

# STAT 3690 Lecture Note

Week Six (Feb 13, 15, & 17, 2023)

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Feb/17 11:38:26

---

## Inference on $\mu$ under the normality assumption (con'd)

### Testing $\mu$ (J&W Sec. 5.2 & 5.3, con'd)

- Sample  $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ,  $n > p$ 
  - $\mathcal{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$
- $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$  v.s.  $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$
- Multivariate case (with unknown  $\boldsymbol{\Sigma}$ )
  - Name of approach: LRT
  - Test statistic:  $T(\mathcal{X}) = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) (\sim T^2(p, n-1) = \frac{(n-1)p}{n-p} F(p, n-p)$  under  $H_0$ )
  - Level  $\alpha$  rejection region (with respect to  $T(\mathcal{X})$ ):  $R_\alpha = \{T(\mathcal{X}) : \frac{n-p}{p(n-1)} T(\mathcal{X}) \geq F_{1-\alpha, p, n-p}\}$ , i.e., reject  $H_0$  if  $T(\mathcal{X}) \geq \frac{p(n-1)}{n-p} F_{1-\alpha, p, n-p}$ 
    - \*  $F_{1-\alpha, p, n-p}$ : the  $(1-\alpha)$ -quantile of  $F(p, n-p)$
  - $p$ -value:  $p(\mathcal{X}) = 1 - F_{F(p, n-p)}\{\frac{n-p}{p(n-1)} T(\mathcal{X})\}$ 
    - \*  $F_{F(p, n-p)}$ : the cdf of  $F(p, n-p)$

---

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

(mu_hat <- colMeans(dataset))

# Test mu = mu_0
mu_0 <- c(25, 50, 3)
n = nrow(dataset)
p = ncol(dataset)
(test.stat <- drop(
  n * t(mu_hat - mu_0) %*% solve(cov(dataset)) %*% (mu_hat - mu_0)
))
(crit.point = (n-1)*p/(n-p)*qf(.95, p, n-p))
```

```
test.stat >= cri.point
(p.val = 1-pf((n-p)/(n-1)/p*test.stat, p, n-p))
```

- Report: Testing hypotheses  $H_0 : \boldsymbol{\mu} = [25, 50, 3]^\top$  v.s.  $H_1 : \boldsymbol{\mu} \neq [25, 50, 3]^\top$ , we carried on the LRT and obtained 249718 as the value of test statistic with  $[7.819, \infty)$  as the corresponding level .05 rejection region. In addition, the  $p$ -value was almost 0. So, at the .05 level, there was a strong statistical evidence implying the rejection of  $H_0$ , i.e., we believed that the population mean vector was not  $[25, 50, 3]^\top$ .

### $(1 - \alpha) \times 100\%$ confidence region (CR) for $\boldsymbol{\mu}$ (J&W Sec. 5.4)

- $\Pr\{(1 - \alpha) \times 100\% \text{ CR covers } \boldsymbol{\mu}\} \geq 1 - \alpha$ 
  - CR is a set made of observations and is hence random
  - $\boldsymbol{\mu}$  is fixed
  - $(1 - \alpha) \times 100\%$  CR covers  $\boldsymbol{\mu}$  with probability at least  $(1 - \alpha) \times 100\%$
- Inverted from the level  $\alpha$  rejection region for  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$  v.s.  $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ . Specifically,
  1. Take the rejection region as a function of  $\boldsymbol{\mu}_0$ ;
  2. Replace  $\boldsymbol{\mu}_0$  with  $\boldsymbol{\mu}$ ;
  3. Take the complement.
- Eventually,  $(1 - \alpha) \times 100\%$  CR
  - $= \{\boldsymbol{\mu} : n(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) < \chi_{1-\alpha, p}^2\}$  if  $\boldsymbol{\Sigma}$  is known
  - $= \{\boldsymbol{\mu} : \frac{n(n-p)}{p(n-1)}(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) < F_{1-\alpha, p, n-p}\}$  if  $\boldsymbol{\Sigma}$  is not known

### Testing $\mathbf{A}\boldsymbol{\mu}$ (J&W pp. 279)

- $\mathbf{A}$  is of  $q \times p$  and  $\text{rk}(\mathbf{A}) = q$ , i.e.,  $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top > 0$
- Known: iid  $\mathbf{A}\mathbf{X}_i \sim \text{MVN}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ .
- LRT for  $H_0 : \mathbf{A}\boldsymbol{\mu} = \boldsymbol{\nu}_0$  v.s.  $H_1 : \mathbf{A}\boldsymbol{\mu} \neq \boldsymbol{\nu}_0$ 
  - Test statistic:  $T(\mathcal{X}) = n(\mathbf{A}\bar{\mathbf{X}} - \boldsymbol{\nu}_0)^\top (\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)^{-1}(\mathbf{A}\bar{\mathbf{X}} - \boldsymbol{\nu}_0) (\sim T^2(q, n-1) = \frac{(n-1)q}{n-q} F(q, n-q))$  under  $H_0$
  - Level  $\alpha$  rejection region (with respect to  $T(\mathcal{X})$ ):  $R_\alpha = \{T(\mathcal{X}) : \frac{n-q}{q(n-1)} T(\mathcal{X}) \geq F_{1-\alpha, q, n-q}\}$
  - $p$ -value:  $p(\mathcal{X}) = 1 - F_{F(q, n-q)}\{\frac{n-q}{q(n-1)} T(\mathcal{X})\}$
- Multiple comparison
  - Interested in  $H_0 : \mu_1 = \dots = \mu_p$  v.s.  $H_1 : \text{Not all entries of } \boldsymbol{\mu} \text{ are equal.}$ 
    - \*  $\mu_k$ : the  $k$ th entry of  $\boldsymbol{\mu}$
  - Take

$$\boldsymbol{\nu}_0 = \mathbf{0}_{(p-1) \times 1}, \quad \mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix}_{(p-1) \times p}.$$

- $p = 2$  (i.e.,  $\mathbf{A} = [1, -1]$ ): the case of A/B testing

---

```
options(digits = 4)
install.packages(c("dslabs", "tidyverse"))
library(dslabs)
library(tidyverse)
data("gapminder")
dataset = gapminder[
  !is.na(gapminder$infant_mortality) &
  gapminder$region == 'South America' &
```

```

gapminder$year %in% 2000:2008,
c('country', 'year', "life_expectancy")] %>%
spread(year, life_expectancy)
(dataset = as.matrix(dataset[, -1]))
n = nrow(dataset); p = ncol(dataset)
(mu_hat <- colMeans(dataset))

# Test H0:A %%% mu = nu_0
(nu_0 <- as.matrix(rep(0, p-1)))
(A = cbind(rep(1, p-1), -diag(p-1)))

(test.stat <- drop(
  n * t(A %%% mu_hat - nu_0) %%%
  solve(A %%% cov(dataset) %%% t(A)) %%%
  (A %%% mu_hat - nu_0)
))
(cri.point = (n-1)*(p-1)/(n-p+1)*qf(.95, p-1, n-p+1))
test.stat >= cri.point
(p.val = 1-pf((n-p+1)/(n-1)/(p-1)*test.stat, p-1, n-p+1))

```

- Report: Testing hypotheses  $H_0$ : the average life expectancy over south american countries doesn't vary with time v.s.  $H_1$ : otherwise, we carried on the LRT and obtained 628.5 as the value of test statistic and  $[132.9, \infty)$  as the corresponding level .05 rejection region. In addition, the  $p$ -value was .002858. So, at the .05 level, there was a strong statistical evidence against  $H_0$ , i.e., we believed that the average life expectancy over south american countries does vary with time.

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

- $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} \text{MVN}_p(\mu, \Sigma)$  with unknown  $\Sigma$  and  $n > p$
- $\mathbf{A}$  is of  $q \times p$  and  $\text{rk}(\mathbf{A}) = q$ , i.e.,  $\mathbf{A}\Sigma\mathbf{A}^\top > 0$
- Then iid  $\mathbf{A}\mathbf{X}_i \sim \text{MVN}_q(\nu, \mathbf{A}\Sigma\mathbf{A}^\top)$
- $(1 - \alpha) \times 100\%$  CR for  $\nu$  is  $\{\nu : \frac{n(n-q)}{q(n-1)}(\mathbf{A}\bar{\mathbf{x}} - \nu)^\top(\mathbf{A}\mathbf{S}\mathbf{A}^\top)^{-1}(\mathbf{A}\bar{\mathbf{x}} - \nu) < F_{1-\alpha, q, n-q}\}$
- Special case:  $\mathbf{A} = \mathbf{a}^\top \in \mathbb{R}^{1 \times p}$ , i.e.,  $\mathbf{A}$  is a row vector. Then
  - $(1 - \alpha) \times 100\%$  confidence interval (CI) for scalar  $\nu = \mathbf{a}^\top \mu$  is  $\{\nu : n(\mathbf{a}^\top \bar{\mathbf{x}} - \nu)^2 / (\mathbf{a}^\top \mathbf{S} \mathbf{a}) < F_{1-\alpha, 1, n-1}\}$ , i.e.,

$$\left( \mathbf{a}^\top \bar{\mathbf{x}} - t_{1-\alpha/2, n-1} \sqrt{\mathbf{a}^\top \mathbf{S} \mathbf{a} / n}, \quad \mathbf{a}^\top \bar{\mathbf{x}} + t_{1-\alpha/2, n-1} \sqrt{\mathbf{a}^\top \mathbf{S} \mathbf{a} / n} \right)$$

\* E.g., when  $\mathbf{A} = [1, 0, \dots, 0]$ , it is the CI for the first entry of  $\mu$ , say  $\mu_1$

- Checking the coverage probability of the previous CI for each  $\mu_k$

```

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  $(1 - \alpha_k) \times 100\%$  CIs for scalars  $\mathbf{a}_k^\top \boldsymbol{\mu}$ , say  $\text{CR}_k$ ,  $k = 1, \dots, m$ , simultaneously
- Make sure  $\Pr(\bigcap_k \{\mathbf{a}_k^\top \boldsymbol{\mu} \in \text{CR}_k\}) \geq 1 - \alpha$
- Bonferroni correction
  - Bonferroni inequality (optional) :

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

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

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

- Appropriate for small  $m$
- Scheffé's method
  - Let  $\text{CI}_{\mathbf{a}} = (\mathbf{a}^\top \bar{\mathbf{x}} - c \sqrt{\mathbf{a}^\top \mathbf{S} \mathbf{a} / n}, \mathbf{a}^\top \bar{\mathbf{x}} + c \sqrt{\mathbf{a}^\top \mathbf{S} \mathbf{a} / n})$  for all  $\mathbf{a} \in \mathbb{R}^p$ . Then we may find that  $c = \sqrt{p(n-1)(n-p)^{-1} F_{1-\alpha, p, n-p}}$ .
  - Derivation by Cauchy-Schwarz inequality (optional):  $\{\mathbf{a}^\top (\bar{\mathbf{x}} - \boldsymbol{\mu})\}^2 = [(\mathbf{S}^{1/2} \mathbf{a})^\top \{\mathbf{S}^{-1/2} (\bar{\mathbf{x}} - \boldsymbol{\mu})\}]^2 \leq \{(\mathbf{a}^\top \mathbf{S} \mathbf{a})^\top / n\} \{n(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})\} \Rightarrow$

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

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

---

```

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: After the Bonferroni correction, the resulting 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%.

**The confidence region for  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_p]^\top$  vs. simultaneous confidence intervals for  $\mu_1, \dots, \mu_p$**

- $\mathbf{X}_1, \dots, \mathbf{X}_n \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with known  $\boldsymbol{\Sigma}$  and  $n > p$
- $(1 - \alpha) \times 100\%$  CR for  $\boldsymbol{\mu}$ :  $\{\boldsymbol{\mu} : n(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) < \frac{p(n-1)}{n-p} F_{1-\alpha, p, n-p}\}$ 
  - CR covering  $\boldsymbol{\mu}$  with a probability at least  $1 - \alpha$
  - With a coverage probability closer to  $(1 - \alpha) \times 100\%$
- $(1 - \alpha) \times 100\%$  simultaneous  $\text{CI}_k$  for  $\mu_k$ :  $\left(\bar{x}_k - c\sqrt{S_{kk}/n}, \bar{x}_k + c\sqrt{S_{kk}/n}\right)$  with  $\bar{x}_k$  the  $k$ th entry of  $\bar{\mathbf{x}}$  and  $S_{kk}$  the  $(k, k)$ -th entry of  $\mathbf{S}$ 
  - $c = \sqrt{\frac{p(n-1)}{n-p} F_{1-\alpha, p, n-p}}$  (Scheffé) and  $t_{1-\alpha/(2p), n-1}$  (Bonferroni)
  - $\text{CI}_1 \times \dots \times \text{CI}_p$  covering  $\boldsymbol{\mu}$  with a probability at least  $1 - \alpha$
  - Clearly indicating the range for each  $\mu_k$

```

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) # entries of interest
A = rbind(a1, a2)
(mu_hat <- colMeans(dataset))
(sample_cov <- cov(dataset))
c = sqrt(p*(n-1)/(n-p) * qf(1-alpha, p, n-p))

# Plot the CR for the population mean vector mu
car::ellipse(center = mu_hat, shape = sample_cov/n, radius = c, add = F,
             xlab = "infant_mortality", ylab = "life_expectancy")

# Plot the simultaneous CIs with Scheffe correction
(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)
))
abline(v = Scheffe[1,1], col="red")
abline(v = Scheffe[1,2], col="red")
abline(h = Scheffe[2,1], col="red")
abline(h = Scheffe[2,2], col="red")

# Plot the simultaneous CIs with Bonferroni correction
(Bonferroni <- cbind(
  A %*% mu_hat - qt(1-alpha/2/nrow(A), n-1) * sqrt(diag(A %*% sample_cov %*% t(A))/n),
  A %*% mu_hat + qt(1-alpha/2/nrow(A), n-1) * sqrt(diag(A %*% sample_cov %*% t(A))/n)
))
abline(v = Bonferroni[1,1], col="green")
abline(v = Bonferroni[1,2], col="green")
abline(h = Bonferroni[2,1], col="green")
abline(h = Bonferroni[2,2], col="green")

```

## Comparing two population mean vectors (J&W Sec. 6.3)

- Two independent samples following two distributions with equal covariance

$$\begin{aligned}
 & - \mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \\
 & - \mathbf{X}_{21}, \dots, \mathbf{X}_{2n_2} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})
 \end{aligned}$$

- Let  $\bar{\mathbf{X}}_i$  and  $\mathbf{S}_i$  be the sample mean and sample covariance for the  $i$ th sample,  $i = 1, 2$ .
- Hypotheses  $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$  v.s.  $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

- Test statistic following LRT

$$T(\mathcal{X}) = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^\top \{(n_1^{-1} + n_2^{-1})\mathbf{S}_{\text{pool}}\}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \sim \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F(p, n_1 + n_2 - p - 1) \text{ under } H_0$$

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

- Level  $\alpha$  rejection region

$$\left\{ T(\mathcal{X}) : T(\mathcal{X}) \geq \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{1-\alpha, p, n_1 + n_2 - p - 1} \right\}$$

- $p$ -value

$$1 - F_{F_{1-\alpha, p, n_1 + n_2 - p - 1}} \left[ \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T(\mathcal{X}) \right]$$

---

```
options(digits = 4)
install.packages(c("dslabs"))
library(dslabs)
data("gapminder")
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
(crit.val <- (n1+n2-2)*p/(n1+n2-p-1)*qf(1-alpha, p, n1+n2-p-1))
lrt >= crit.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 and  $[6.255, \infty)$  as the corresponding rejection region. In addition, the  $p$ -value was

4.952e-14. 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 infant mortality and/or life expectancy.

---