4. Inferences about a Mean Vector

4.1 The Plausibility of μ_0 as a Value for a Normal Population Mean

- The univariate theory for determining whether a specific value of μ_0 is a plausible value for the population mean μ.
  - ***Test*** of the competing ***hypotheses***

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

Here H_0 is the null hypothesis and H_1 is the (two-side) alternative hypothesis.

– Rejecting H_0 when $|t|$ is large is equivalent to rejecting H_0 in favor of H_1, at significance level α if

$$t^2 = \frac{(\bar{X} - \mu_0)^2}{s^2/n} = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0) > t^2_{n-1}(\alpha/2)$$

– If H_0 is not rejected, we conclude that μ_0 is a plausible value for the normal population mean. From the well-known correspondence between acceptance region for test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ and confidence interval for μ we have

$$\left\{\text{Do not reject } H_0 : \mu = \mu_0 \text{ at level } \alpha \text{ or } \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| \leq t_{n-1}(\alpha/2)\right\}$$

is equivalent to

$$\left\{\mu_0 \text{ lies in the } 100(1-\alpha)\% \text{ confidence interval } \bar{x} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}\right\}$$

- A natural generalization of the squared distance above is its multivariate analog

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

  The statistics T^2 is called Hotelling's T^2.

- If the observed statistical distance T^2 is too large— that is, if $\hat{\mathbf{x}}$ is "too far" from μ_0—the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected.

- $$T^2 \text{ is distributed as } \frac{(n-1)p}{n-p} F_{p,n-p}$$

  where $F_{p,n-p}$ denotes a random variable with an F-distribution with p and $n - p$ degree of freedom.

To summarize, we have the following:

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \Sigma)$ population. Then with $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j$ and $\mathbf{S} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$,

$$\begin{aligned} \alpha &= P\left[T^2 > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] \\ &= P\left[n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right] \end{aligned}$$

whatever the true $\boldsymbol{\mu}$ and Σ. Here $F_{p,n-p}(\alpha)$ is the upper $(100\alpha)th$ percentile of the $F_{p,n-p}$ distribution.

Example 4.1(Evaluating T^2) Let the data matrix for a random sample of size $n = 3$ from a bivariate normal population be

$$\boldsymbol{X} = \begin{bmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{bmatrix}$$

Evaluate the observed T^2 for $\boldsymbol{\mu}_0' = [9, 5]$. What is the sampling distribution of T^2 in this case ?

Example 4.2 (Testing a multivariate mean vector with T^2) Perspiration from 20 healthy females was analyzed. Three components, X_1 =sweat rate, X_2 =sodium content, and X_3 =potassium content, were measured, and the results, which we call the *sweat data*, are presented in Table 5.1.

Test the hypothesis $H_0 : \boldsymbol{\mu}' = [4, 50, 10]$ against $H_1 : \boldsymbol{\mu}' \neq [4, 50, 10]$ at level of significance $\alpha = .10$.

Table 5.1 Sweat Data

Individual	X_1 (Sweat rate)	X_2 (Sodium)	X_3 (Potassium)
1	3.7	48.5	9.3
2	5.7	65.1	8.0
3	3.8	47.2	10.9
4	3.2	53.2	12.0
5	3.1	55.5	9.7
6	4.6	36.1	7.9
7	2.4	24.8	14.0
8	7.2	33.1	7.6
9	6.7	47.4	8.5
10	5.4	54.1	11.3
11	3.9	36.9	12.7
12	4.5	58.8	12.3
13	3.5	27.8	9.8
14	4.5	40.2	8.4
15	1.5	13.5	10.1
16	8.5	56.4	7.1
17	4.5	71.6	8.2
18	6.5	52.8	10.9
19	4.1	44.1	11.2
20	5.5	40.9	9.4

Source: Courtesy of Dr. Gerald Bargman.

4.2 Confidence Regions and Simultaneous Comparisons of Component Means

- The region $R(\boldsymbol{X})$ is said to be a $100(1-\alpha)\%$ ***confidence region*** if, before the sample is selected,

$$P[(\boldsymbol{X}) \text{ will cover the true } \theta] = 1-\alpha$$

This probability is calculated under the true, but unknown value of $\boldsymbol{\theta}$.

- The confidence region for the mean $\boldsymbol{\mu}$ of a p-dimension normal population is available. Before the sample is selected,

$$P\left[n(\bar{\boldsymbol{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}) \le \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)\right] = 1-\alpha$$

whatever the values of the unknown $\boldsymbol{\mu}$ and Σ. In words, $\bar{\boldsymbol{X}}$ will be within $[(n-1)pF_{p,n-p}(\alpha)/(n-p)]^{1/2}$ of $\boldsymbol{\mu}$, with probability $1-\alpha$, provided that distance is defined in terms of $n\mathbf{S}^{-1}$.

- For $P \geq 4$,we cannot graph the joint confidence region for $\boldsymbol{\mu}$. However we can calculate the axes of the confidence ellipsoid and their relative lengths.

- These are determined from the eigenvalues λ_i and eigenvectors $\mathbf{e}_i$ of $\mathbf{S}$. The direction and lengths of the axes of

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq c^2 = \frac{p(n-1)}{n-p}F_{p,n-p}(\alpha)$$

are determined by going

$$\sqrt{\lambda}c/\sqrt{n} = \sqrt{\lambda_i}\sqrt{p(n-1)F_{p,n-p}(\alpha)/n(n-p)}$$

units along the eigenvectors $\mathbf{e}_i$. Beginning at the center $\bar{\mathbf{x}}$, the axes of the confidence ellipsoid are

$$\pm\sqrt{\lambda_i}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}\mathbf{e}_i \quad \text{where} \quad \mathbf{S}\mathbf{e}_i = \lambda_i\mathbf{e}_i, i = 1, 2, \ldots, p$$

The ratio of the λ_i's will help identify relative amount of elongation along pair of axes

Example 4.3 (Constructing a confidence ellipse for μ) Data for radiation from microwave oven were introduced in Example 3.10 and 3.17. Let

$$x_1 = (\text{measured ration with door closed})^{\frac{1}{4}}$$

and

$$x_2 = (\text{measured ration with door open})^{\frac{1}{4}}$$

Construct the confidence ellipse for $\boldsymbol{\mu} = [\mu_1, \mu_2]$.

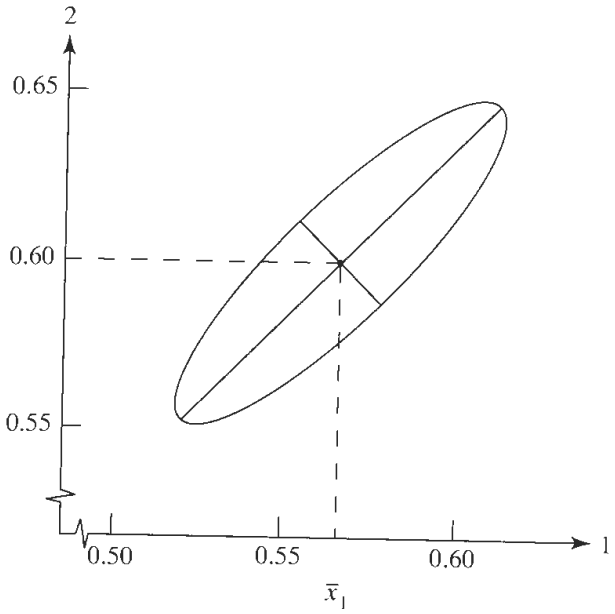

Figure 5.1 A 95% confidence ellipse for μ based on microwave-radiation data.

Simultaneous Confidence Statements

- While the confidence region $n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq c^2$, for c a constant, correctly assesses the joint knowledge concerning plausible value of $\boldsymbol{\mu}$, any summary of conclusions ordinarily includes confidence statement about the individual component means.

- In so doing, we adopt the attitude that all of the separate confidence statements should hold simultaneously with specified high probability.

- It is the guarantee of a specified probability against *any* statement being incorrect that motivates the term ***simultaneous confidence intervals***

- Let $\boldsymbol{X}$ have an $N_p(\boldsymbol{\mu}, \Sigma)$ distribution and form the linear combination

$$Z = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p = \mathbf{a}'\boldsymbol{X}$$

Hence

$$\mu_Z = \mathrm{E}(Z) = \mathbf{a}'\boldsymbol{\mu} \quad \text{and} \quad \sigma_Z^2 = \mathrm{Var}(Z) = \mathbf{a}'\Sigma\mathbf{a}$$

- Moreover, by Result 3.2, Z has an $N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\Sigma\mathbf{a})$ distribution. If a random sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ from the $N_p(\boldsymbol{\mu}, \Sigma)$ and $Z_j = \mathbf{a}'\boldsymbol{X}_j, j = 1, 2, \ldots, n$. Then the sample mean and variance of the observed values $z_1, z_2, \ldots, z_n$ are

$$\bar{z} = \mathbf{a}'\bar{\mathbf{x}} \quad \text{and} \quad s_z^2 = \mathbf{a}'\mathbf{S}\mathbf{a}$$

where $\hat{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and covariance matrix of the $\mathbf{x}_j$'s respectively.

- For $\mathbf{a}$ fixed and σ_Z^2 unknown, a $100(1-\alpha)\%$ confidence interval for $\mu_Z = \mathbf{a}'\boldsymbol{\mu}$ is based on student's t-ratio

$$t = \frac{\bar{z} - \mu_Z}{s_Z/\sqrt{n}} = \frac{\sqrt{n}(\mathbf{a}'\bar{\mathbf{x}} - \mathbf{a}'\boldsymbol{\mu})}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}$$

- and leads to the statement

$$\bar{z} - t_{n-1}(\alpha/2)\frac{s_z}{\sqrt{n}} \leq \mu_Z \leq \bar{z} + t_{n-1}(\alpha/2)\frac{s_z}{\sqrt{n}}$$

or

$$\mathbf{a}'\bar{\mathbf{x}} - t_{n-1}(\alpha/2)\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}} \leq \mathbf{a}'\boldsymbol{\mu} \leq \mathbf{a}'\bar{\mathbf{x}} + t_{n-1}(\alpha/2)\frac{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}}{\sqrt{n}}$$

where $t_{n-1}(\alpha/2)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n-1$ d.f.

- Clearly, we could make serval confidence statements about the components of $\boldsymbol{\mu}$, each with associated confidence coefficient $1-\alpha$, by choosing different coefficient vector $\mathbf{a}$. However, the confidence associated with all of the statements taken together is not $1-\alpha$.

- Intuitively, it would be desirable to associate a "collective" confidence coefficient of $1-\alpha$ with the confidence intervals that can be generated by all choices of $\mathbf{a}$. However, a price must be paid for the convenience of a large simultaneous confidence coefficient: intervals that are wider than the interval for a specific choice of $\mathbf{a}$.

- Given a data set $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and a particular $\mathbf{a}$, the confidence interval is that set of $\mathbf{a}'\boldsymbol{\mu}$ values for which

$$|t| = \left| \frac{\sqrt{n}(\mathbf{a}'\bar{\mathbf{x}} - \mathbf{a}'\boldsymbol{\mu})}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}}} \right| \leq t_{n-1}(\alpha/2)$$

- A simultaneous confidence region is given by the set $\mathbf{a}'\boldsymbol{\mu}$ values such that t^2 is relatively small for all choice of $\mathbf{a}$. It seems reasonable to expect that the constant $t^2(\alpha/2)$ will be replaced by a large value c^2, when statements are developed for many choices of $\mathbf{a}$.

- Considering the values of $\mathbf{a}$ for which $t^2 \leq c^2$, we are naturally led to the determination of

$$\max_{\mathbf{a}} t^2 = \max_{\mathbf{a}} \frac{n(\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu}))^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} = n(\bar{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = T^2$$

with the maximum occurring for $\mathbf{a}$ proportional to $\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$.

Result 4.3. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from an $N_p(\boldsymbol{\mu}, \Sigma)$ population with Σ positive definite. Then simultaneously for all $\mathbf{a}$, the interval

$$\left(\mathbf{a}'\bar{\boldsymbol{X}} - \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'\mathbf{S}\mathbf{a}}, \quad \mathbf{a}'\bar{\boldsymbol{X}} + \sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)\mathbf{a}'\mathbf{S}\mathbf{a}}\right)$$

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

Example 4.4 (Simultaneous confidence intervals as shadows of the confidence ellipsoid In Example 4.3, we obtain the 95% confidence ellipse for the means of the four roots of the door-closed and door-open microwave radiation measurements. Obtain the 95% simultaneous T^2 intervals for the two component means.

Example 4.5 (Constructing simultaneous confidence intervals and ellipse) The scores obtained by $n = 87$ college students on the College Level Examination Program (CLEP) subtest X_1, and the College Qualification Test (CQT) subtests X_2 and X_3 are given in Table 5.3 for X_1 =social science and history, X_2 =verbal, and X_3 =science. Construct simultaneous confidence intervals for μ_1, μ_2 and μ_3.

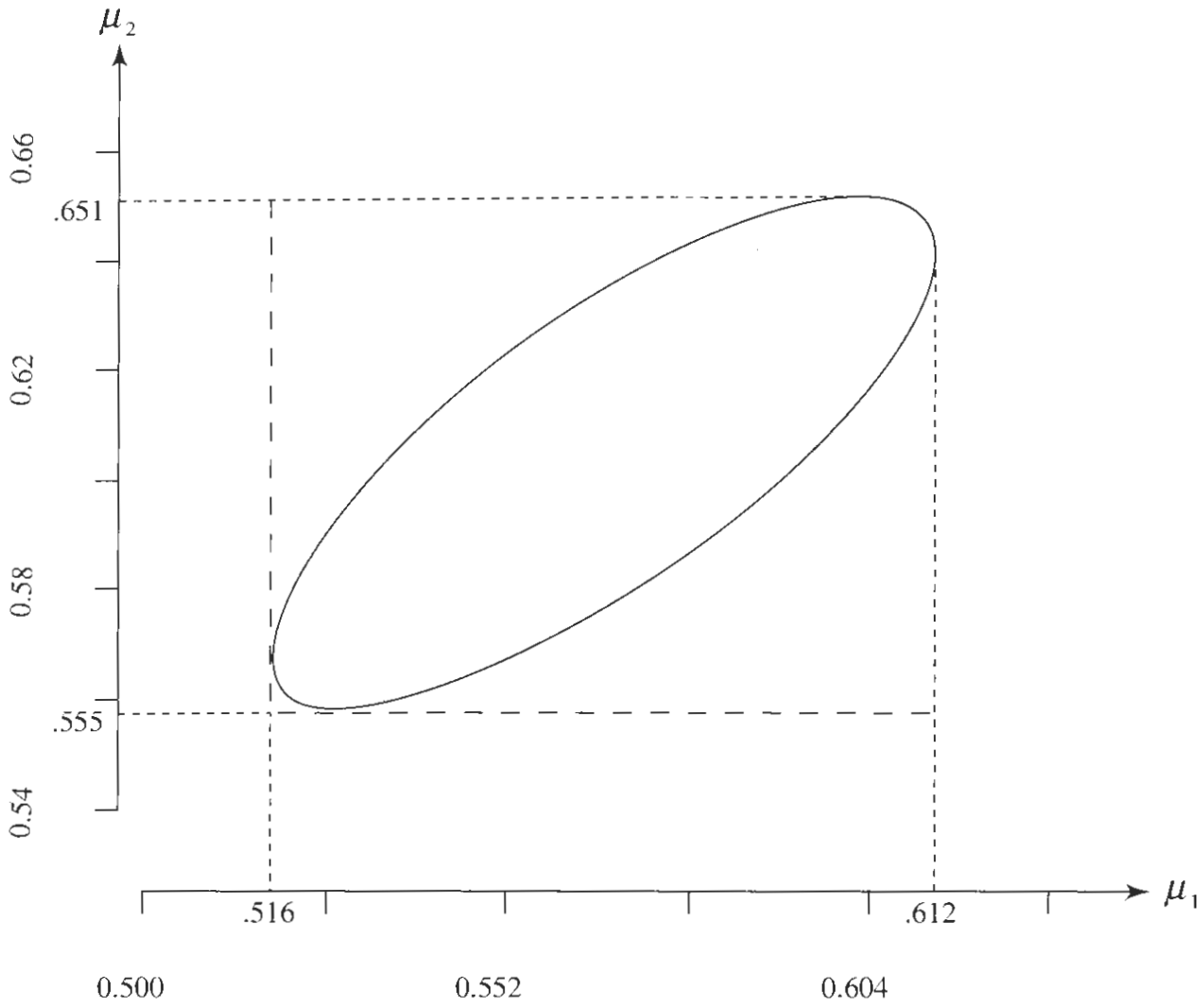

Figure 5.2 Simultaneous T^2-intervals for the component means as shadows of the confidence ellipse on the axes—microwave radiation data.

Table 5.2 College Test Data

Individual	X_1 (Social science and history)	X_2 (Verbal)	X_3 (Science)	Individual	X_1 (Social science and history)	X_2 (Verbal)	X_3 (Science)
1	468	41	26	45	494	41	24
2	428	39	26	46	541	47	25
3	514	53	21	47	362	36	17
4	547	67	33	48	408	28	17
5	614	61	27	49	594	68	23
6	501	67	29	50	501	25	26
7	421	46	22	51	687	75	33
8	527	50	23	52	633	52	31
9	527	55	19	53	647	67	29
10	620	72	32	54	647	65	34
11	587	63	31	55	614	59	25
12	541	59	19	56	633	65	28
13	561	53	26	57	448	55	24
14	468	62	20	58	408	51	19
15	614	65	28	59	441	35	22
16	527	48	21	60	435	60	20
17	507	32	27	61	501	54	21
18	580	64	21	62	507	42	24
19	507	59	21	63	620	71	36
20	521	54	23	64	415	52	20
21	574	52	25	65	554	69	30
22	587	64	31	66	348	28	18
23	488	51	27	67	468	49	25
24	488	62	18	68	507	54	26
25	587	56	26	69	527	47	31
26	421	38	16	70	527	47	26
27	481	52	26	71	435	50	28
28	428	40	19	72	660	70	25
29	640	65	25	73	733	73	33
30	574	61	28	74	507	45	28
31	547	64	27	75	527	62	29
32	580	64	28	76	428	37	19
33	494	53	26	77	481	48	23
34	554	51	21	78	507	61	19
35	647	58	23	79	527	66	23
36	507	65	23	80	488	41	28
37	454	52	28	81	607	69	28
38	427	57	21	82	561	59	34
39	521	66	26	83	614	70	23
40	468	57	14	84	527	49	30
41	587	55	30	85	474	41	16
42	507	61	31	86	441	47	26
43	574	54	31	87	607	67	32
44	507	53	23				

Source: Data courtesy of Richard W. Johnson.

4.4 Large Sample Inference about a Population Mean Vector

Result 4.4. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ. When $n - p$ is large, the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ is rejected in favor of $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, at a level of significance approximately α, if the observed

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \chi_p^2(\alpha)$$

Here $\chi_p^2(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with $p.d.f$.

Result 4.5. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample from a population with mean $\boldsymbol{\mu}$ and positive definite covariance matrix Σ. If $n - p$ is large,

$$\mathbf{a}'\bar{\boldsymbol{X}} \pm \sqrt{\chi_p^2(\alpha)}\sqrt{\frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{n}}$$

will contain $\mathbf{a}'\boldsymbol{\mu}$, for every $\mathbf{a}$, with probability approximately $1 - \alpha$.

Example 4.6 (Constructing large sample simultaneous confidence intervals) A music educator tested thousands of Finnish students on their native musical ability in order to set national norms in Finland. Summary statistics for part of the data set are given in Table 5.5. These statistics are based on a sample of $n = 96$ Finnish 12th graders. Construct 90% simultaneous confidence intervals for individual mean components $\mu_i, i = 1, 2, \ldots, 7$.

Table 5.5 Musical Aptitude Profile Means and Standard Deviations for 96 12th-Grade Finnish Students Participating in a Standardization Program

Variable	Raw score	
	Mean ($\bar{x}_i$)	Standard deviation ($\sqrt{s_{ii}}$)
X_1 = melody	28.1	5.76
X_2 = harmony	26.6	5.85
X_3 = tempo	35.4	3.82
X_4 = meter	34.2	5.12
X_5 = phrasing	23.6	3.76
X_6 = balance	22.0	3.93
X_7 = style	22.7	4.03

Source: Data courtesy of V. Sell.

4.5 Paired Comparisons

Paired Comparisons

In the single response (univariate) case, let X_{j1} denote the response to treatment 1, and let X_{j2} denote the response to treatment 2 for the jth trial. That is, (X_{j1}, X_{j2}) are measurements recorded on the jth unit or jth pair of like units. By design, the n differences

$$D_j = X_{j1} - X_{j2}, j = 1, 2, \ldots, n$$

Should reflect only the differences D_j represent independent observations from an $N(\delta, \sigma_d^2)$ distribution. the variable

$$t = \frac{\bar{D} - \delta}{s_d/\sqrt{n}}$$

where $\bar{D} = \frac{1}{n}\sum_{j=1}^{n} D_j$ and $s_d^2 = \frac{1}{n-1}\sum_{j=1}^{n}(D_j - \bar{D})^2$ has a t-distribution with $n-1$ d.f.

- An α-level test of

$$H_0 : \delta = 0 \quad \text{vs} \quad H_1 : \delta \neq 0$$

may be conducted by comparing $|t|$ with $t_{n-1}(\alpha/2)$-the upper $100(\alpha/2)$th percentile of a t-distribution with n-1 d.f.

- A $100(1-\alpha)\%$ confidence interval for the mean difference $\delta = \mathrm{E}(X_{j1} - X_{j2})$ is provided the statement

$$\bar{D} - t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}} \leq \delta \leq \bar{D} + t_{n-1}(\alpha/2)\frac{s_d}{\sqrt{n}}$$

- Multivariate extension of the paired-comparison procedure to distinguish between p response,two treatments, and n experimental units. The p paired-difference random variables become

$$\begin{aligned} D_{j1} &= X_{1j1} - X_{2j1} \\ D_{j2} &= X_{1j2} - X_{2j2} \\ &\vdots \qquad\quad \vdots \\ D_{jp} &= X_{1jp} - X_{2jp} \end{aligned}$$

Let $\boldsymbol{D}_j' = [D_{j1}, D_{j2}, \ldots, D_{jp}]$, and assume for $j = 1, 2, \ldots, n$ that

$$\mathrm{E}(\boldsymbol{D}_j) = \boldsymbol{\delta} = [\delta_1, \delta_2, \cdots, \delta_p]' \quad \text{and} \quad \mathrm{Cov}(\boldsymbol{D}_j) = \Sigma_d$$

If, in addition, $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n$ are independent $N_p(\boldsymbol{\delta}, \Sigma_d)$ random vectors, inference about the vector of mean differences $\boldsymbol{\delta}$ can based upon a T^2 statistics

$$T^2 = n(\bar{\boldsymbol{D}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\boldsymbol{D}} - \boldsymbol{\delta})$$

where $\bar{\boldsymbol{D}} = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{D}_j$ and $\mathbf{S}_d = \frac{1}{n-1} \sum_{j=1}^{n} (\boldsymbol{D}_j - \bar{\boldsymbol{D}})(\boldsymbol{D}_j - \bar{\boldsymbol{D}})'$.

Result 4.6 Let the differences $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_n$ be a random sample from $N_p(\boldsymbol{\delta}, \Sigma_d)$ population. Then

$$T^2 = n(\bar{\boldsymbol{D}} - \boldsymbol{\delta})' \mathbf{S}_d^{-1} (\bar{\boldsymbol{D}} - \boldsymbol{\delta})$$

is distributed as an $[(n-p)p/(n-p)F_{p,n-p}]$ random variable, whatever the true $\boldsymbol{\delta}$ and Σ_d.

Given the observed differences $\boldsymbol{D}_j' = [D_{j1}, D_{j2}, \ldots, D_{jp}], j = 1, 2, \ldots, n,$

- an α-level test of

$$H_0 : \boldsymbol{\delta} = 0 \quad \text{versus} \quad H_1 : \boldsymbol{\delta} = 0$$

for an $N_p(\boldsymbol{\delta}, \Sigma_d)$ population rejects H_0 if the observed

$$T^2 = n\bar{\boldsymbol{D}}'\mathbf{S}_d^{-1}\bar{\boldsymbol{D}} > \frac{(n-p)p}{n-p}F_{p,n-p}(\alpha)$$

where $F_{p,n-p}(\alpha)$ is the upper $(100\alpha)\%$th percentile of an F-distribution with p and $n-p$ d.f.

- A $100(1-\alpha)\%$ confidence region for $\boldsymbol{\delta}$ consists of all $\boldsymbol{\delta}$ such that

$$(\bar{\boldsymbol{D}} - \boldsymbol{\delta})'\mathbf{S}_d^{-1}(\bar{\boldsymbol{D}} - \boldsymbol{\delta}) > \frac{(n-p)p}{n(n-p)}F_{p,n-p}$$

Also $100(1-\alpha)\%$ simultaneous confidence intervals for the individual mean differences δ_i are given by

$$\delta_i : \bar{\boldsymbol{D}}_i \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\sqrt{\frac{s_{D_i}^2}{n}}$$

where $\bar{\boldsymbol{D}}_i$ is the ith element of $\bar{\boldsymbol{D}}$ and $s_{D_i}^2$ is the ith diagonal element of $\mathbf{S}_d$.[22]

Example 4.7 (Checking for a mean difference with paired observations) Muicipal wasterwater treatment plants are required by law to monitor their discharges into rivers and streams on a regular basis. Concern about the reliability of data from one of these self-monitoring programs led to a study in which samples of effluent were divided and sent to two laboratories for testing. One-half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solid (SS) were obtained, for $n = 11$ sample splits, from the two laboratories . The data are displayed in Table 6.1.

Do the two laboratories's chemical analyses agree ? If differences exist, what is their nature?

Table 6.1 Effluent Data

Sample j	Commercial lab x_{1j1} (BOD)	Commercial lab x_{1j2} (SS)	State lab of hygiene x_{2j1} (BOD)	State lab of hygiene x_{2j2} (SS)
1	6	27	25	15
2	6	23	28	13
3	18	64	36	22
4	8	44	35	29
5	11	30	15	31
6	34	75	44	64
7	28	26	42	30
8	71	124	54	64
9	43	54	34	56
10	33	30	29	20
11	20	14	39	21

Source: Data courtesy of S. Weber.

4.6 Comparing Mean Vectors from Two Populations

Sample	Summary statistics	
(Population 1)		
$\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1n_1}$	$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{x}_{1j}$	$\mathbf{S}_1 = \frac{1}{n_1-1}\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)'$
(Population 2)		
$\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2n_1 2}$	$\bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \mathbf{x}_{2j}$	$\mathbf{S}_2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'$

Assumptions Concerning the Structure of the Data

1. The sample $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from a p-variate population with mean $\boldsymbol{\mu}_1$ and covariance matrix Σ_1.

2. The sample $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$ is a random sample of size n_1 from a p-variate population with mean $\boldsymbol{\mu}_2$ and covariance matrix Σ_2.

3. Also, $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ are independent of $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$.

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0 \quad \textbf{vs} \quad H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq 0$$

Further Assumption When n_1 and n_2 Are Small

1. Both populations are multivariate normal.
2. Also, $\Sigma_1 = \Sigma_2$ (same covariance matrix)

Set the estimate of Σ as

$$\begin{aligned}
\mathbf{S}_{pooled} &= \frac{\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)(\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)' + \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)(\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)'}{n_1 + n_2 - 2} \\
&= \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2
\end{aligned}$$

Then

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

Hence

$$\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma$$

is an estimator of $\text{Cov}(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$.

Result 4.7 If $\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1n_1}$ is a random sample of size n_1 from $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $\boldsymbol{X}_{21}, \boldsymbol{X}_{22}, \ldots, \boldsymbol{X}_{2n_2}$ is an independent random sample size n_2 from $N_p(\boldsymbol{\mu}_2, \Sigma)$, then

$$T^2 = [\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} [\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]$$

is distributed as

$$\frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$$

Consequently,

$$P(T^2 \leq c^2) = 1 - \alpha$$

where

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

Example 4.8 (Constructing a confidence region for the difference of two mean vectors) Fifty bars of soap are manufactured in each of two ways. Two characteristics, X_1 =lather and X_2 =mildness, are measured. The summary statistics for bars produced by method 1 and 2 are

$$\mathbf{x}_1 = [8.3\ 4.1]', \mathbf{x}_2 = [10.2\ 3.9]'$$

$$\mathbf{S}_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Obtain a 95% confidence region for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

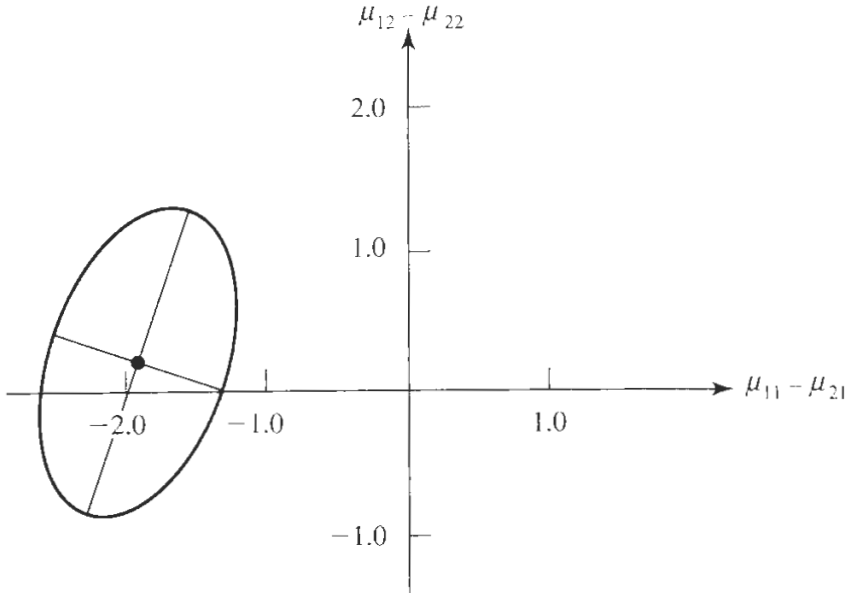

Figure 6.1 95% confidence ellipse for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Simultaneous Confidence Intervals

Result 4.8 Let $c^2 = [(n_1 + n_2 - 2)p/(n_1 + n_2 - p - 1)]F_{p,n_1+n_2-p-1}(\alpha)$. With probability $1 - \alpha$.

$$\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm c\sqrt{\mathbf{a}'\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{pooled}\mathbf{a}}$$

will cover $\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for all $\mathbf{a}$. In particular $\mu_{1i} - \mu_{2i}$ will be covered by

$$(\bar{\boldsymbol{X}}_{1i} - \bar{\boldsymbol{X}}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s_{ii,pooled}} \quad \text{for } i = 1, 2, \ldots, p$$

The Two-Sample Situation When $\Sigma_1 \neq \Sigma_2$

Result 4.9 Let the sample sizes be such that $n_1 - p$ and $n_2 - p$ are large. Then, an approximate $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is given by all $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ satisfying

$$T^2 = [\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)]' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{S}_{pooled} \right]^{-1} [\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)] \leq \chi^2_p(\alpha)$$

where $\chi^2_p(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with p d.f. Also $100(1-\alpha)\%$ simultaneous confidence interval for all linear combinations $\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ are provided by

$$\mathbf{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{belongs to} \quad \mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \pm \sqrt{\chi^2_p(\alpha)} \sqrt{\mathbf{a}' \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right) \mathbf{a}}.$$

4.7 Testing for Equality of Covariance Matrics

With g populations, the null hypothesis is

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

where Σ_l is the covariance matrix for the lth population, $l = 1, 2, \ldots, g$, and Σ is the presumed common covariance matrix. The alternative hypothesis is that at least two of the covariance matrices are not equal.

- Assuming multivariate normal populations, a likelihood ratio statistic for testing above is given by

$$\Gamma = \prod_l \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{pooled}|} \right)^{(n_l-1)/2}$$

Here n_l is the sample size for the lth group, $\mathbf{S}_l$ is the lth group sample covariance matrix and $\mathbf{S}_{pool}$ is the pooled sample covariance matrix given by

$$\mathbf{S}_{pool} = \frac{1}{\sum_l (n_l - 1)} \{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 + \cdots + (n_g - 1)\mathbf{S}_g\}$$

- Box's test is based on his χ^2 approximation to the sampling distribution of $-2 \ln \Gamma$. Setting $-2 \ln \Gamma = M$(Box's M statistics) gives

$$M = \left[\sum_l (n_l - 1) \right] \ln |\mathbf{S}_{pooled}| - \sum_l [(n_l - 1) \ln |\mathbf{S}_l|.$$

Box's Test for Equailty of Covariance Matrices

Set

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

where p is the number of variables and g is the number of groups. Then

$$C = (1-u)M = (1-u)\left\{\left[\sum_l (n_l - 1)\right] \ln |\mathbf{S}_{pooled}| - \sum_l [(n_l - 1) \ln |\mathbf{S}_l]\right\}$$

has an approximate χ^2 distribution with

$$\nu = g\frac{1}{2}p(p+1) - \frac{1}{2}p(p+1) = \frac{1}{2}p(p+1)(g-1)$$

degrees of freedom. At significance level α, reject H_0 if $C > \chi^2_{p(p+1)(g-1)/2}(\alpha)$.

Example 4.9 (Testing equality of covariance matrices—nursing homes) The Wisconsin Department of Health and Social Services reimburse nursing homes in the state for the services provided. The department develops a set of formulas for the rates for each facility, based on factors such as level of care, mean wage rate, and average wage rate in the state.

Nursing homes can be classified on the basis of ownership (private party, nonprofit organization, and government) and certification (skilled nursing facility, intermediate care facility, or combination of the two).

One purpose of a recent study was to investigate the effects of ownership or certification (or both) on cost.s. Four costs, computed on a per-patient-day basis and measured in hours per patient day, were selected for analysis $X_1 =$ cost of nursing labor, $X_2 =$cost dietary labor, $X_3 =$cost of plant operation and maintenance labor, and $X_4 =$cost of housekeeping and laundry labor. A total of $n = 516$ observations on each of the $p = 4$ cost variables were initially separated according to ownership. Summary statistics for each of the $g = 3$ groups are given in the following table.

Group	Number of observations	Sample mean vectors
$l = 1$ (private)	$n_1 = 271$	$\bar{\mathbf{x}}_1 = [2.066 \ .480 \ .082 \ .360]'$
$l = 2$ (nonprofit)	$n_2 = 138$	$\bar{\mathbf{x}}_2 = [2.167 \ .596 \ .124 \ .418]'$
$l = 3$(government)	$n_3 = 107$	$\bar{\mathbf{x}}_3 = [2.273 \ .521 \ .125 \ .283]'$

Sample covariance matrices

$$\mathbf{S}_1 = \begin{bmatrix} .291 & & & \\ -.001 & .011 & & \\ .002 & .000 & .001 & \\ .010 & .003 & .000 & .010 \end{bmatrix}; \quad \mathbf{S}_2 = \begin{bmatrix} .561 & & & \\ -.011 & .025 & & \\ .001 & .004 & .005 & \\ .037 & .007 & .002 & .019 \end{bmatrix};$$

$$\mathbf{S}_3 = \begin{bmatrix} .261 & & & \\ -.030 & .017 & & \\ .003 & -.000 & .004 & \\ .018 & .006 & .001 & .013 \end{bmatrix};$$

Assuming multivariate normal data, test hypothesis $H_0 : \Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma$.