Big/n$ estimates μ and $s^2=\left(n-1\right)^{-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ estimates σ^2. The following results hold

(2.2) $$E\left\{\bar{\mathbf{X}}\right\}=\boldsymbol{\mu}$$

$$E\left\{\mathbf{S}\right\}=\boldsymbol{\Sigma}$$

which says that the sample mean vector and the sample covariance matrix are unbiased for the true mean vector and the true covariance matrix. It can be shown that the maximum likelihood estimators $\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}$ are given by

(2.3) $$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$$

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{n-1}{n} \right) \mathbf{S}$$

which is similar to the univariate case, where the sample mean is the maximum likelihood estimator, but the estimator of the variance uses the divisor of *n* instead of *n*-1. Unless otherwise stated we will use **S** as the estimator of Σ.

We next give Result 1 which gives the sampling distribution of the sample mean vector and the sample covariance matrix.

Result 1. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then the following three facts hold:

1. $\bar{\mathbf{X}}$ is distributed as $N_p\left(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}\right)$.
2. (*n*-1)**S** is distributed as a *Wishart random matrix* with parameter $\boldsymbol{\Sigma}$ and degrees-of-freedom n-1.
3. $\bar{\mathbf{X}}$ and **S** are independent.

We do not have time in this class to discuss the Wishart distribution, but is a multivariate analog to the chi-square distribution in the univariate case. Note that Result 1 is similar to the results in the univariate case, since the same mean (vector) is normally distributed, and the sample mean (vector) and sample variance (matrix) are independent. Also note that

(2.4) $$\hat{V}\left\{\bar{\mathbf{X}}\right\} = n^{-1}\mathbf{S}$$

is an unbiased estimator of the covariance matrix of $\bar{\mathbf{X}}$.

In reviewing the univariate case in SURV 615, we looked at testing the hypothesis that the true mean was equal to some specified value, under the assumption that the variance was known. We can do a similar thing in the multivariate case.

Result 2. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, and assume that $\boldsymbol{\Sigma}$ is known, and we want to test the hypothesis that

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 .$$

Then the best test rejects H_0 at the α level when

$$\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \left(\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right) \geq \chi^2_{p,\alpha}$$

where $\chi^2_{p,\alpha}$ denotes the value where a chi-square random variable with p degrees of freedom exceeds the value with probability α.

We can also view the test statistic as a function of $\boldsymbol{\mu}_0$. When this is done, the set of values of $\boldsymbol{\mu}_0$ which yield test statistics less than $\chi^2_{p,\alpha}$ map out the interior of an ellipsoid. This gives us a *confidence ellipsoid.* To make this more clear, we will work an example in the two dimensional case.

Example 3. Consider the data in Example 2, for SAT Math Scores and SAT Verbal scores. Based on the sample of 917 respondents we computed

$$\bar{\mathbf{X}} = \begin{pmatrix} 448 \\ 409 \end{pmatrix} \qquad \mathbf{S} = \begin{pmatrix} 14755 & 11031 \\ 11031 & 14666 \end{pmatrix}$$

Assume, for illustrative purposes, that

$$n^{-1}\boldsymbol{\Sigma} = n^{-1}\mathbf{S}=\begin{pmatrix} 16.0905 & 12.0294 \\ 12.0294 & 15.9935 \end{pmatrix}$$

and want to test the hypothesis that

$$H_0 : \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} = \begin{pmatrix} 450 \\ 400 \end{pmatrix} \quad \text{versus} \quad H_A : \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} \neq \begin{pmatrix} 450 \\ 400 \end{pmatrix}$$

Then

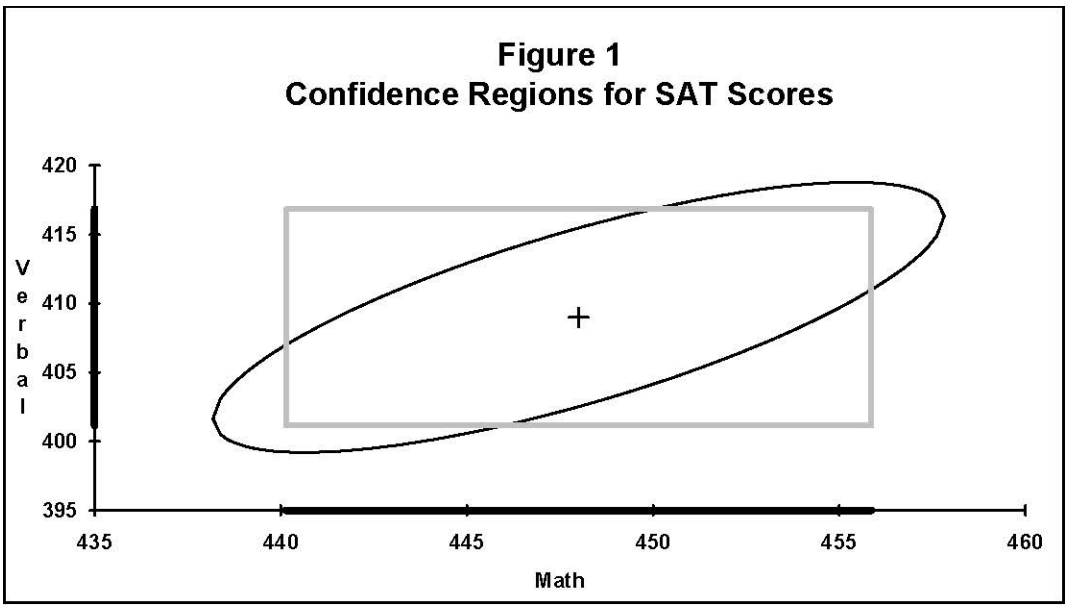
$$\left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} 450 \\ 400 \end{pmatrix} \right)' \begin{pmatrix} 16.0905 & 12.0294 \\ 12.0294 & 15.9935 \end{pmatrix}^{-1} \left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} 450 \\ 400 \end{pmatrix} \right) = 15.98$$

Since $\chi^2_{2,0.05} = 5.99$ we reject H_0.

Now consider the equation

$$\left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} \right)' \begin{pmatrix} 16.0905 & 12.0294 \\ 12.0294 & 15.9935 \end{pmatrix}^{-1} \left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} \right) = 5.99$$

as a function of μ_{Math} and μ_{Verbal}. We have graphed this is Figure 1 below.



The graph of the equation is the ellipse in Figure 1. The + represents the point $\bar{\mathbf{X}}$. Any points inside the ellipse correspond to null hypotheses that would not be rejected by the data at the 5% level. Therefore we can call this a *95% confidence ellipsoid.* It is interesting to compare the confidence ellipsoid to what one would get if you were to form 95% confidence intervals for the SAT Verbal and SAT Math score separately. These confidence intervals are plotted on their respective axes in solid black lines, and the gray box in figure 1 corresponds to the implied joint region the form. It can be shown that the probability of falling in the gray box is *less than 95%,* which says that even though the confidence statements are each true marginally, when taken together (simultaneously) they do not have a combined probability of 95%. This is a problem, because we are used to thinking in terms of rectangular confidence regions, and not in terms of confidence ellipsoids. It is possible to form rectangular confidence regions which have the correct joint coverage probability. We will discuss this in the next section.

In the univariate case we examined a hypothesis test for the variance. There exist similar tests for the covariance matrix in the multivariate case, but we do not have time to examine them. The references above treat the subject in depth.

3. Hotelling's T^2 Statistic.

In the previous section we gave Result 2, which said how to test hypotheses about the mean vector when the covariance matrix is known. In practice the covariance matrix will seldom be known, so we will estimate it by using **S** as an estimate of Σ. In the univariate case we saw how when we replaced a known variance in the denominator of the Z statistic with the sample variance, which had a multiple of a chi-square distribution, we got a random variable which had a *t*-distribution instead of a normal distribution. The analogous thing happens in the multivariate case when we replace the known covariance matrix in Result 2, with the sample covariance matrix which has a multiple of a Wishart distribution which is the multivariate analog of the chi-square distribution. We give this in the next result.

Result 3. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are both unknown. We are interested in testing the hypothesis

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 .$$

Define *Hotelling's T^2 Statistic as*

$$T^2 = \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \left(\frac{1}{n}\mathbf{S}\right)^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)$$

$$= n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)$$

Then under H_0

$$T^2 \sim \frac{(n-1)p}{(n-p)} F^p_{n-p}$$

and the best test of H_0 versus H_A rejects H_0 when

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

where $F^{p}_{n-p}(\alpha)$ is the upper α percentile on the *F*-distribution with numerator degrees-of-freedom *p* and denominator degrees-of-freedom *n-p*.

To demonstrate the use of the T^2 statistics we work Example 4.

Example 4. Again consider the data in Example 2, for SAT Math Scores and SAT Verbal scores where based on the sample of 917 respondents we computed

$$\bar{\mathbf{X}} = \begin{pmatrix} 448 \\ 409 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 14755 & 11031 \\ 11031 & 14666 \end{pmatrix}$$

and want to test the hypothesis that

$$H_0 : \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} = \begin{pmatrix} 450 \\ 400 \end{pmatrix} \text{ versus } H_A : \begin{pmatrix} \mu_{Math} \\ \mu_{Verbal} \end{pmatrix} \neq \begin{pmatrix} 450 \\ 400 \end{pmatrix}$$

Then we get

$$T^2 = (917)\left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} 450 \\ 400 \end{pmatrix}\right)' \begin{pmatrix} 14755 & 11031 \\ 11031 & 14666 \end{pmatrix}^{-1} \left(\begin{pmatrix} 448 \\ 409 \end{pmatrix} - \begin{pmatrix} 450 \\ 400 \end{pmatrix}\right) = 15.98$$

$$\frac{(n-1)p}{(n-p)} F^{p}_{n-p}(0.05) = \frac{(916)2}{915}(3.00) = 6.01$$

so we reject H_0.

We got similar results in Examples 4, and Example 3. This is because of the large sample sizes involved relative to the dimension of the multivariate normal distribution. Notice that

$$\frac{(n-1)\,p}{(n-p)} F^{p}_{n-p} \rightarrow \chi^2_p \quad \text{as} \quad n \rightarrow \infty \tag{3.1}$$

which says that the distribution with known variance is very close to that with estimated variance when the sample size is large.

It is useful to write the Hotelling T^2 statistic as

$$T^2 = \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right)' \left[\hat{V}\left\{\bar{\mathbf{X}}\right\}\right]^{-1} \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_0\right) \tag{3.2}$$

$$T^2 \sim \chi^2_p \quad \text{as} \quad n \rightarrow \infty$$

and it is in this form that it is often used in the analysis of large complex surveys, where $\hat{V}\left\{\bar{\mathbf{X}}\right\}$ denotes the estimated covariance matrix of the sample mean. The estimate $\hat{V}\left\{\bar{\mathbf{X}}\right\}$ could be computed by SUDAAN for example. When it is done in this way, expression (3.2) is usually called a *Wald Statistic*, to test the hypothesis that $\boldsymbol{\mu}_0$ is the true mean vector. The value of the Hotelling T^2 under Normality is that the distributional theory is exact, and can be used even for relatively small samples.

Simultaneous Confidence Intervals

The Hotelling T^2 statistic has an alternative, and interesting motivation. First consider an arbitrary linear combination of the sample mean vector, say

$$\mathbf{c}'\bar{\mathbf{X}} = \sum_{i=1}^{p} c_i \bar{X}_i \tag{3.3}$$

Then

(3.4) $$E\left\{\mathbf{c}'\bar{\mathbf{X}}\right\} = \mathbf{c}'\boldsymbol{\mu}$$

$$\hat{V}\left\{\mathbf{c}'\bar{\mathbf{X}}\right\} = n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}$$

and form the univariate *t*-statistic

(3.5) $$t_c = \frac{\mathbf{c}'\bar{\mathbf{X}} - \mathbf{c}'\boldsymbol{\mu}}{\sqrt{n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}}} = \frac{\mathbf{c}'\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)}{\sqrt{n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}}}$$

and

(3.6) $$t_c^2 = \frac{\mathbf{c}'\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)'\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)\mathbf{c}}{n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}}$$

where the subscript **c** on the t-statistic denotes the dependence on **c**. If we then consider the set of all such t_c^2 as we let **c** vary in $\Re^p$ then it can be shown that

(3.7) $$\max_{\mathbf{c}\in\Re^p}\left(t_c^2\right) = T^2 = n\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}\right)$$

so that the Hotelling T^2 statistic is actually the *largest squared univariate t- statistic* that can be constructed from the data. This leads to the following result for rectangular confidence intervals.

Result 4. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ be a random sample from an $N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ distribution, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are both unknown, and let **c** be any *p*-dimensional column vector. Then simultaneously for all values of **c** in $\Re^p$, the confidence interval

$$\left[\mathbf{c}'\bar{\mathbf{X}} - \sqrt{\frac{(n-1)\,p}{(n-p)} F^{p}_{n-p}(\alpha)} \sqrt{n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}}, \mathbf{c}'\bar{\mathbf{X}} + \sqrt{\frac{(n-1)\,p}{(n-1)} F^{p}_{n-p}(\alpha)} \sqrt{n^{-1}\mathbf{c}'\mathbf{S}\mathbf{c}} \right]$$

will contain $\mathbf{c}'\boldsymbol{\mu}$ with probability $1-\alpha$.

Such intervals are called *simultaneous T^2 confidence intervals.* Result 4 is very useful because it shows how to produce rectangular confidence intervals which have *at least* the correct joint coverage probability. This is particularly useful when exploring the data, without a preset hypothesis. In the next example we demonstrate their use in the two dimensional case.

Example 5. We again examine the data from Example 3. From before we have

$$\bar{X}_{Math} = 448$$

$$\bar{X}_{Verbal} = 409$$

$$\hat{V}\left\{\bar{X}_{Math}\right\} = 16.0905$$

$$\hat{V}\left\{\bar{X}_{Verbal}\right\} = 15.9935$$

univariate 95% Confidence interval for μ_{Math} [440.138, 455.862]

univariate 95% Confidence interval for μ_{Verbal} [401.162, 416.838]

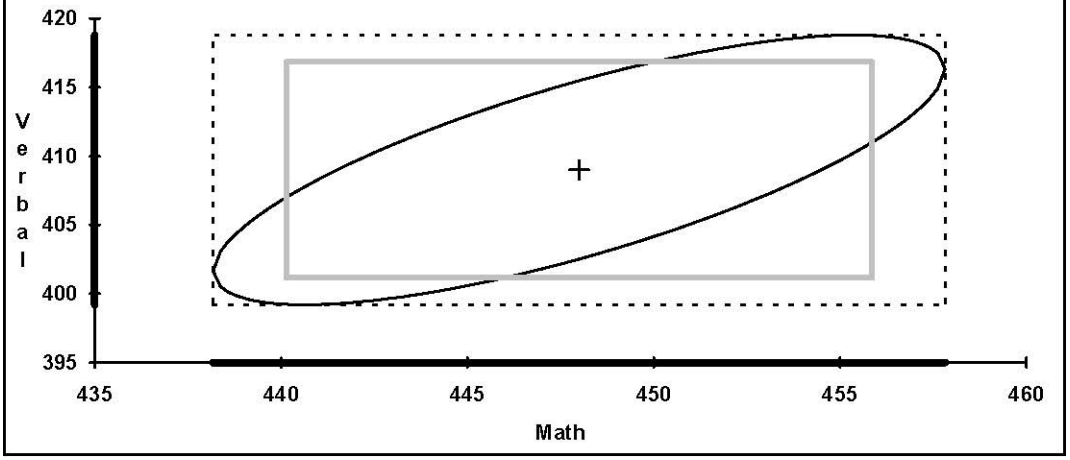
where these were obtained by the usual formula of taking plus/minus 1.96 times the standard error. These were the confidence intervals graphed in Figure 1, which produced the gray box. In Figure 2 below we have reproduced the gray box, but additionally graph the simultaneous confidence intervals

simultaneous 95% Confidence interval for μ_{Math} [438.1825, 457.8174]

simultaneous 95% Confidence interval for μ_{Verbal} [399.2122, 418.7878]

where we used Result 4, taking plus/minus $\sqrt{6.01}$ times the standard error, where we got $\frac{(n-1)p}{(n-1)} F_{n-p}^{p}(\alpha) = 6.01$ from Example 4. These simultaneous confidence intervals are graphed as the solid lines on the axes of Figure 2, and the dashed box represents the joint rectangular interval which they create. Note how the rectangle created by the simultaneous intervals inscribes the 95% confidence ellipsoid.

Figure 2
Simultaneous Confidence Regions for SAT Scores



There are other ways to form rectangular confidence regions which have at least the right joint coverage probability, one of which is called *Bonferroni's Method.*

Pooled Tests

We have been considering a single sample from a multivariate normal distribution, but we can also consider testing the mean vectors for two different multivariate normal distributions. In particular let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{n_1}$ be a random sample of size n_1 from a $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ distribution, and let $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{n_2}$ be a second random sample (independent of the first sample) of size n_2 from a $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ distribution. Notice that we are

assuming that the covariance matrices for the two distributions are the same. Assume we are interested in testing the hypothesis

(3.8) $$H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta} \quad \text{versus} \quad H_A : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}$$

Define

(3.9) $$\bar{\mathbf{X}} = n_1^{-1} \sum_{i=1}^{n_1} \mathbf{X}_i$$

$$\mathbf{S}_1 = (n_1 - 1)^{-1} \sum_{i=1}^{n_1} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

$$\bar{\mathbf{Y}} = n_2^{-1} \sum_{i=1}^{n_2} \mathbf{Y}_i$$

$$\mathbf{S}_2 = (n_2 - 1)^{-1} \sum_{i=1}^{n_2} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})'$$

$$\mathbf{S}_{Pooled} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 + n_2 - 2)}$$

$$T^2 = (\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \boldsymbol{\delta})' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{Pooled} \right]^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}} - \boldsymbol{\delta})$$

Then

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

and the best level α test of the hypothesis rejects H_0 when

$$T^2 > \frac{(n_1+n_2-2)\,p}{(n_1+n_2-p-1)} F^{p}_{n_1+n_2-p-1}(\alpha) \tag{3.11}$$

To demonstrate this we work the following example.

Example 6. We continue the analysis of SAT Math Score, SAT Verbal Scores, and AFQT percentile score begun in Example 2, but now consider the sample of 917 respondents, as composed of two separate independent samples of men and women. Note that strictly this is not correct given the sampling design of the NLSY, but we will use it for illustrative purposes. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{407}$ be a random sample of 407 men with a $N_3(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ distribution, and let $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_{510}$ be a second random sample of 510 women with a $N_3(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ distribution. We are interested in testing the hypothesis that men and women have the same mean vector of scores

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

By calculation we get

$$\bar{\mathbf{X}} = \begin{pmatrix} 72.412776 \\ 484.69287 \\ 425.40541 \end{pmatrix}$$

$$\mathbf{S}_1 = \begin{pmatrix} 570.58782 & 2331.9744 & 2298.0835 \\ 2331.9744 & 15281.125 & 11725.802 \\ 2298.0835 & 11725.802 & 15422.434 \end{pmatrix}$$

$$\bar{\mathbf{Y}} = \begin{pmatrix} 62.186275 \\ 418.80392 \\ 396.45098 \end{pmatrix}$$

$$\mathbf{S}_2 = \begin{pmatrix} 671.18134 & 2312.8558 & 2387.185 \\ 2312.8558 & 12433.341 & 9650.1672 \\ 2387.185 & 9650.1672 & 13718.421 \end{pmatrix}$$

$$\mathbf{S}_{Pooled} = \begin{pmatrix} 626.5464 & 2321.339 & 2347.6492 \\ 2321.339 & 13696.948 & 10571.159 \\ 2347.649 & 10571.159 & 14474.518 \end{pmatrix}$$

$$T^2 = 90.13$$

$$\frac{(n_1 + n_2 - 2)}{(n_1 + n_2 - p - 1)} F^{p}_{n_1+n_2-p-1}(0.05) = \frac{(915)3}{(913)}(2.61) = 7.85$$

so we reject null hypothesis.

In practice one would want to test the hypothesis that the covariance matrices for the two populations were equal, but we will not give the details of how to do that in this class. The references give above contain more information on the likelihood ratio tests that can be used to test for the equality of the two covariance matrices.