Inference on the mean of multivariate normal

Now we start on methodology of statistical inference, that is, using information in multivariate sample data to draw conclusions on population properties of interest.

Unless otherwise stated, we consider the dimension of the random vectors p ias fixed, of moderate size, while the sample size $n > p$.

In this section we discuss statistics inference methods used for one sample of multivariate variables. The sample data are treated as independent observations from a multivariate normal distribution.

The next section will consider inference methods on multiple samples, where sample data are treated as independent observations from several multivariate normal distributions.

1 One-sample test on the mean

We are interested in estimating the components $\mu_1, \cdots, \mu_p$ of the mean vector $\boldsymbol{\mu}$ jointly, taking the covariance relationship among the variables into consideration.

One common type of statistical inference on the mean is formulated as hypothesis testing. Although hypothesis testing appears to be a small part of statistical data analysis, the test statistic discussed here is closely related to more general inference such as providing confidence regions for $\boldsymbol{\mu}$.

The hypothesis test to be considered is

$$\begin{cases} H_o: & \boldsymbol{\mu} = \boldsymbol{\mu}_o \\ H_a: & \boldsymbol{\mu} \neq \boldsymbol{\mu}_o \end{cases}$$

2 Review: univariate hypothesis test and t-statistic

First recall the one sample t-test and its t-statistic on testing the mean in the univariate case $p = 1$, which you should have come across in previous courses.

For a random sample of size n, assume

$$X_1, \cdots, X_n \quad i.i.d. \quad \sim N(\mu, \sigma^2)$$

The hypotheses of interest is

$$H_o: \ \mu = \mu_o \quad vs. \quad H_a: \ \mu \neq \mu_o$$

Then under H_o, the t-test statistic T_1 stated below is of the Student's t-distribution with degrees of freedom $n-1$,

$$T_1 = \frac{\bar{X} - \mu_o}{s/\sqrt{n}} \ \sim \ t_{n-1} \qquad under \ H_o$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean — an unbiased estimator of the population mean μ with $\mathbb{E}(\bar{X}) = \mu$, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance — an unbiased estimator of the population variance σ^2 with $\mathbb{E}(s^2) = \sigma^2$, t_k represents the Student t distribution with k degrees of freedom, and $\sim$ denotes the probability distribution.

The rationale of hypothesis testing inference is that, if the X_i's are indeed independent draws from a $N(\mu_0, \sigma^2)$ random variable, then T_1 would be in the interval where the t_{n-1} distribution has high probability, that is, T_1 should be not too far away from 0. If the observed value of T_1 lands away from 0, then we conclude that the sample is not likely from a $N(\mu_0, \sigma^2)$ distribution. The cutoff of the decision boundary depends on our tolerance of the error that the data are from a $N(\mu_0, \sigma^2)$ distribution but fall to the far tail of the distribution just by chance.

Formally, at test level $\alpha = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true})$, which is the Type-I error that we decide to tolerate, H_o is rejected when the value of the sample test statistic $|T_1| \geq t_{n-1,\alpha/2}$, where $t_{n-1,\alpha/2} > 0$ is the upper $\frac{\alpha}{2}$ quantile of t_{n-1} distribution, defined by $\mathbb{P}\left(t_{n-1} > t_{n-1,\alpha/2}\right) = \frac{\alpha}{2}$.

Without the normality assumption of the X_i's, the t test statistic T_1 is approximately of t_{n-1} distribution when n is moderately large, in the spirit of the Central Limit Theorem.

Remarks on t-distributions and t-statistic

- Recall that, a random variable Y is of t-distribution of degrees of freedom d if

  $$Y = \frac{Z}{\sqrt{W/d}}$$

  with

  $$Z \sim N(0,1), \qquad W \sim \chi^2_d, \qquad Z \perp\!\!\!\perp W,$$

  where $\perp\!\!\!\perp$ denotes independent relationship, $\sim$ the probability distribution of the random variable, $N(0,1)$ the standard normal distribution, and χ^2_d the chi-square distribution with degrees of freedom d.

- It is clear that t distribution is closely related to normal distribution and chi-square distribution.

  Recall that, a random variable X is of chi-square distribution with degrees of freedom k, if

  $$X = Z_1^2 + \cdots + Z_k^2, \qquad with \quad Z_i's \quad i.i.d. \ \sim N(0,1)$$

  That is, a chi-square random variable with degrees of freedom k can be expressed as the sum of squares of k independent standard normal random variables.

  Consequently the density function of χ^2_k-distribution can be derived from the density of standard normal (likely done in STAT245x0) by applying variable substitution method to the density functions.

  χ^2_k distribution is a special case of gamma distribution with the shape parameter $\frac{k}{2}$ and rate parameter $\frac{1}{2}$.

- From the density functions of normal distribution and chi-square distribution, the density of t distribution can be derived. The density function of t-distribution of degrees of freedom d is

  $$f(t) = \frac{1}{\sqrt{d}\, B\left(\frac{1}{2}, \frac{d}{2}\right)} \left(1 + \frac{t^2}{d}\right)^{-\frac{d+1}{2}}, \quad t \in \mathbb{R}.$$

  In the above, $B(\cdot,\cdot)$ denotes the Beta function

  $$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}dx, \ a,b > 0,$$

  where $\Gamma(\cdot)$ is the Gamma function

  $$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$$

  Some most useful properties of the Gamma function are: $\Gamma(1) = 1$, $\Gamma(z+1) = z\Gamma(z)$, thus $\Gamma(k) = (k-1)!$ for integer k, and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

- The distribution of the test statistic t asymptotically approaches the standard normal distribution as $n \to \infty$. The proof of the asymptotic distribution uses the exponential limit formula

$$e^x = \lim_{x\to\infty}\left(1+\frac{x}{n}\right)^x$$

and the Gamma function property

$$\lim_{n\to\infty}\frac{\Gamma(n+c)}{\Gamma(n)n^c} = 1$$

for scalar c.

- The above leads to the useful result that if $\{X_1, \cdots, X_n\}$ is an i.i.d. sample from a general probability distribution with mild regularity conditions such as having second moment, then $\bar{X}$ is asymptotically of normal distribution for large n, by the Central Limit Theorem, and the t statistic is asymptotically approaching $N(0,1)$.

If we square the t-statistic then rewrite the test statistic as

$$T_1^2 = (\bar{x}-\mu_o)\left(\frac{s^2}{n}\right)^{-1}(\bar{x}-\mu_o) = (\bar{x}-\mu_o)\left[Var(\bar{x}-\mu_o)\right]^{-1}(\bar{x}-\mu_o)$$

We then obtain a convenient form suitable for generalization of the t-statistic to multidimensional cases.

3 Hotelling's T^2 statistc

The multivariate normal generalization of the t-test is to test H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_o$, when $\boldsymbol{\mu}, \boldsymbol{\mu}_o \in \mathbb{R}^p$ with $p \geq 2$.

Recall that observed multivariate data can be displayed as

$$\begin{array}{l}\text{observation 1}\\ \text{observation 2}\\ \vdots\\ \text{observation } j\\ \vdots\\ \text{observation } n\end{array}\begin{bmatrix}\boldsymbol{x}_1'\\ \boldsymbol{x}_2'\\ \vdots\\ \boldsymbol{x}_j'\\ \vdots\\ \boldsymbol{x}_n'\end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np}\end{bmatrix}$$

The data can be viewed as realizations of n i.i.d. multivariate vector,

$$\begin{array}{l}\text{random vector 1}\\ \text{random vector 2}\\ \vdots\\ \text{random vector } j\\ \vdots\\ \text{random vector } n\end{array}\begin{bmatrix}\boldsymbol{X}_1'\\ \boldsymbol{X}_2'\\ \vdots\\ \boldsymbol{X}_j'\\ \vdots\\ \boldsymbol{X}_n'\end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p}\\ X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp}\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np}\end{bmatrix}$$

Assuming the random sample $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ are i.i.d. from an $N_p(\boldsymbol{\mu}, \Sigma)$, a p-variate multivariate normal distribution. The sample mean vector is

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$$

and sample variance-covariance matrix is

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^n (\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})'$$

Denote the proposed null mean vector

$$\boldsymbol{\mu}_0 = \begin{bmatrix}\mu_{10}\\ \vdots\\ \mu_{p0}\end{bmatrix}$$

The test statistic is **Hotelling's $\boldsymbol{T}^2$**,

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

To test the hypotheses

$$H_o:\ \boldsymbol{\mu} = \boldsymbol{\mu}_o \quad vs. \quad H_a:\ \boldsymbol{\mu} \neq \boldsymbol{\mu}_o$$

Under the multivariate normality assumption, the test statistic T^2 is related to the F distribution under the null,

$$T^2 \ \sim\ \frac{(n-1)p}{n-p}\,F_{p,n-p} \qquad under \quad H_o$$

With the observed data $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ with sample mean $\bar{\boldsymbol{x}}$, given a test level α, we would reject H_o if

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

where $F_{p,n-p}(\alpha)$ is the upper 100α percentile of $F_{p,n-p}$, that is, $\mathbb{P}\{F_{n,n-p} > F_{p,n-p}(\alpha)\} = \alpha$.

Remarks on T^2

- Analogous to the t-statistic in the univariate test, T^2 is a sample statistic, its probability distribution does not involve $\boldsymbol{\mu}$ or Σ, which consist of unknown parameters.
- T^2 is invariant under non-singular linear transformation of the variables $\boldsymbol{X}_{p\times 1} \to C_{p\times p}\boldsymbol{X}_{p\times 1} + d_{p\times 1}$ (exercise).
- The $\boldsymbol{S}$ in Hotelling's T^2 is the $p \times p$ sample covariance matrix.

  Elementwise,

$$\boldsymbol{S} = \left[\frac{1}{n-1}\sum_{j=1}^n (x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)\right]_{i,k=1,\cdots,p} = [s_{ik}]_{i,k=1,\cdots,p}$$

  where s_{ik} is the sample covariance between the ith and kth variables, as defined at the beginning of the course.

  In vector-matrix form of the observed sample data,

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{j=1}^n (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$$

  The p-variate vector $\boldsymbol{x}_j$ is the jth observation. In other words, $(n-1)\boldsymbol{S}$ can be written as the sum of n matrices, each of the n matrices is of dimensions $p \times p$, the jth matrix has the form

$$(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' = \begin{bmatrix} x_{j1}-\bar{x}_1\\ x_{j2}-\bar{x}_2\\ \vdots\\ x_{jk}-\bar{x}_k\\ \vdots\\ x_{jp}-\bar{x}_p\end{bmatrix}\left[x_{j1}-\bar{x}_1,\ x_{j2}-\bar{x}_2,\ \cdots,\ x_{jk}-\bar{x}_k,\ \cdots,\ x_{jp}-\bar{x}_p\right] = \left[(x_{ji}-\bar{x}_i)(x_{jk}-\bar{x}_k)\right]_{i,k=1,\cdots,p}$$

  where $j = 1, \cdots, n$ index the observations.

- Write the sample covariance matrix as $\boldsymbol{S} = [s_{ik}]_{p\times p}$, and assume that $\boldsymbol{S}$ is of full rank, that is, the p columns are linearly independent. Denote the inverse of the sample covariance matrix as

$$\boldsymbol{S}^{-1} = [s^{ik}]_{p\times p}$$

Then Hotelling's T^2 can be written as a sum of p^2 univariate terms:

$$T^2 = (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \left(\frac{\mathbf{S}}{n}\right)^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_o) = n\sum_{i=1}^{p}\sum_{k=1}^{p} s^{ik}(x_i - \mu_{i0})(x_k - \mu_{k0})$$

- Hotelling's T^2 is of $\frac{(n-1)p}{n-p}F_{p,n-p}$ distribution under the null hypothesis.

The F-distribution of T^2 is derived from the independence between $\bar{\boldsymbol{X}}$ & $\boldsymbol{S}$, and the distribution of $\boldsymbol{S}^{-1}$ (an inverse Wishart distribution, not covered in this class).

The proof is complicated and beyond the coverage of this course, thus omitted.

A complete, detailed proof can be found in Chapter 3 of Muirhead (1982).

- When $n \to \infty$, T^2 is asymptotically of χ^2_p distribution.

However the test using T^2 is more useful when n is not very large.

- Review of F distribution
  - A random variable is of F-distribution with degrees of freedom d_1 and d_2, denoted as F_{d_1,d_2}, if it is the ratio of two independent Chi-square random variables divided by their degrees of freedom respectively,

$$\frac{Y_1/d_1}{Y_2/d_2}, \qquad with \quad Y_1 \sim \chi^2_{d_1}, \quad Y_1 \sim \chi^2_{d_2}$$

and the numerator is independent of the denominator,

$$Y_1 \perp\!\!\!\perp Y_2.$$

In the above, χ^2_d denotes the Chi-square distribution with degrees of freedom d.

  - The density of F_{d_1,d_2} is

$$f(y)_{d_1,d_2} = \frac{1}{B\left(\frac{d_1}{2},\frac{d_2}{2}\right)}\left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} y^{\frac{d_1}{2}-1}\left(1+\frac{d_1}{d_2}y\right)^{-\frac{d_1+d_2}{2}}, \qquad y>0.$$

  - In particular, when $d_1 = p = 1,\ d_2 = n-p = n-1$, the $F_{p,n-p}$ density becomes

$$f(y)_{1,n-1} = \frac{1}{B\left(\frac{1}{2},\frac{n-1}{2}\right)}\left(\frac{1}{n-1}\right)^{\frac{1}{2}} y^{-\frac{1}{2}}\left(1+\frac{y}{n-1}\right)^{-\frac{n}{2}}, \qquad y>0,$$

which is the density of $(t_{n-1})^2$, the square of a t-distribution random variable with degrees of freedom $n-1$.

- Note that $p < n$ is assumed in the above derivation and application.

When $p > n$, inverse of covariance matrix does not exist. Even if pseudo-inverse matrix is used, T^2 test loses power (Bai and Saranadasa 1996). Various remedies have been studied and proposed.

4 Hotelling's T^2 and likelihood ratio tests

Hotelling's T^2 for multivariate normal is equivalent to a likelihood ratio test.

Let θ be a parameter of interest, $\theta \in \Theta$, the whole parameter space. To test if the true parameter value of θ lies inside a specific subspace Θ_0, the hypothesis test can be expressed as

$$\begin{cases} H_o: & \theta \in \Theta_0 \subset \Theta \\ H_a: & \theta \in \Theta \setminus \Theta_o \end{cases}$$

For example, we may test on $\Theta_0 = \{(0,0)\}, \Theta = \mathbb{R}^2$.

Likelihood Ratio

The Likelihood Ratio test statistic is the ratio of the maximum likelihood under H_0 and the unrestricted maximum likelihood under a more general $H_o \cup H_a$.

$$\Lambda = \frac{\max_{\theta\in\Theta_0} L(\theta)}{\max_{\theta\in\Theta} L(\theta)}, \qquad where \quad \Theta_0 \subset \Theta.$$

- Possible values of Λ are in $[0,1]$.
- When H_0 is true, Λ is close to 1.
- When H_a is true, Λ is closer to 0.
- Consequently, H_0 is rejected if $\Lambda < c$, where c is a critical value depending on the test level α ($type\text{-}I$ error).
- When the sample size n is large, often $-2\ln\Lambda$ is used as the likelihood ratio test statistic, with asymptotic χ^2 distribution under the null,

$$-2\ln\Lambda \ \sim\ \chi^2_{v-v_0} \qquad under \quad H_o \qquad \text{(asymptotically)}$$

where

$$v_0 = dim(\Theta_0), \quad v = dim(\Theta).$$

Maximum likelihood and Likelihood Ratio for multivariate normal

To test the mean parameter vector of multivariate normal, the likelihood ratio test has a simpler form.

We have derived (in lecture notes "The multivariate normal distribution" in week 2) that the maximum of the likelihood function for a multivariate normal sample of size n is

$$\max_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu},\Sigma) = L(\hat{\boldsymbol{\mu}},\hat{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}|^{n/2}} e^{-np/2}$$

where the ML estimators for the population mean vector and variance-covariance matrix are

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}, \qquad \hat{\Sigma} = \boldsymbol{S}_n = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$$

Therefore, under the null hypothesis $H_o : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the maximum likelihood has the form

$$\max_{\Sigma} L(\boldsymbol{\mu}_0,\Sigma) = L(\hat{\boldsymbol{\mu}}_0,\hat{\Sigma}_0) = \frac{1}{(2\pi)^{np/2}|\hat{\Sigma}_0|^{n/2}} e^{-np/2}$$

where the ML estimator for the variance-covariance matrix under the null H_o is

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'.$$

For $\boldsymbol{\mu}_0 \in \mathbb{R}^p$, consider the hypotheses

$$\begin{cases} H_o: & \boldsymbol{\mu} = \boldsymbol{\mu}_0 \\ H_a: & \boldsymbol{\mu} \neq \boldsymbol{\mu}_0 \end{cases}$$

The resulting likelihood ratio has a simple form.

$$\Lambda = \frac{\max\limits_{\Sigma} L(\boldsymbol{\mu}_0, \Sigma)}{\max\limits_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu}, \Sigma)} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|}\right)^{n/2} = \left(\frac{\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right|}{\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\right|}\right)^{n/2}$$

The statistics $\Lambda^{2/n}$ is called the *Wilks' lambda*.

Wilks' Lambda and Hotelling's T^2

The following gives the relation between the likelihood ratio and Hotelling's T^2.

$$\text{Wilks' Lambda} \qquad \Lambda^{2/n} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_0|} = \left(1 + \frac{T^2}{n-1}\right)^{-1}$$

$$\text{Hotelling's } T^2 \qquad T^2 = (n-1)\left(\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|} - 1\right)$$

Proof. In the following we derive the relation between Λ and T^2.

We use the following matrix algebra result on block matrix determinants:

$$\left|\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right| = |A_{11}|\,|A_{22} - A_{21}A_{11}^{-1}A_{12}| = |A_{22}|\,|A_{11} - A_{12}A_{22}^{-1}A_{21}| \tag{1}$$

under the conditions that the diagonal block are invertible square matrices: $|A_{11}| \neq 0, |A_{22}| \neq 0$.

The results in (1) follow from the Schur complements of block matrices.

Assuming sample variance-covariance matrix $\boldsymbol{S}$ is of full rank, (1) can be applied to the block matrix

$$\left|\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right| = \left|\begin{array}{cc} \sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' & \sqrt{n}(\bar{\boldsymbol{x}} - \mu_0) \\ \sqrt{n}(\bar{\boldsymbol{x}} - \mu_0)' & -1 \end{array}\right|,$$

We obtain

$$\begin{aligned} |A_{22}|\,|A_{11} - A_{12}A_{22}^{-1}A_{21}| &= (-1)\left|\textstyle\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' + n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\right| \\ &= (-1)\left|\textstyle\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\right| \end{aligned} \tag{2}$$

and

$$\begin{aligned} |A_{11}|\,|A_{22} - A_{21}A_{11}^{-1}A_{12}| &= \left|\textstyle\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right| \\ &\times \left|-1 - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\left(\textstyle\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right)^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right| \end{aligned} \tag{3}$$

Notice that

$$\begin{aligned} &\left|-1 - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\left(\textstyle\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right)^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right| \\ &= -1 - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\left[(n-1)\boldsymbol{S}\right]^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)' \\ &= (-1)\left(1 + \frac{1}{n-1}n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\right) \\ &= (-1)\left(1 + \frac{1}{n-1}T^2\right) \end{aligned}$$

where Hotelling's $T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)$.

From (1), The right hand sides of (2) and (3) are equal.

$$(-1)\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\right| = (-1)\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right|\left(1 + \frac{1}{n-1}T^2\right)$$

The equality above leads to the desired result

$$\Lambda^{2/n} = \frac{\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'\right|}{\left|\sum_{j=1}^{n}(\boldsymbol{x}_j - \boldsymbol{\mu}_0)(\boldsymbol{x}_j - \boldsymbol{\mu}_0)'\right|} = \left(1 + \frac{1}{n-1}T^2\right)^{-1}$$

which shows the desired relationship between Λ and T^2. □

Usefulness of the alternative expression of T^2

There is a need to compute the inverse matrix $\boldsymbol{S}^{-1}$ in evaluating Hotelling's T^2 using $T^2 = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)'$. Calculating $\boldsymbol{S}^{-1}$ is computationally demanding, unless the dimensions is quite low, thus often avoided. The new formula $T^2 = (n-1)(|\hat{\Sigma}_0|/|\hat{\Sigma}| - 1)$ is often easier to evaluate.

5 Confidence regions

Let $\boldsymbol{\theta}$ be the vector of parameters of interest, $\boldsymbol{\theta} \in \Theta$, the parameter space. In the univariate case, a parameter estimate $\hat{\theta} = \hat{\theta}(X)$ based on a random sample $X = \{X_1, \cdots, X_n\}$ is often presented along with a confidence interval $\mathbb{P}(L(X) < \theta < U(X)) = 1 - \alpha$, where L and U can be calculated from the sample X, similar to $\hat{\theta}$ which can be evaluated from sample values of X. The region (L, U) is a $100(1-\alpha)\%$ confidence interval for the unknown parameter θ.

Analogously, in multivariate inference, a $100(1-\alpha)\%$ **confidence region** $R(\boldsymbol{X})$ for $\boldsymbol{\theta}$ based on a p-variate random sample $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$ is defined as

$$\mathbb{P}(\boldsymbol{\theta} \in R(\boldsymbol{X})) = 1 - \alpha,$$

in the sense that among all possible sample data $\boldsymbol{X}$, the parameter vector $\boldsymbol{\theta}$ is in the the region $R(\boldsymbol{X})$ about $(1-\alpha)100\%$ of the time. The probability in the above equality is with respect to the true (unknown) parameter $\boldsymbol{\theta}$.

Confidence region of the mean vector by T^2

Recall

$$n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) = T^2 \;\sim\; \frac{(n-1)\,p}{n-p}\,F_{p,n-p} \qquad \textit{under } H_0.$$

where $F_{d_1,d_2}(\alpha)$ is the upper $\alpha \times 100\%$ percentile of the F distribution F_{d_1,d_2}, so $\mathbb{P}[F_{d_1,d_2} \geq F_{d_1,d_2}(\alpha)] = \alpha$.

The $(1-\alpha)100\%$ T^2 confidence region of the mean $\boldsymbol{\mu}$ based on the sample mean $\bar{\boldsymbol{X}}$ is derived from

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

Remarks on T^2 confidence region

- When $\bar{\boldsymbol{X}}$ is multivariate normal and $\boldsymbol{\mu}$ is known (only in theory), the $\bar{\boldsymbol{x}}$ values satisfying the inequality

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha) \tag{4}$$

  gives an elliptically shaped region centered at $\boldsymbol{\mu}$, in which sample mean vector $\bar{\boldsymbol{x}}$ falls with probability $1-\alpha$.

- When $\bar{\boldsymbol{x}}$ is from an i.i.d. sample of multivariate normal and $\boldsymbol{\mu}$ is unknown (as in practice), plausible values of $\boldsymbol{\mu}$ satisfying (4) form a random region of ellipsoid centered at sample mean $\bar{\boldsymbol{x}}$, and the region includes the true population mean vector $\boldsymbol{\mu}$ with confidence level $(1-\alpha)100\%$. This is the T^2 **confidence region** of the mean vector.

- When $p = 1$, the inequality gives the usual $(1-\alpha)100\%$ confidence interval constructed by t-test.

- The Euclidean distance between vectors $\boldsymbol{x} = (x_1, \cdots, x_p)'$ and $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)'$ is

$$\sqrt{(x_1 - \mu_1)^2 + \cdots + (x_p - \mu_p)^2} = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})'(\boldsymbol{x} - \boldsymbol{\mu})}$$

  The Mahalanobis distance ($a.k.a.$ **statistical distance**), a type of standardized distance, of variable vector $\boldsymbol{x}$ to constant vector $\boldsymbol{\mu}$ is defined as

$$\sqrt{(\boldsymbol{x} - \boldsymbol{\mu})' \widehat{Cov}(\boldsymbol{x})^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}$$

  where $\widehat{Cov}(\boldsymbol{x})$ denotes the sample covariance matrix of $\boldsymbol{x}$.

- The Mahalanobis distance takes into consideration the covariance structure of the data. Hotelling's T^2 uses the Mahalanobis distance between the mean $\bar{\boldsymbol{x}}$ and $\boldsymbol{\mu}_0$ to test $H_0: \boldsymbol{\mu} = \boldsymbol{\mu}_0$. The T^2 Confidence Region uses the statistical distance to construct a confidence region for the population mean $\boldsymbol{\mu}$.

- For multivariate normal $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, the contours of constant density function

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

  are ellipsoids in $\mathbb{R}^p$, centered at $\boldsymbol{\mu}$ with axes proportional to $\sqrt{\lambda_i}\boldsymbol{e}_i$, where $\Sigma \boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, i = 1, \cdots, p$.
  Consequently, the constant contours of T^2 confidence region are also ellipsoids in $\mathbb{R}^p$, centered at $\bar{\boldsymbol{x}}$ with axes proportional to $\sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)}\,\boldsymbol{e}_i$, where $\boldsymbol{S}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$ for $i = 1, \cdots, p$.

6 Simultaneous confidence intervals of component means

Other than the above ellipsoidal confidence region formed by $\bar{\boldsymbol{x}}$, in many cases it is still desirable to have a statement

"$\mu_k \in (a_k, b_k)$ for all $k = 1, \cdots, p$ "

with a reasonable level of confidence.

Such statement must involve all $\bar{x}_k$, forming a hyper-rectangular region $\in \mathbb{R}^p$. Compared with the advantages of ellipsoidal confidence region using statistical distance, these hyper-rectangles (orthotopes) are often easier to form and to compute.

Confidence interval for a linear combination of component means

To relate a component mean $\bar{x}_k$ with the mean vector $\bar{\boldsymbol{x}}$, notice that $\bar{x}_k$ can be expressed as $\boldsymbol{a}'\bar{\boldsymbol{x}}$, where $\boldsymbol{a} = [\,0 \cdots 0\ 1\ 0 \cdots 0\,]'$ has the kth component $= 1$ and 0 for other components.

First we will derive a confidence interval for $\boldsymbol{a}'\bar{\boldsymbol{\mu}}$, a linear combination of the component means.

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then any linear combination of its components (written as $\boldsymbol{a}'\boldsymbol{X}$) is a univariate normal,

$$\boldsymbol{a}'\boldsymbol{X} \sim N(\mu, \sigma^2), \qquad with \quad \mu = \boldsymbol{a}'\boldsymbol{\mu}, \quad \sigma^2 = \boldsymbol{a}'\Sigma\boldsymbol{a}$$

Especially, the mean vector $\bar{\boldsymbol{X}}$ of a random sample has the same property.

$$\bar{\boldsymbol{X}} \sim N_p\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right) \qquad \Rightarrow \quad \boldsymbol{a}'\bar{\boldsymbol{X}} \sim N\left(\boldsymbol{a}'\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{a}'\Sigma\boldsymbol{a}\right)$$

If $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ is a p-variate random sample from a common distribution with sample mean $\bar{\boldsymbol{x}}$ and sample covariance matrix $\boldsymbol{S}$, then for any $\boldsymbol{a} \in \mathbb{R}^p \setminus \{\boldsymbol{0}\}$,

$$\begin{cases} y_1 = \boldsymbol{a}'\boldsymbol{x}_1 \\ y_2 = \boldsymbol{a}'\boldsymbol{x}_2 \\ \vdots \\ y_n = \boldsymbol{a}'\boldsymbol{x}_n \end{cases}$$

is a random sample from the univariate normal variable $Y = \boldsymbol{a}'\boldsymbol{X}$, with univariate sample mean and sample variance

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \boldsymbol{a}'\bar{\boldsymbol{x}}, \qquad s_y^2 = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}.$$

For a fixed vector $\boldsymbol{a} \in \mathbb{R}^p$, the t-statistic

$$t = \frac{\bar{y} - \mu_y}{\sqrt{s_y^2/n}} = \frac{\boldsymbol{a}'\bar{\boldsymbol{x}} - \boldsymbol{a}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}/n}} \sim t_{n-1}$$

since $\mu_y = \boldsymbol{a}'\boldsymbol{\mu}$ is the true mean of Y. Therefore a $(1-\alpha)100\%$ confidence interval of parameter $\mu_y = \boldsymbol{a}'\boldsymbol{\mu}$ is

$$\boldsymbol{a}'\bar{\boldsymbol{x}} - t_{n-1,\alpha/2}\frac{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}}{\sqrt{n}} \;\leq\; \boldsymbol{a}'\boldsymbol{\mu} \;\leq\; \boldsymbol{a}'\bar{\boldsymbol{x}} + t_{n-1,\alpha/2}\frac{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}}{\sqrt{n}} \tag{5}$$

In other words, for an observed $\bar{\boldsymbol{x}}$ and fixed $\boldsymbol{a} \in \mathbb{R}^p$, the $(1-\alpha)100\%$ confidence interval consists of plausible $\boldsymbol{\mu}$ values for which $\left|\frac{\boldsymbol{a}'\bar{\boldsymbol{x}} - \boldsymbol{a}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}/n}}\right| \leq t_{n-1,\alpha/2}$, or equivalently,

$$\frac{n[\boldsymbol{a}'(\bar{\boldsymbol{x}} - \boldsymbol{\mu})]^2}{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} \leq t_{n-1,\alpha/2}^2 \tag{6}$$

Thus, if we only want to construct only one confidence interval for a specific μ_k, we choose $\boldsymbol{a} = [0 \cdots 0\ 1\ 0 \cdots 0]'$ with the kth component $= 1$ and 0 otherwise. Then the above inequality (6) becomes

$$\frac{n[\boldsymbol{a}'(\bar{\boldsymbol{x}} - \boldsymbol{\mu})]^2}{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} = \frac{(\bar{x}_k - \mu_k)^2}{s_{kk}/n} \leq t_{n-1,\alpha/2}^2$$

which yields a confidence interval for a single μ_k.

6.1 Hyper-rectangle by marginal confidence intervals of component means (naive)

If we ignore the multivariate nature and treat the component variables as independent, a marginal confidence interval of each component mean μ_k can be obtains from the t-distribution,

$$\bar{x}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \tag{7}$$

This can be derived from univariate one-sample t-statistic. A single such interval contains its mean parameter μ_k with $(1-\alpha)100\%$ confidence, ignoring or without considering probability statements on other components. However in general we do not know the confidence level that the p intervals contain all μ_k simultaneously.

The only exception is when the component variables are independent, then

$$\begin{aligned}
&P\left\{\bar{x}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}},\ \forall k = 1,\cdots,p\right\}\\
&= \prod_{k=1}^{p} P\left\{\bar{x}_k - t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + t_{n-1,\alpha/2}\sqrt{\frac{s_{kk}}{n}}\right\}\\
&= (1-\alpha)^p \ < \ 1-\alpha
\end{aligned}$$

That is, the p marginal intervals in the form of (7) contain all μ_k simultaneously would have confidence level $(1-\alpha)^p 100\%$ when independence holds. The coverage probability of the p simultaneous confidence intervals would be $(1-\alpha)^p < 1-\alpha$, and independence among component variables is not true or interesting in multivariate studies.

6.2 Bonferroni simultaneous confidence intervals of component means

In order to have a desirable confidence level for the p confidence intervals in (7) to contain all μ_k simultaneously, we need to pay the price of making the intervals wider, according to the commonly used method of Bonferroni.

Let C_k = the event that the confidence interval of μ_k contains μ_k with confidence level $(1-\alpha_k)100\%$. So

$$P(C_k = true) = 1 - \alpha_k$$

for a given $k = 1,\cdots,p$. To achieve the desired confidence level for simultaneous confidence intervals, we need

$$P(C_k = true, k = 1,\cdots,p) \ \ge \ 1-\alpha$$

The Bonferroni method is based on the following probability derivation.

$$\begin{aligned}
P(C_k = true, k = 1,\cdots,p) &= 1 - P(C_k = false, for\ some\ k)\\
&\ge 1 - \sum_{k=1}^{p} P(C_k = false) \ = \ 1 - \sum_{k=1}^{p}[1 - P(C_k = true)]\\
&= 1 - \sum_{k=1}^{p}\alpha_k = 1 - (\alpha_1 + \cdots + \alpha_p)
\end{aligned}$$

The desired $(1-\alpha)100\%$ confidence level for p simultaneous confidence intervals can be guaranteed if $\sum_{i=1}^{p}\alpha_i \le \alpha$.

$$\alpha_1 + \cdots + \alpha_p \ \le \ \alpha \qquad \Longrightarrow \qquad P(C_k = true, k = 1,\cdots,p) \ \ge \ 1-\alpha$$

A common, convenient choice is to pick each $\alpha_k = \alpha/p$, then we obtain the Bonferroni confidence intervals

$$\bar{x}_k - t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}} \ \le \ \mu_k \ \le \ \bar{x}_k + t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}}, \qquad k = 1,\cdots,p.$$

The statement is true simultaneously for all $k = 1,\cdots,p$, with confidence level at least $(1-\alpha)100\%$.

$$P\left(\bar{x}_k - t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}} \le \ \mu_k \ \le \bar{x}_k + t_{n-1,\alpha/2p}\sqrt{\frac{s_{kk}}{n}},\ k = 1,\cdots,p\right) \ge 1-\alpha$$

6.3 Asymptotic approximation and simultaneous confidence intervals

When the sample size n is large relative to p (i.e. when $n-p$ is large), the Central Limit Theorem leads to

$$n(\bar{X}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{X}-\boldsymbol{\mu}) \sim \chi^2_p \qquad \text{(approximately asymptotically)}$$

without assuming normality of the distribution.

The χ^2 approximation gives the asymptotic simultaneous confidence intervals

$$\bar{x}_k - \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + \sqrt{\chi^2_p(\alpha)}\sqrt{\frac{s_{kk}}{n}}, \qquad k = 1,\cdots,p.$$

The above asymptotic statement comes from the multivariate normal property (exercise)

$$\boldsymbol{X} \sim N_p(\boldsymbol{\mu},\Sigma) \qquad \Rightarrow \quad (\boldsymbol{X}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_p$$

and the more general confidence interval result for any linear combination $\boldsymbol{a}'\boldsymbol{\mu}$, that when $n-p$ is sufficiently large,

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

is a $(1-\alpha)100\%$ confidence interval of $\boldsymbol{a}'\boldsymbol{\mu}$ asymptotically approximately.

6.4 Simultaneous confidence intervals of several component means by T^2

The inequality (6) is developed for a fixed $\boldsymbol{a}$. In the following we will use Hotelling's T^2 test statistic to construct an analogous result for all $\boldsymbol{a}$, which leads to hyper-rectangular simultaneous confidence intervals of $\bar{\mu}_k, k = 1,\cdots,p$.

The left hand side of (6) is in the form

$$\frac{(\boldsymbol{v}'\boldsymbol{w})^2}{\boldsymbol{v}'B\boldsymbol{v}}, \qquad with \quad \boldsymbol{v} = \boldsymbol{a},\ \boldsymbol{w} = \bar{\boldsymbol{x}} - \boldsymbol{\mu},\ B = \frac{1}{n}\boldsymbol{S}.$$

A useful extension of the Cauchy-Schwarz Inequality

$$(\boldsymbol{v}'\boldsymbol{w})^2 \le \|\boldsymbol{v}\|^2\|\boldsymbol{w}\|^2 = (\boldsymbol{v}'\boldsymbol{v})(\boldsymbol{w}'\boldsymbol{w})$$

is the Maximization Lemma,

$$\frac{(\boldsymbol{v}'\boldsymbol{w})^2}{\boldsymbol{v}'B\boldsymbol{v}} \le \boldsymbol{w}B^{-1}\boldsymbol{w}$$

where the equality holds if and only if $\boldsymbol{v} \propto B^{-1}\boldsymbol{w}$ ($\propto$ means proportional to).

The Maximization Lemma implies that the left side of (6) achieves maximum when we choose $\boldsymbol{a} \propto n\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})$.

Therefore for all $\boldsymbol{a} \in \mathbb{R}^p$, we have

$$\frac{n[\boldsymbol{a}'(\bar{\boldsymbol{x}} - \boldsymbol{\mu})]^2}{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} \le \max_{\boldsymbol{a}} \frac{n[\boldsymbol{a}'(\bar{\boldsymbol{x}} - \boldsymbol{\mu})]^2}{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}} = n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) = T^2 \tag{8}$$

The last expression above is in the form of Hotelling's T^2 statistic. Since $\boldsymbol{\mu}$ is the true mean,

$$n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu}) = T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p}$$

Thus the T^2 in (8) satisfies the probability property

$$P\left(T^2 \le \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right) = 1-\alpha,$$

where $F_{d_1,d_2}(\alpha)$ denotes the upper α quantile of the F_{d_1,d_2} distribution, as defined earlier.

Since the left side inequality in (8) holds for all $\boldsymbol{a} \in \mathbb{R}^p$,

$$\left|\frac{\boldsymbol{a}'\bar{\boldsymbol{x}}-\boldsymbol{a}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}/n}}\right| \le T^2, \qquad \forall \boldsymbol{a} \in \mathbb{R}^p,$$

from the probability property of T^2 we obtain

$$P\left(\left|\frac{\boldsymbol{a}'\bar{\boldsymbol{x}}-\boldsymbol{a}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}/n}}\right| \le \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\right) \ge 1-\alpha, \qquad \forall \boldsymbol{a} \in \mathbb{R}^p.$$

In other words,

$$P\left(|\boldsymbol{a}'\bar{\boldsymbol{x}}-\boldsymbol{a}'\boldsymbol{\mu}| \le \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\,\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}}\right) \ge 1-\alpha, \qquad \forall \boldsymbol{a} \in \mathbb{R}^p.$$

The probability statement gives an at least $(1-\alpha)100\%$ confidence interval of parameter $\boldsymbol{a}'\boldsymbol{\mu}$, for any $\boldsymbol{a} \in \mathbb{R}^p$,

$$P\left(\boldsymbol{a}'\bar{\boldsymbol{x}} - \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\,\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}} \;\le\; \boldsymbol{a}'\boldsymbol{\mu} \;\le\; \boldsymbol{a}'\bar{\boldsymbol{x}} + \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\,\frac{\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}}{n}}\right) \ge 1-\alpha. \qquad (9)$$

In particular, when we choose vector $\boldsymbol{a}$ to be $\boldsymbol{a} = [0\ \cdots\ 0\ 1\ 0\ \cdots\ 0]'$ with the kth component $= 1$ and 0 otherwise, then

$$\boldsymbol{a}\bar{\boldsymbol{x}} = \bar{x}_k, \qquad \boldsymbol{a}\bar{\boldsymbol{\mu}} = \mu_k, \qquad \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} = s_{kk} = s_k^2,$$

we obtain a $(1-\alpha)100\%$ confidence interval for μ_k, the populations mean parameter for the kth variable. Because (9) holds for any and all $\boldsymbol{a} \in \mathbb{R}^p$, we can obtain such confidence interval for $\mu_k, k = 1, \cdots, p$ simultaneously,

$$\bar{x}_k - \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\,\frac{s_{kk}}{n}} \le \mu_k \le \bar{x}_k + \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\,\frac{s_{kk}}{n}}, \qquad \text{for all } k = 1, \cdots, p. \qquad (10)$$

The intervals in (10) are called the simultaneous T^2 confidence intervals for the means. The simultaneous confidence intervals form a hyper-rectangle in $\mathbb{R}^p$.

Remarks on the T^2 confidence intervals for the means

- In fact, any confidence interval on a linear combination of the μ_k's can be obtained "for free" from the general statement in (9). For example, we can obtain confidence intervals for differences of the means of two component variables $\mu_k - \mu_i$.
- The simultaneous confidence intervals in (10) derived using T^2 is a hypercube or hyper-rectangle in $\mathbb{R}^p$, compared with the T^2 confidence region (4) which is an ellipsoid in $\mathbb{R}^p$. Both are useful.
- Because the statement in (9) is so general, the result for any specific $\boldsymbol{a} \in \mathbb{R}^p$ is on the conservative side, ending up with wider confidence intervals, sometime too wide to be practically informative.

6.5 Summary of simultaneous confidence intervals of component mean

We have discussed four types of hyper-rectangle confidence regions for component means. The hype-rectangle is formed by p simultaneous confidence intervals for component mean parameter $\mu_k, k = 1, \ldots, p$, in the form

$$\bar{x}_k \pm (\text{"Multiplier"})\sqrt{\frac{s_{kk}}{n}}, \qquad k = 1, \cdots, p.$$

Different methods are based on different probability distributions, thus use different multipliers. The multipliers are multiple of critical values of various probability distributions.

- Marginal confidence intervals using t statistics, ignoring dependence among component variables.

$$\text{Multiplier} = t_{n-1,\frac{\alpha}{2}}$$

- Bonferroni simultaneous confidence intervals using t statistics

$$\text{Multiplier} = t_{n-1,\frac{\alpha}{2p}}$$

- Asymptotic simultaneous confidence intervals applies χ^2 statistic,

$$\text{Multiplier} = \sqrt{\chi^2_p(\alpha)}$$

- Hotelling's T^2 induced simultaneous confidence intervals use F statistic,

$$\text{Multiplier} = \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}$$

These simultaneous confidence intervals are useful for small to moderate p. When p is in the hundreds or higher, these hyper-rectangles simultaneous confidence intervals are likely too large to be useful. Other criteria are sought and developed, under the subject of multiple testing (not covered in this course).

Note: Related chapter in Johnson and Wichern: Chapter 5.