Lecture 5

Inferences about a Mean Vector

Readings: Zelterman, 2015, Chapters 5, 6, 7

DSA 8070 Multivariate Analysis

September 19 - September 23, 2022

Whitney Huang
Clemson University

Agenda

1. Confidence Intervals/Region for Population Means
2. Hypothesis Testing for Mean Vector
3. Multivariate Paired Hotelling's T-Square

Overview

In this week we consider estimation and inference on population mean vector

We will explore the following questions:

- What is the sampling distribution of $\bar{\boldsymbol{X}}_n$?
- How to construct confidence intervals/region for population means
- How to conduct hypothesis testing for population means

Review: Sampling Distribution of Univariate Sample Mean $\bar{X}_n$

Suppose $X_1, X_2, \cdots, X_n$ is a random sample from a univariate population distibution with mean $\mathbb{E}(X) = \mu$ and variance $\mathbb{Var}(X) = \sigma^2$. The sample mean $\bar{X}_n$ is a function of random sample and therefore has a distribution

- $\bar{X}_n \dot{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n})$ when the sample size n is "sufficiently" large $\Rightarrow$ This is the central limit theorem (CLT)
- The result above is exact if the population follows a normal distribution, i.e., $X \sim \mathrm{N}(\mu, \sigma^2)$
- The standard error $\sqrt{\mathbb{Var}(\bar{X}_n)} = \frac{\sigma}{\sqrt{n}}$ provides a measure estimation precision. In practice, we use $\frac{s}{\sqrt{n}}$ instead where s is the sample standard deviation

Sampling Distribution of Multivariate Sample Mean Vector $\bar{X}_n$

Suppose $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ is a random sample from a multivariate population distibution with mean vector $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and covariance matrix = $\boldsymbol{\Sigma}$.

- $\bar{\boldsymbol{X}}_n \dot{\sim} \mathrm{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$ when the sample size n is "sufficiently" large ⇒ This is the multivariate version of CLT
- The result above is exact if the population follows a normal distribution, i.e., $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- Again, the estimation precision improves with a larger sample size. Like the univariate case we would need to replace $\boldsymbol{\Sigma}$ by its estimate S, the sample covariacne matrix

Review: Interval Estimation of Univariate Population Mean μ

The general format of a confidence interval (CI) estimate of a population mean is

Sample mean ± multiplier × standard error of mean.

For variable X, a CI estimate of its population mean μ is

$$\bar{X}_n \pm t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}},$$

Here the multiplier value is a function of the confidence level, α, the sample size n

Constructing Confidence Intervals for Mean Vector

We will still use the general recipe

Sample mean ± multiplier × standard error of mean.

The multiplier value also depends the strategy used for dealing with the multiple inference issue

- One at a Time CIs: a CI for μ_j is computed as

$$\bar{x}_j \pm t_{n-1}(\alpha/2)\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p$$

- Bonferroni Method: a CI for μ_j is computed as

$$\bar{x}_j \pm t_{n-1}(\alpha/2p)\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p$$

- Simultaneous CIs: a CI for μ_j is computed as

$$\bar{x}_j \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p$$

Example: Mineral Content Measurements [source: Penn Stat Univ. STAT 505]

This example uses the dataset that includes mineral content measurements at two different arm bone locations for $n = 64$ women. We'll determine confidence intervals for the two different population means. Sample means and standard deviations for the two variables are:

Variable	Sample size	Mean	Std Dev
domradius (X_1)	$n = 64$	$\bar{x}_1 = 0.8438$	$s_1 = 0.1140$
domhumerus (X_2)	$n = 64$	$\bar{x}_2 = 1.7927$	$s_2 = 0.2835$

Let's apply the three methods we learned to construct 95% CIs

Mineral Content Measurements Example Cont'd

- **One at a Time CIs**: $\bar{x}_j \pm t_{n-1}(\alpha/2)\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p.$
  Therefore 95% CIs for μ_1 and μ_2 are:

$$\mu_1: \quad 0.8438 \pm \underbrace{1.998}_{t_{63}(0.025)} \times \frac{0.1140}{\sqrt{64}} = [0.815, 0.872]$$

$$\mu_2: \quad 1.7927 \pm 1.998 \times \frac{0.2835}{\sqrt{64}} = [1.722, 1.864]$$

- **Bonferroni Method**: $\bar{x}_j \pm t_{n-1}(\alpha/2p)\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p.$

$$\mu_1: \quad 0.8438 \pm \underbrace{2.296}_{t_{63}(0.0125)} \times \frac{0.1140}{\sqrt{64}} = [0.811, 0.877]$$

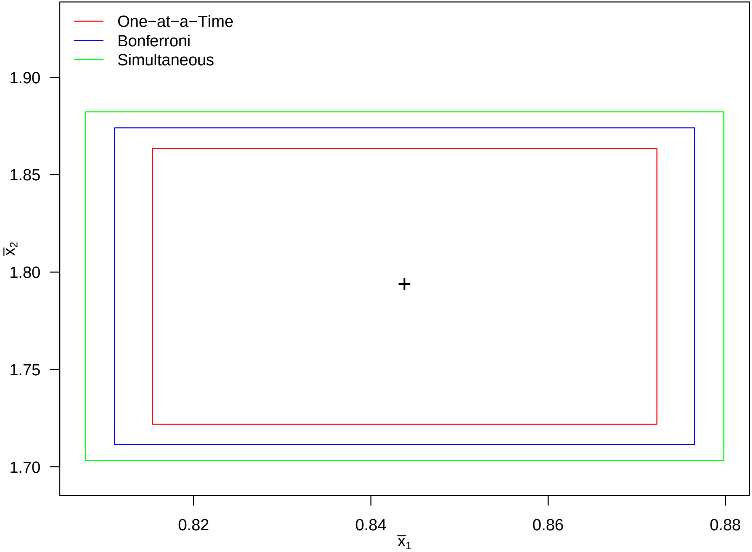
$$\mu_2: \quad 1.7927 \pm 2.296 \times \frac{0.2835}{\sqrt{64}} = [1.711, 1.874]$$

- **Simultaneous CIs**: $\bar{x}_j \pm \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)}\frac{s_j}{\sqrt{n}}, \quad j = 1, \cdots, p$

$$\mu_1: \quad 0.8438 \pm 2.528 \times \frac{0.1140}{\sqrt{64}} = [0.808, 0.880]$$

$$\mu_2: \quad 1.7927 \pm 2.528 \times \frac{0.2835}{\sqrt{64}} = [1.703, 1.882]$$

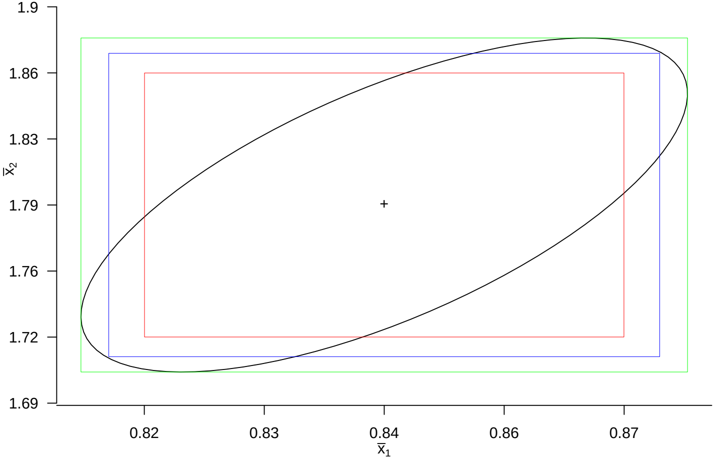
95 % CIs Based on Three Methods



Confidence Ellipsoid

A confidence ellipsoid for $\boldsymbol{\mu}$ is the set of $\boldsymbol{\mu}$ satisfying

$$n(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$$



Hypothesis Testing for Mean

- Recall: for univariate data, t statistic

$$t = \frac{\bar{X}_n - \mu_0}{s/\sqrt{n}} \Rightarrow t^2 = \frac{\left(\bar{X}_n - \mu_0\right)^2}{s^2/n} = n\left(\bar{X}_n - \mu_0\right)\left(s^2\right)^{-1}\left(\bar{X}_n - \mu_0\right)$$

Under $H_0 : \mu = \mu_0$

$$t \sim t_{n-1}, \quad t^2 \sim F_{1,n-1}$$

- Extending to multivariate by analogy:

$$T^2 = n\left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right)^T \boldsymbol{S}^{-1}\left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right)$$

Under $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

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

Note: T^2 here is the so-called Hotelling's T-Square

Hypothesis Testing for Mean Vector μ

1. State the null

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

   and the alternative

$$H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

2. Compute the test statistic

$$F = \frac{n-p}{(n-1)p} n \left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right)^T \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}_0\right)$$

3. **Compute the P-value.** Under H_0 : $\quad F \sim F_{p,n-p}$

4. **Draw a conclusion**: We do (or do not) have enough statistical evidence to conclude $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ at α significant level

Example: Women's Dietary Intake [source: Penn Stat Univ. STAT 505]

The recommended intake and a sample mean for all women between 25 and 50 years old are given below:

Variable	Recommended Intake ($\boldsymbol{\mu}_0$)	Sample Mean ($\bar{\boldsymbol{x}}_n$)
Calcium	1000 mg	624.0 mg
Iron	15 mg	11.1 mg
Protein	60 g	65.8 g
Vitamin A	800 μg	839.6 μg
Vitamin C	75 mg	78.9 mg

Here we would like to test, at $\alpha = 0.01$ level, if the $\boldsymbol{\mu} = \boldsymbol{\mu}_0$

Women's Dietary Intake Example Analysis

1. State the null

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$$

   and the alternative

$$H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

2. Compute the test statistic

$$F = \frac{n-p}{(n-1)p} n \left(\bar{\boldsymbol{x}}_n - \boldsymbol{\mu}_0\right)^T \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{x}}_n - \boldsymbol{\mu}_0\right) = 349.80$$

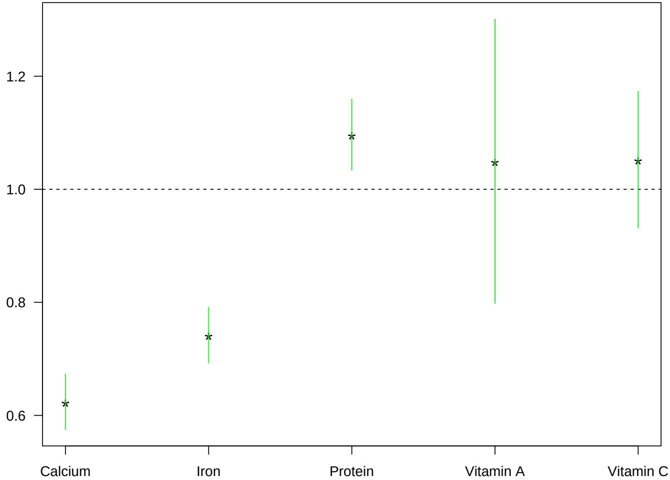
3. **Compute the P-value.** Under $H_0 : \quad F \sim F_{p,n-p} \Rightarrow$ p-value $= \mathbb{P}\mathrm{r}(F_{p,n-p} > 349.80) = 3 \times 10^{-191} < \alpha = 0.01$

4. **Draw a conclusion**: We do have enough statistical evidence to conclude $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ at α significant level

Profile Plots

1. Standardize each of the observations by dividing their hypothesized means

2. Plot either simultaneous or Bonferroni CIs for the population mean of these standardized variables



Spouse Survey Data Example

A sample ($n = 30$) of husband and wife pairs are asked to respond to each of the following questions:

1. What is the level of passionate love you feel for your partner?
2. What is the level of passionate love your partner feels for you?
3. What is the level of companionate love you feel for your partner?
4. What is the level of companionate love your partner feels for you?

Responses were recorded on a typical five-point scale: 1) None at all 2) Very little 3) Some 4) A great deal 5) Tremendous amount.

We will try to address the following question: Do the husbands respond to the questions in the same way as their wives?

Multivariate Paired Hotelling's T-Square

Let $\boldsymbol{X}_F$ and $\boldsymbol{X}_M$ be the responses to these 4 questions for females and males, respectively. Here the quantities of interest are $\mathbb{E}(\boldsymbol{D}) = \boldsymbol{\mu}_D$, the average differences across all husband and wife pairs.

1. State the null $H_0 : \boldsymbol{\mu}_D = 0$ and the alternative hypotheses $H_a : \boldsymbol{\mu}_D \neq \mathbf{0}$

2. Compute the test statistic

$$F = \frac{n-p}{(n-1)p} n \bar{\boldsymbol{D}}_n^T \boldsymbol{S}_{\boldsymbol{D}}^{-1} \bar{\boldsymbol{D}}_n$$

3. **Compute the P-value**. Under $H_0 : \quad F \sim F_{p,n-p}$

4. **Draw a conclusion**: We do (or do not) have enough statistical evidence to conclude $\boldsymbol{\mu}_D \neq \mathbf{0}$ at α significant level

Confidence Intervals/Region for Population Means
Hypothesis Testing for Mean Vector
Multivariate Paired Hotelling's T-Square

Spouse Survey Data Example Analysis

1. State the null

$$H_0 : \boldsymbol{\mu}_D = \mathbf{0}$$

and the alternative

$$H_a : \boldsymbol{\mu}_D \neq \mathbf{0}$$

2. Compute the test statistic

$$F = \frac{n-p}{(n-1)p} n \bar{\boldsymbol{D}}_n^T \boldsymbol{S}_{\boldsymbol{D}}^{-1} \bar{\boldsymbol{D}}_n = 2.942$$

3. **Compute the P-value.** Under $H_0 : \quad F \sim F_{p,n-p} \Rightarrow$ p-value $= \mathbb{P}\mathrm{r}(F_{p,n-p} >) = 0.0394 < \alpha = 0.05$

4. **Draw a conclusion**: We do have enough statistical evidence to conclude $\boldsymbol{\mu}_D \neq \mathbf{0}$ at 0.05 significant level