STAT 3690 Lecture 12

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Feb 18, 2022

$(1-\alpha)\times 100\%$ CR for $\boldsymbol{\nu}=\mathbf{A}\boldsymbol{\mu}$

- $\mathbf{X}_1,\ldots,\mathbf{X}_n \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$
  - Unknown $\boldsymbol{\Sigma}$
  - $n>p$
- $\mathbf{A}$ is of $q\times p$ and $\mathrm{rk}(\mathbf{A})=q$, i.e., $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top>0$
- Then iid $\mathbf{A}\mathbf{X}_i\sim MVN_q(\boldsymbol{\nu},\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$
- $(1-\alpha)\times 100\%$ CR for $\boldsymbol{\nu}$ is $\{\boldsymbol{\nu}:\frac{n(n-q)}{q(n-1)}(\mathbf{A}\bar{\boldsymbol{x}}-\boldsymbol{\nu})^\top(\mathbf{A}\mathbf{S}\mathbf{A}^\top)^{-1}(\mathbf{A}\bar{\boldsymbol{x}}-\boldsymbol{\nu})<F_{1-\alpha,q,n-q}\}$
- Special case: $\mathbf{A}=\boldsymbol{a}^\top\in\mathbb{R}^p$
  - $(1-\alpha)\times 100\%$ confidence interval (CI) for scalar $\nu=\boldsymbol{a}^\top\boldsymbol{\mu}$ is

$$\{\nu: n(\boldsymbol{a}^\top\bar{\boldsymbol{x}}-\nu)^2(\boldsymbol{a}^\top\mathbf{S}\boldsymbol{a})^{-1}<F_{1-\alpha,1,n-1}\}=\left(\boldsymbol{a}^\top\bar{\boldsymbol{x}}-t_{1-\alpha/2,n-1}\sqrt{\boldsymbol{a}^\top\mathbf{S}\boldsymbol{a}/n},\boldsymbol{a}^\top\bar{\boldsymbol{x}}+t_{1-\alpha/2,n-1}\sqrt{\boldsymbol{a}^\top\mathbf{S}\boldsymbol{a}/n}\right)$$

  - Check the coverage probability of CI for each entry of $\boldsymbol{\mu}$

```
options(digits = 4)
install.packages(c("MASS"))
set.seed(1)
B = 5e3L
n = 5e2L
Mu = (1:10)^2; (p = length(Mu))
(Sigma = diag(p)+.5)
alpha <- .05
(A = diag(p))
cover = matrix(0, ncol = p, nrow = B)
for (b in 1:B){
  sample = MASS::mvrnorm(n, Mu, Sigma)
  mu_hat = colMeans(sample)
  sample_cov = cov(sample)
  LB = A %*% mu_hat - qt(1-alpha/2, n-1)* sqrt(diag(A %*% sample_cov %*% t(A))/n)
  RB = A %*% mu_hat + qt(1-alpha/2, n-1)* sqrt(diag(A %*% sample_cov %*% t(A))/n)
  cover[b,] = (LB < Mu) * (Mu < RB)
}
(cover_prob_indiv = colMeans(cover))
(cover_prob_simul = mean(apply(cover, 1, prod)))
```

Simultaneous confidence intervals

- Interested in $(1-\alpha_k)$ CIs for scalars $\boldsymbol{a}_k^\top \boldsymbol{\mu}$, say CI_k, $k = 1, \ldots, m$, simultaneously
- Make sure $\Pr(\bigcap_k \{\boldsymbol{a}_k^\top \boldsymbol{\mu} \in \text{CI}_k\}) \geq 1 - \alpha$
- Bonferroni correction
  - Bonferroni inequality:

$$\Pr(\bigcap_{k=1}^m \{\boldsymbol{a}_k^\top \boldsymbol{\mu} \in \text{CI}_k\}) = 1 - \Pr(\bigcup_{k=1}^m \{\boldsymbol{a}_k^\top \boldsymbol{\mu} \notin \text{CI}_k\}) \geq 1 - \sum_{k=1}^m \Pr(\boldsymbol{a}_k^\top \boldsymbol{\mu} \notin \text{CI}_k) = 1 - \sum_{k=1}^m \alpha_k$$

  - Taking α_k such that $\alpha = \sum_{k=1}^m \alpha_k$, e.g., $\alpha_k = \alpha/m$, i.e.,

$$(\boldsymbol{a}_k^\top \bar{\boldsymbol{x}} - t_{1-\alpha/(2m),n-1}\sqrt{\boldsymbol{a}_k^\top \mathbf{S} \boldsymbol{a}_k/n}, \boldsymbol{a}_k^\top \bar{\boldsymbol{x}} + t_{1-\alpha/(2m),n-1}\sqrt{\boldsymbol{a}_k^\top \mathbf{S} \boldsymbol{a}_k/n})$$

  - Working for small m
- Scheffé's method
  - Let $\text{CI}_{\boldsymbol{a}} = (\boldsymbol{a}^\top \bar{\boldsymbol{x}} - c\sqrt{\boldsymbol{a}^\top \mathbf{S} \boldsymbol{a}/n}, \boldsymbol{a}^\top \bar{\boldsymbol{x}} + c\sqrt{\boldsymbol{a}^\top \mathbf{S} \boldsymbol{a}/n})$ for all $\boldsymbol{a} \in \mathbb{R}^p$
  - Derive that $c = \sqrt{p(n-1)(n-p)^{-1} F_{1-\alpha,p,n-p}}$
    - By Cauchy–Schwarz: $\{\boldsymbol{a}^\top(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\}^2 = [(\mathbf{S}^{1/2}\boldsymbol{a})^\top \{\mathbf{S}^{-1/2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\}]^2 \leq \{(\boldsymbol{a}^\top \mathbf{S} \boldsymbol{a})^\top/n\}\{n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\} \Rightarrow$

$$\begin{aligned}\Pr(\bigcap_{k=1}^m \{\boldsymbol{a}_k^\top \boldsymbol{\mu} \in \text{CI}_k\}) \geq \Pr(\bigcap_{\boldsymbol{a} \in \mathbb{R}^p} \{\boldsymbol{a}^\top \boldsymbol{\mu} \in \text{CI}_{\boldsymbol{a}}\}) &= 1 - \Pr(\bigcup_{\boldsymbol{a} \in \mathbb{R}^p} \{\boldsymbol{a}^\top \boldsymbol{\mu} \notin \text{CI}_{\boldsymbol{a}}\}) \\ &= 1 - \Pr(\bigcup_{\boldsymbol{a} \in \mathbb{R}^p} [\{\boldsymbol{a}^\top(\bar{\mathbf{X}} - \boldsymbol{\mu})\}^2/\{(\boldsymbol{a}^\top \mathbf{S} \boldsymbol{a})^\top/n\} > c^2]) \\ &\geq 1 - \Pr(\{n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) > c^2\})\end{aligned}$$

    - $\Pr(\{n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) > c^2\}) = \alpha \Rightarrow c = \sqrt{p(n-1)(n-p)^{-1} F_{1-\alpha,p,n-p}}$
  - Working for large even infinite m

```r
options(digits = 4)
install.packages(c("dslabs"))
library(dslabs)
data("gapminder")
dataset = gapminder[
  !is.na(gapminder$infant_mortality) &
    gapminder$year == 2012,
  c('infant_mortality', "life_expectancy")]
dataset = as.matrix(dataset)

n = nrow(dataset); p = ncol(dataset)

alpha <- .05
a1 = c(1,0); a2 = c(0,1)
A = rbind(a1, a2)
(mu_hat <- colMeans(dataset))
(sample_cov <- cov(dataset))

# Simultaneous CIs without correction
c = qt(1-alpha/2, n-1)
```

```
(NOcorrection <- cbind(
    A %*% mu_hat - c * sqrt(diag(A %*% sample_cov %*% t(A))/n),
    A %*% mu_hat + c * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))

# Simultaneous CIs with Bonferroni correction
m = nrow(A)
c = qt(1-alpha/2/m, n-1)
(Bonferroni <- cbind(
    A %*% mu_hat - c * sqrt(diag(A %*% sample_cov %*% t(A))/n),
    A %*% mu_hat + c * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))

# Simultaneous CIs with Scheffe correction
c = sqrt(p*(n-1)/(n-p) * qf(1-alpha, p, n-p))
(Scheffe <- cbind(
    A %*% mu_hat - c * sqrt(diag(A %*% sample_cov %*% t(A))/n),
    A %*% mu_hat + c * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))
```

- Report: CIs (21.82, 29.82) and (69.92, 72.70) cover the mean infant mortality and mean life expectancy, simultaneously, with probability at least 95%.

Comparing two multivariate means (J&W Sec. 6.3)

- Two independent samples of (potentially) different sizes from two distributions with equal covariance
  - $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
  - $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2} \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- Let $\bar{\mathbf{X}}_i$ and $\mathbf{S}_i$ be the sample mean and sample covariance for the ith sample
- Hypotheses $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ v.s. $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$
- Test statistic following LRT

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \{(n_1^{-1} + n_2^{-1})\mathbf{S}_{\text{pool}}\}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F(p, n_1+n_2-p-1)$$

  - $\mathbf{S}_{\text{pool}} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$
- Rejection region at level α

$$\left\{ x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2} : (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \{(n_1^{-1} + n_2^{-1})\mathbf{S}_{\text{pool}}\}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \geq \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{1-\alpha, p, n_1+n_2-p-1} \right\}$$

- p-value

$$1 - F_{F_{1-\alpha,p,n_1+n_2-p-1}} \left[\frac{n_1+n_2-p-1}{p(n_1+n_2-2)} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \{(n_1^{-1} + n_2^{-1})\mathbf{S}_{\text{pool}}\}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \right]$$

```
options(digits = 4)
install.packages(c("dslabs"))
library(dslabs)
data("gapminder")
```

```r
dataset1 = gapminder[
  !is.na(gapminder$infant_mortality) &
    gapminder$continent == "Africa" &
    gapminder$year == 2012,
  c('infant_mortality', "life_expectancy")]
dataset1 = as.matrix(dataset1)

dataset2 = gapminder[
  !is.na(gapminder$infant_mortality) &
    gapminder$continent == "Asia" &
    gapminder$year == 2012,
  c('infant_mortality', "life_expectancy")]
dataset2 = as.matrix(dataset2)

n1 <- nrow(dataset1); n2 <- nrow(dataset2); p <- ncol(dataset1)

(mu_hat1 <- colMeans(dataset1))
(mu_hat2 <- colMeans(dataset2))
(S1 <- cov(dataset1))
(S2 <- cov(dataset2))
S_pool <- ((n1 - 1)*S1 + (n2 - 1)*S2)/(n1+n2-2)

(lrt <- t(mu_hat1-mu_hat2) %*%
  solve((n1^-1 + n2^-1)*S_pool) %*%
  (mu_hat1-mu_hat2))

alpha <- .05
(cri.val <- (n1+n2-2)*p/(n1+n2-p-1)*qf(1-alpha, p, n1+n2-p-1))
lrt >= cri.val
(p.val = 1-pf((n1+n2-p-1)/(n1+n2-2)/p*lrt, p, n1+n2-p-1))
```

- Report: Testing hypotheses H_0: In 2012 Asia and Africa shared the identical mean value in both infant mortality and life expectancy v.s. H_1: otherwise, we carried on the LRT and obtained 87.65 as the value of test statistic. The corresponding p-value (resp. rejection region) was 4.952e-14 (resp. $[6.255, \infty)$). So, at the .05 level, there was a strong statistical evidence against H_0, i.e., we rejected H_0 and believed that in 2012 Asia and Africa didn't share the identical mean value in either infant mortality or life expectancy.