Applied Multivariate and Longitudinal Data Analysis

Chapter 2: Inference about the mean vector(s)

Ana-Maria Staicu
SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

In this chapter we will discuss inference about one population mean vector as well as compare two or more population mean vectors. We discuss one way MANOVA and two-way MANOVA. We will consider typical component-by-component comparison as well as profile analysis. Recall by *inference* we mean reaching conclusions concerning a population parameter using information from data.
Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ be two population mean vectors. In this chapter we will study

- how to carry out inference about $\boldsymbol{\mu}_1$ (hypothesis testing and confidence intervals)
- how to formally assess a hypothesis of the form $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$
- how to formally assess a hypothesis of the form
  H_0 : *the change in the components of* $\boldsymbol{\mu}_1$ **is the same** *as the change in the components of* $\boldsymbol{\mu}_2$
- how to construct confidence intervals for the mean difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

in a variety of scenarios.

Motivating example: Fisher's or Anderson's Iris data.

Source: Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179-188.

Data set: Available in R as 'iris'

This data set consists of the measurements of the variables sepal length and width, and petal length and width (in centimeters), respectively, for 50 flowers from each of 3 species (setosa, versicolor and virginica) of iris. The first few rows of the data matrix is shown below.

Id	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Question 1: Estimate the mean of (SL, SW, PL, PW) for setosa flowers.

Question 2: Provide confidence intervals.

Question 3: Compare the mean of these characteristics across species.

1 Inference for a single population: Confidence intervals. Hypothesis testing.

Review of the univariate case: Let $X_1, X_2, \ldots, X_n$ be a sample from a normal distribution with mean μ and unknown variance. Let $\alpha \in (0, 1)$.

Constructing confidence intervals / bands for μ as well as hypothesis testing procedures mainly rely on the result:

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is the Student t distribution with $n - 1$ degrees of freedom.

Then we can construct confidence intervals for μ by using an estimate for the standard error and appropriate critical value. Specifically let $t_{n-1}(\alpha/2)$ be the upper tail probability corresponding to the t_{n-1} distribution defined as $P(t_{n-1} > t_{n-1}(\alpha/2)) = \alpha/2$. Then the $100(1-\alpha)\%$ confidence interval (CI) for μ is

$$\left(\bar{X} - t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}},\ \bar{X} + t_{n-1}(\alpha/2)\frac{S}{\sqrt{n}}\right).$$

The $100(1-\alpha)\%$ CI for μ can be written as $|\sqrt{n}(\bar{X} - \mu)/S| \leq t_{n-1}(\alpha/2)$ or equivalently

$$n(\bar{X} - \mu)^2/S^2 \leq F_{1,n-1}(\alpha)$$

since $t_{n-1}(\alpha/2)^2 = F_{1,n-1}(\alpha)$ where $F_{1,n-1}(\alpha)$ is the upper tail probability corresponding to the $F_{1,n-1}$ distribution defined as $P(F_{1,n-1} > F_{1,n-1}(\alpha)) = \alpha$.

This remark allows us to extend the confidence interval to confidence region, which is more appropriate when the population parameter of interest (e.g. mean vector) has a higher dimension than one.

Definition: When the parameter, denoted generically by $\boldsymbol{\theta}$, has dimension $p \geq 2$, the $100(1-\alpha)\%$ confidence region based on same sample $\mathbf{X}$, denoted by $R(\mathbf{X})$, is defined as

$$P\big(R(\mathbf{X}) \text{ will cover the true } \boldsymbol{\theta}\big) = 1 - \alpha,$$

where the probability is calculated using the true parameter value θ.

CASE 1: Normal parent distribution.

Assume $\boldsymbol{\mu}$ is the mean p-dimensional vector parameter of interest. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a *random sample from a p-variate normal parent distribution with mean $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$.*

Let $\bar{\mathbf{X}} = \frac{1}{n}(\mathbf{X}_1 + \ldots + \mathbf{X}_n)$ be the sample mean and $S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$ be the sample covariance matrix. Most of the inferential procedures that we study use the test statistic (called *Hotelling's T^2*)

$$T^2 := n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$$

and rely on the result that

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

It follows that $P\{T^2 \leq \frac{p(n-1)}{n-p} F_{p,n-p}(\alpha)\} = 1 - \alpha$ for critical value $F_{p,n-p}(\alpha)$.

Confidence region for $\boldsymbol{\mu}$.

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be the observed data. The $100(1-\alpha)\%$ confidence region for $\boldsymbol{\mu}$ is the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\intercal} \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha);$$

here $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean and sample covariance for the particular observed data.

Remarks:

- The confidence region is exact for any value of n
- For large n the limiting distribution $\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \approx \chi^2_p$, i.e. the two variables on each side of $\approx$ have approximately the same distribution

Example: In class demonstration using iris data.

While the (ellipsoidal) region $n(\bar{\mathbf{x}}-\boldsymbol{\mu})^\intercal \mathbf{S}^{-1}(\bar{\mathbf{x}}-\boldsymbol{\mu}) \leq c$ for some $c>0$ assesses joint knowledge concerning plausible values of $\boldsymbol{\mu}$, it is common to summarize the conclusions for the individual components of $\boldsymbol{\mu}$ separately. In doing so, we adopt the principle that the individual intervals hold simultaneously with a specified high probability.

Definition: Suppose $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_p)^T$. The intervals $I_1,\ldots,I_p$ are simultaneous $100(1-\alpha)\%$ confidence intervals for $\mu_1,\ldots,\mu_p$ if

$$P(I_k \text{ will contain true } \mu_k \textit{ simultaneously} \text{ for every } k=1,\ldots,p) = 1-\alpha.$$

Key ideas: Consider linear combinations $\mathbf{a}^\intercal\boldsymbol{\mu}$; the $100(1-\alpha)\%$ CI for $\mathbf{a}^\intercal\boldsymbol{\mu}$ is

$$\left|\frac{\sqrt{n}(\mathbf{a}^\intercal\bar{\mathbf{X}}-\mathbf{a}^\intercal\boldsymbol{\mu})}{\sqrt{\mathbf{a}^\intercal\mathbf{S}\mathbf{a}}}\right| \leq t_{n-1}(\alpha/2).$$

By conveniently choosing $\mathbf{a}$ we can obtain individual confidence intervals for the components of $\boldsymbol{\mu}$. However these confidence intervals are not simultaneous; the confidence associated with all the statements taken together is smaller than $(1-\alpha)$.

In order to obtain simultaneous confidence intervals we need to evaluate

$$\max_{\mathbf{a}}\left|\frac{\sqrt{n}(\mathbf{a}^\intercal\bar{\mathbf{X}}-\mathbf{a}^\intercal\boldsymbol{\mu})}{\sqrt{\mathbf{a}^\intercal\mathbf{S}\mathbf{a}}}\right|$$

which is an equivalent problem to evaluating

$$\max_{\mathbf{a}}\left|\frac{n(\mathbf{a}^\intercal\bar{\mathbf{X}}-\mathbf{a}^\intercal\boldsymbol{\mu})^2}{\mathbf{a}^\intercal\mathbf{S}\mathbf{a}}\right| = n(\bar{\mathbf{X}}-\boldsymbol{\mu})^\intercal\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu});$$

the right hand side terms is T^2, and its sampling distribution is $\frac{(n-1)p}{(n-p)}F_{p,n-p}$.

Thus, simultaneously *for all* p-dimensional vectors $\mathbf{a}$, the interval

$$\mathbf{a}^\intercal\bar{\mathbf{X}} \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\ \sqrt{\mathbf{a}^\intercal\mathbf{S}\mathbf{a}/n}$$

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

Simultaneous confidence intervals for the components of $\boldsymbol{\mu}$

Let $\bar{\mathbf{X}}=(\bar{X}_1,\ldots,\bar{X}_p)$. The simultaneous $100(1-\alpha)\%$ confidence intervals for μ_k, $k=1,\ldots,p$ are

$$\bar{X}_k \pm \sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\ \sqrt{S_{kk}/n}, \quad k=1,\ldots,p,$$

where S_{kk} is the kth element of the diagonal of the sample covariance $\mathbf{S}$.

We can compare these simultaneous confidence intervals to the individual intervals (*one-at-a time*) for $\mu_k, k = 1, \ldots, p$. Assume we are concerned with a smaller number of parameters; these could be individual components of the population mean vector, or more generally linear combinations of the population mean vector $\mathbf{a}_j^T \boldsymbol{\mu}$ for $j = 1, \ldots, m$ for some m. In these situations, it may be be possible to do better than the simultaneous confidence intervals. Specifically it may be possible to find a critical value that yields shorter intervals while still preserving the correct level of confidence; the method is commonly referred to as *Bonferroni correction*.

Key idea: Key ideas: Let C_j denote a confidence statement about the value of $\mathbf{a}_j^T \boldsymbol{\mu}$ with level of confidence $(1 - \alpha')$. Then $P(C_j \; true) = 1 - \alpha'$. It follows that:

$$P(all \; C_j \; true) = 1 - P(at \; least \; one \; C_j \; false) \geq \ldots = 1 - m\alpha'.$$

One can preserve an overall level of confidence of $1 - \alpha$ by choosing the individual level of confidence $1 - \alpha'$ such that $\alpha' = \alpha/m$.

Bonferroni-corrected confidence intervals for the components of $\boldsymbol{\mu}$.

The simultaneous $100(1 - \alpha)\%$ confidence intervals for μ_k, $k = 1, \ldots, p$ are

$$\bar{X}_k \pm t_{n-1}\left(\frac{\alpha}{2p}\right) \sqrt{S_{kk}/n}, \quad k = 1, \ldots, p$$

where S_{kk} is the kth element of the diagonal of the sample covariance $\mathbf{S}$.

Hypothesis testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ (using Hotelling's T^2).

We want to test the above null hypothesis versus the alternative that $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ using significance level α. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be the observed data. For this we will use the test statistic $T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\intercal \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$. Recall that when the null hypothesis is true ($\boldsymbol{\mu} = \boldsymbol{\mu}_0$) then

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

So we would reject H_0 at level α if

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^\intercal \mathbf{s}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha).$$

There is a general principle used to develop powerful testing procedure, and this is called *likelihood ratio (LR) method*. Intuitively, the test derived by this principle is equal to the ratio between the maximum likelihood under the null hypothesis ($\boldsymbol{\mu} = \boldsymbol{\mu}_0$) and the maximum likelihood under the the alternative hypothesis; the LR testing procedures have several optimal properties for large samples. A LR test rejects H_0 in favor of the alternative if the LR is smaller than some suitably chosen constant.

Hypothesis testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ (using Wilks's lambda).

The *likelihood ratio test* for testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ is

$$\Lambda := \left(\frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|} \right)^{n/2}$$

where $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ and $\widehat{\boldsymbol{\Sigma}}_0 := \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)^T$ are the estimated covariance under the alternative H_1 and null H_0 hypotheses respectively. Note that $\widehat{\boldsymbol{\Sigma}}$ was previously denoted by S; the new notation is required by the set up. Also $|A|$ denotes the determinant of the square matrix A.

The equivalent statistic $\Lambda^{2/n} = |\widehat{\boldsymbol{\Sigma}}|/|\widehat{\boldsymbol{\Sigma}}_0|$ is also known as the *Wilks's lambda* statistic.

It turns out that $\Lambda^{2/n} = \{1 + T^2/(n-1)\}^{-1}$, where T^2 is the Hotelling's statistic.
So we would reject H_0 at level α if

$$\Lambda < c_\alpha.$$

where c_α is the lower $100\alpha\%$ percentile of the distribution of Λ.

CASE 2: Unknown parent distribution. Large sample size n.

Assume $\boldsymbol{\mu}$ is the mean p-dimensional vector parameter of interest. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a *sample from a p-multivariate parent distribution with mean $\boldsymbol{\mu}$ and some unknown covariance matrix $\boldsymbol{\Sigma}$*. When the sample size n is large, the inference about $\boldsymbol{\mu}$ is done as above, with the difference that the distribution of the test statistics is approximate and thus the confidence intervals are *approximately* accurate.

When the sample size is large, the limiting distribution, $\frac{(n-1)p}{n-p}F_{p,n-p}$, converges to chi-square distribution χ^2_p. Thus

- the hypothesis testing is based on the χ^2_p asymptotic null distribution of the T^2 test;
- the confidence intervals/region use the critical value from χ^2_p.

In the lack of information about the latent parent distribution (that is multivariate normal) one needs to check the multivariate normality assumption.

Assessing normality. Most of the techniques that we study in this course assume that the parent distribution is multivariate normal or that the sample size sufficiently large (in which case the normality assumption is less crucial). However the quality of the inferences relies on how close the parent distribution is to the multivariate normal. Thus it is important to validate the normality assumption.
Some diagnostic plots:

- Assess whether the individual components (characteristics) come from univariate normal distribution. For this use : quantile-quantile plot (`qqplot()`), goodness-of-fit test, Shapiro-Wilk test
- Assess whether pairs of components (pairs of characteristics) have ellipsoidal appearance. Contour and perspective plots for assessing bivariate normality
- Chisquare Q-Q plot to check the multivariate normality.

  The variable $d^2 = (\mathbf{X} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu})$ has a chi-square distribution with p degrees of freedom, and for large samples the observed Mahalanobis distances have an approximate chi-square distribution. This result can be used to evaluate (subjectively) whether a data point may be an outlier and whether observed data may have a multivariate normal distribution.

  A qq-plot can be used to plot the Mahalanobis distance for the sample vs the theoretical quantile of a chi-square distribution with p degrees of freedom. This should resemble a straight-line for data from a multivariate normal distribution. Outliers will show up as points on the upper right side of the plot for which the Mahalanobis distance is notably greater than the chi-square quantile value.
- Other testing procedures to check for normality exist. See the R code part.

Exercise: Application to the iris data.

2 Treatment mean comparison, when treatments arise from factors

Dogs anesthetics example (Johnson and Wichern, 2007). Improved anesthetics are often developed by first studying their effects on animals. In one study 19 dogs were initially given the drug pentobarbitol. Each dog was then administered carbon dioxide CO_2 at each of two pressure levels. Next, halothane (H) was added and then administration of CO_2 was repeated. The response, milliseconds between heartbeats was measured for the four treatment combinations. The data are included in the folder called"Data" that is available on the course website.

Treatment	CO_2 pressure	Halothane
1	high	present
2	low	present
3	high	absent
4	low	absent

Id	trt 1	trt 2	trt 3	trt 4
1	426	609	556	600
2	253	236	392	395
3	359	433	349	357
4	432	431	522	600
5	405	426	513	513
6	324	438	507	539
7	310	312	410	456
8	326	326	350	504
9	375	447	547	548
10	286	286	403	422
11	349	382	473	497
12	429	410	488	547
13	348	377	447	514
14	412	473	472	446
15	347	326	455	468
16	434	458	637	524
17	364	367	432	469
18	420	395	508	531
19	397	556	645	625

Denote the mean response under treatment j as μ_j, $j = 1, ..., 4$.

Hypotheses:

1. Null effect: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
2. Main effect of halothane. $H_0 : (\mu_1 + \mu_2) - (\mu_3 + \mu_4) = 0$
3. Main effect of CO_2. $H_0 : (\mu_1 + \mu_3) - (\mu_2 + \mu_4)$
4. Interaction of CO_2 and halothane. $H_0 : (\mu_1 + \mu_4) - (\mu_2 + \mu_3) = 0$
5. In general, *no treatment effect* can be written as $H_0 : \mathbf{C}\boldsymbol{\mu} = 0$ for a $q \times 4$ matrix $\mathbf{C}$

In class: Write down the matrix $\mathbf{C}$ for points 2–4 above.

In general, assume that multiple competing treatments are assigned to the same experimental unit for many units and of interest is to formally assess whether there is significant difference between the two corresponding mean responses. Specifically, for the ith experimental unit denote by X_{i1} the *scalar* response for the first treatment, X_{i2} the *response* for the second treatment, and so on for p competing treatments. Here i indexes the experimental unit, $i = 1, \ldots, n$.

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be the vector of responses for the ith experimental unit. It is assumed that the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is IID from a p-variate normal distribution with some unknown mean vector $\boldsymbol{\mu}$ and unknown covariance matrix. The parameter of interest is $\mathbf{C}\boldsymbol{\mu}$, for some appropriately defined $q \times p$ matrix $\mathbf{C}$.

A $100(1-\alpha)\%$ **confidence region** for $\mathbf{C}\boldsymbol{\mu}$:

$$n(\mathbf{C}\bar{\mathbf{X}} - \mathbf{C}\boldsymbol{\mu})^{\intercal}(\mathbf{C}\mathbf{S}\mathbf{C}^{\intercal})^{-1}(\mathbf{C}\bar{\mathbf{X}} - \mathbf{C}\boldsymbol{\mu}) \leq \frac{(n-1)q}{n-q} F_{q,n-q}(\alpha);$$

here $\bar{\mathbf{X}}$ and $\mathbf{S}$ are the sample mean and covariance of $\mathbf{X}_i$'s.

$100(1-\alpha)\%$ **simultaneous confidence intervals** are

$$\mathbf{c}^{\intercal}\bar{\mathbf{X}} \pm \sqrt{\frac{(n-1)q}{n-q} F_{q,n-q}(\alpha)} \sqrt{\frac{\mathbf{c}^{\intercal}\mathbf{S}\mathbf{c}}{n}}$$

will contain $\mathbf{c}^{\intercal}\boldsymbol{\mu}$ *simultaneously* for all the columns $\mathbf{c}$ of the matrix $\mathbf{C}$ with probability $1-\alpha$.

Testing $H_0 : \mathbf{C}\boldsymbol{\mu} = 0$ is done using the test statistic:

$$T^2 := n(\mathbf{C}\bar{\mathbf{X}})^T(\mathbf{C}\mathbf{S}\mathbf{C}^T)^{-1}\mathbf{C}\bar{\mathbf{X}}.$$

If the null hypothesis is true then $T^2 \sim \frac{(n-1)q}{n-q} F_{q,n-q}$. Thus we reject H_0 at level α if the observed value of T^2 exceeds the critical value $\frac{(n-1)q}{(n-q)} F_{q,n-q}(\alpha)$.

Remark: The above results are presented for the case that the matrix $\mathbf{C}$ has q rows. In the previous data example, the matrix has $p-1$ rows (each row corresponding to one contrast). In that case, the analogous results are as the above with q replaced by $p-1$.

3 Inference for two populations: Confidence intervals. Hypothesis testing.

The ideas of the previous part can be used to develop procedures for comparison of two or more population mean vectors. We will first study such comparisons when the data arise from a paired design (measurements before/after), then when the data arises from multiple treatments, and finally when we have multiple independent populations.

3.1 Paired comparison

Wastewater monitoring example (Johnson and Wichern, 2007). Municipal wastewater treatment plants are required by law to monitor their discharge into rivers and streams on a regular basis. Concerns 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 labs for testing. One half of each sample was sent to the Wisconsin State Laboratory of Hygiene and one half was sent to a private commercial laboratory routinely used in the monitoring program. Measurements of biochemical oxygen demand (BOD) and suspended solids (SS) were obtained for $n = 11$ sample splits from the two laboratories.

Goal: compare commercial lab and state lab.

Commercial BOD	Commercial SS	StateLabHygiene BOD	StateLabHygiene SS
6	27	25	15
6	23	28	13
18	64	36	22
8	44	35	29
11	30	15	31
34	75	44	64
28	26	42	30
71	124	54	64
43	54	34	56
33	30	29	20
20	14	39	21

Assume two competing treatments are assigned to the same experimental unit for many units and of interest is to formally assess whether there is significant difference between the two corresponding mean responses.

Specifically, for the ith experimental unit denote by $\mathbf{X}_{1i}$ the response for the first treatment and $\mathbf{X}_{2i}$ the response for the second treatment, where $i = 1, \ldots, n$. Denote by $\boldsymbol{\mu}_1 = E[\mathbf{X}_{1i}]$ and let $\boldsymbol{\mu}_2 = E[\mathbf{X}_{2i}]$. We want to draw inferences about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = E[\mathbf{X}_{1i} - \mathbf{X}_{2i}]$.

Approach: calculate *pairwise* differences $\mathbf{d}_i = \mathbf{X}_{1i} - \mathbf{X}_{2i}$. Assume that $\mathbf{d}_1, \ldots, \mathbf{d}_n$ is an IID sample from the multivariate normal distribution (or that the sample size is very large) with mean $\boldsymbol{\delta}$ and some unknown covariance Σ. Of interest is to draw inference about $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

Paired test for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

To test the null hypothesis $H_0 : \boldsymbol{\delta} = 0_p$ vs $H_1 : \boldsymbol{\delta} \neq 0_p$, we calculate the T^2 test statistic

$$T^2 := n\bar{\mathbf{d}}^T \mathbf{S}_d^{-1} \bar{\mathbf{d}};$$

here $\bar{\mathbf{d}}$ and $\mathbf{S}_d$ are the sample mean and sample covariance of $\mathbf{d}_i$'s. If the null hypothesis is true then $T^2 \sim \frac{(n-1)p}{n-p} F_{p,n-p}$. Thus we reject H_0 at level α if the observed value of T^2 exceeds the critical value $\frac{(n-1)p}{(n-p)}\ F_{p,n-p}(\alpha)$.

Confidence region for the mean difference, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, is

$$\{\text{all } \boldsymbol{\delta} \in \mathbb{R}^p \text{ such that } n(\bar{\mathbf{d}} - \boldsymbol{\delta})^T \mathbf{S}_d^{-1} (\bar{\mathbf{d}} - \boldsymbol{\delta}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\}$$

Simultaneous confidence intervals for the components of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are:

$$\bar{d}_k \pm \sqrt{\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)}\ \sqrt{s_{d,kk}/n}, \quad k = 1, \ldots, p;$$

here d_k is the kth element of the mean difference vector $\bar{\mathbf{d}}$ and $s_{d,kk}$ is the kth element of the diagonal of $\mathbf{S}_d$.

When the sample size is large, then $\frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)$ is replaced by $\chi^2_p(\alpha)$.

Bonferroni-corrected confidence intervals for the components of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are:

$$\bar{d}_k \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\ \sqrt{s_{d,kk}/n}, \quad k = 1, \ldots, p,$$

Example: in class demonstration using the wastewater data.

3.2 Independent populations comparison

Electricity consumption example (Johnson and Wichern, 2007). A total of $n_1 = 45$ and $n_2 = 55$ homes with and without air conditioning respectively are measured in Wisconsin. Two measurements of electrical usage (in kilowatt hours) were considered. The first is a measure of total on-peak consumption (X_1) during July and the second is a measure of total off-peak consumption (X_2) during July. The resulting summaries are:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 204.4 \\ 556.6 \end{bmatrix} \qquad \mathbf{S}_1 = \begin{bmatrix} 13825.3 & 23823.4 \\ 23823.4 & 73107.4 \end{bmatrix}$$
$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 130.0 \\ 355.0 \end{bmatrix} \qquad \mathbf{S}_1 = \begin{bmatrix} 8632.0 & 19616.7 \\ 19616.7 & 55964.5 \end{bmatrix}$$

Goal: inference on the difference of electricity consumption with and without air conditioning.

Review of univariate two sample t test:

- Setup: $X_{11}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$; $X_{21}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$
- $H_0 : \mu_1 - \mu_2 = \delta_0$
- Assumptions: independent populations; equal variance $\sigma_1^2 = \sigma_2^2$; normal distributions
- Test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \text{ where } S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

  where S_1^2 and S_2^2 are sample variances for the two samples.

- Distribution under H_0: $t \sim t_{n_1+n_2-2}$

The multivariate procedures rely on the direct extension of this result to higher dimensions.

Hotelling's T^2 to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$

- Setup: $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$; $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
- $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}_0$ versus $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \neq \boldsymbol{\delta}_0$
- Assumptions: independent populations; equal covariance $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$; normal distributions
- Test statistic:

$$T^2 = \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}_0\right\}^{\intercal} \left\{\mathbf{S}_p \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{-1} \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}_0\right\}$$

  where $\mathbf{S}_p := \frac{n_1-1}{n_1+n_2-2}\mathbf{S}_1 + \frac{n_2-1}{n_1+n_2-2}\mathbf{S}_2$

- Distribution under H_0:

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

  The confidence region and simultaneous confidence intervals are determined in the same way as before.

What if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$?

- When $n_1 - p$ and $n_2 - p$ are both large, we define the test statistic as

$$T^2 = \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}\right\}^{\intercal} \left\{\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right\}^{-1} \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}\right\}.$$

  Under H_0 we have $T^2 \sim \chi^2_p$ approximately.

- When sample sizes are not large, but the two populations are multivariate normal, then $T^2 \sim \frac{vp}{v-p+1}F_{p,v-p+1}$ approximately, where v is

$$\nu = \frac{p + p^2}{\sum_{l=1}^{2} \frac{1}{n_l}(tr[\{\frac{1}{n_l}\mathbf{S}_l(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2)^{-1}\}^2] + [tr\{\frac{1}{n_l}\mathbf{S}_l(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2)^{-1}\}]^2)}.$$

Some comments:

- If the sample size is moderate to large, Hotelling's T^2 is is still quite robust even if there are slight departures from normality and/or a few outliers are present.

3.3 Independent populations comparison (for variables that have similar values). Profile Analysis

High School data example. School A and School B are two rival traditional middle schools in San Francisco. A researcher wants to get a sense of the level of preparation of the 7th graders in the two schools. She collects a random sample of students from each school and records the students' scores in the four core subjects: Math, ELA, Science and Social studies. The reseracher is interested in the following questions:

- Is the students' typical performance similar across the two schools?
- Is it true that the students perform at the same level for all the subjects on average, in the two schools?
- If the previous question is correct is the typical performance the same in the two schools ?

ID	Math	ELA	Science	SocialStudies	school
1	78.80	81.26	77.81	91.92	A
2	96.64	90.50	87.93	100.00	A
3	79.43	79.09	84.19	87.05	A
4	85.19	80.48	85.48	89.87	A
5	73.96	81.50	76.13	91.08	A
6	79.42	77.55	78.05	88.53	A
–	–	–	–	–	–
101	86.54	84.08	87.67	92.36	B
102	79.77	82.25	78.14	93.04	B
103	80.40	83.84	80.43	93.20	B
104	87.76	83.38	80.93	95.66	B
105	87.39	83.38	82.07	94.82	B
106	89.29	82.61	82.68	94.56	B

In class: Formulate mathematically the goal of the problem.

General setup:

- p variables (whose values are expressed in similar units, such as, questions, tests administered) given to two groups/populations
- $\mathbf{X}_{1j} = (X_{1j1}, \ldots, X_{1jp})^T$ - the response for the jth unit in the first group/population
- $\mathbf{X}_{2k} = (X_{2k1}, \ldots, X_{2kp})^T$ - the response for the kth unit in the second group/population
- Assumption: independent samples, $\mathbf{X}_{1j} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathbf{X}_{2k} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- The sample sizes for the groups are n_1 and n_2, respectively
- Two group mean vectors $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, \ldots, \mu_{1p})^{\mathsf{T}}$ and similarly define $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \ldots, \mu_{2p})^{\mathsf{T}}$

Questions of interest:

- Stage 1: are the two profiles parallel (change between the components of the mean vector is similat across the groups)?

$$\mu_{1i} - \mu_{1,i-1} = \mu_{2i} - \mu_{2,i-1}, \quad i = 2, 3, \ldots, p$$

- Stage 2: If parallel, are the two profiles coincident?

$$\mu_{1i} = \mu_{2i}, \quad , i = 1, 2, \ldots, p$$

- Stage 3: If coincident, are the means equal?

$$\mu_{11} = \ldots = \mu_{1p} = \mu_{21} = \ldots = \mu_{2p}$$

In class: Express (the null hypothesis implied by) each of these questions in form of $\mathbf{C}\boldsymbol{\mu}_1$ and $\mathbf{C}\boldsymbol{\mu}_2$.

Stage 1 (parallel profiles):

- $H_0 : \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$ for appropriate $(p-1) \times p$ dimensional matrix $\mathbf{C}$
- $T^2 = \{\mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^\mathsf{T} \left\{ (\frac{1}{n_1} + \frac{1}{n_2}) \mathbf{C}\mathbf{S}_{\text{pool}}\mathbf{C}^\mathsf{T} \right\}^{-1} \mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
- Under H_0, the test statistic follows $\frac{(n_1+n_2-2)(p-1)}{n_1+n_2-p} F_{p-1,n_1+n_2-p}$ distribution

Hint: use the result for the equality of two mean vectors from independent populations $\mathbf{C}\mathbf{X}_{1j}$ and $\mathbf{C}\mathbf{X}_{2k}$.

Stage 2 (coincident profiles, given they are parallel):

- $H_0 : \mathbf{l}^\mathsf{T}\boldsymbol{\mu}_1 = \mathbf{l}^\mathsf{T}\boldsymbol{\mu}_2$, where $\mathbf{l} = (1, \ldots, 1)^\mathsf{T}$
- $T^2 = \{\mathbf{l}^\mathsf{T}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\}^\mathsf{T} \left\{ (\frac{1}{n_1} + \frac{1}{n_2}) \mathbf{l}^\mathsf{T}\mathbf{S}_{\text{pool}}\mathbf{l} \right\}^{-1} \mathbf{l}^\mathsf{T}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$
- Under H_0, the test statistic follows F_{1,n_1+n_2-2} distribution

Stage 3 (constant profiles, given they are coincident):

- Combine two populations together, assume common mean $\boldsymbol{\mu}$
- $H_0 : \mathbf{C}\boldsymbol{\mu} = 0$
- $T^2 = (\mathbf{C}\bar{\mathbf{x}})^\mathsf{T} \left\{ \frac{1}{n_1+n_2} \mathbf{C}\mathbf{S}\mathbf{C}^\mathsf{T} \right\}^{-1} \mathbf{C}\bar{\mathbf{x}}$, where $\mathbf{S}$ is the covariance matrix based on all $n_1 + n_2$ observations
- Under H_0, the test statistic follows $\frac{(n_1+n_2-1)(p-1)}{n_1+n_2-p+1} F_{p-1,n_1+n_2-p+1}$ distribution

Hint: When the means of the two groups are the same, then we can use the overall sample mean as an estimator for the common mean; the pooled sample covariance becomes the sample covariance of the full sample (obtained by combining the two groups together).

Note that Stage 3 is appropriate only if all the questions are measured on the same scale.

In class: R code to analyze the motivating data application.