

STAT 3690 Lecture Note

Part V: Comparisons of population mean vectors

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

2023/Feb/28 23:54:14

Comparisons of population mean vectors

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

- Two independent samples following two distributions with equal covariance

$$- \mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$- \mathbf{X}_{21}, \dots, \mathbf{X}_{2n_2} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

- 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

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

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

- Reject H_0 at level α when

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

- 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 level .05 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.

Comparing multiple population mean vectors (one-way multivariate analysis of variance (One-way MANOVA), J&W Sec. 6.4)

- Generalization of two-sample problem
 - Model: m independent samples, where
 - * $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
 - * \vdots
 - * $\mathbf{X}_{m1}, \dots, \mathbf{X}_{mn_m} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma})$
 - Hypotheses $H_0 : \boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_m$ v.s. H_1 : otherwise
- Alternatively
 - Model: m independent samples, where
 - * $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu} + \boldsymbol{\tau}_1, \boldsymbol{\Sigma})$
 - * \vdots
 - * $\mathbf{X}_{m1}, \dots, \mathbf{X}_{mn_m} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu} + \boldsymbol{\tau}_m, \boldsymbol{\Sigma})$
 - Identifiability: $\sum_i \boldsymbol{\tau}_i = \mathbf{0}$ otherwise there are infinitely many models that lead to the same data-generating mechanism.
 - Hypotheses $H_0 : \boldsymbol{\tau}_1 = \dots = \boldsymbol{\tau}_m = \mathbf{0}$ v.s. H_1 : otherwise
- Alternatively
 - Model: $\mathbf{X}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \mathbf{E}_{ij}$ with $\mathbf{E}_{ij} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\mathbf{0}, \boldsymbol{\Sigma})$
 - * Identifiability: $\sum_i \boldsymbol{\tau}_i = \mathbf{0}$
 - Hypotheses $H_0 : \boldsymbol{\tau}_1 = \dots = \boldsymbol{\tau}_m = \mathbf{0}$ v.s. H_1 : otherwise

-
- Sample means and sample covariances

- Sample mean for the i th sample $\bar{\mathbf{X}}_{i\cdot} = n_i^{-1} \sum_j \mathbf{X}_{ij}$
- Sample covariance for the i th sample $\mathbf{S}_i = (n_i - 1)^{-1} \sum_j (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})^\top$
- Grand mean $\bar{\mathbf{X}}_{..} = \sum_i n_i \bar{\mathbf{X}}_{i\cdot} / \sum_i n_i = \sum_{ij} \mathbf{X}_{ij} / \sum_i n_i$
- Decomposition of total (corrected) sum of squares and cross products matrix (SSP):

$$\mathbf{SSP}_t = \mathbf{SSP}_w + \mathbf{SSP}_b$$

- * Total (corrected) SSP: $\mathbf{SSP}_t = \sum_{ij} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{..})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{..})^\top = \mathbf{SSP}_w + \mathbf{SSP}_b$
 - * Within-group SSP: $\mathbf{SSP}_w = \sum_i (n_i - 1) \mathbf{S}_i = \sum_{ij} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})^\top$
 - * Between-group SSP: $\mathbf{SSP}_b = \sum_i n_i (\bar{\mathbf{X}}_{i\cdot} - \bar{\mathbf{X}}_{..})(\bar{\mathbf{X}}_{i\cdot} - \bar{\mathbf{X}}_{..})^\top$
-

- ML estimator of $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma})$

- Unconstrained
 - * $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{X}}_{i\cdot} = n_i^{-1} \sum_j \mathbf{X}_{ij}$
 - * $\hat{\boldsymbol{\Sigma}} = (\sum_i n_i)^{-1} \mathbf{SSP}_w$
- Under H_0
 - * $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{X}}_{..}$ for each i
 - * $\hat{\boldsymbol{\Sigma}} = (\sum_i n_i)^{-1} \mathbf{SSP}_t$

- Likelihood ratio

$$\lambda = \left\{ \frac{\det(\mathbf{SSP}_w)}{\det(\mathbf{SSP}_t)} \right\}^{\sum_i n_i / 2}$$

monotonic increasing with respect to the Wilk's lambda test statistic

$$\Lambda = \frac{\det(\mathbf{SSP}_w)}{\det(\mathbf{SSP}_t)} = \frac{\det(\mathbf{SSP}_w)}{\det(\mathbf{SSP}_w + \mathbf{SSP}_b)}$$

- $\Lambda \sim$ Wilk's lambda distribution $\Lambda(\boldsymbol{\Sigma}, \sum_i n_i - m, m - 1)$ under H_0
 - * Since $\mathbf{SSP}_w \sim W_p(\boldsymbol{\Sigma}, \sum_i n_i - m)$ and $\mathbf{SSP}_b \sim W_p(\boldsymbol{\Sigma}, m - 1)$ under H_0
 - * Bartlett's approximation (when $\sum_i n_i - m$ is large)

$$\{(p + m)/2 - \sum_i n_i + 1\} \ln \Lambda \approx \chi^2(p(m - 1))$$

- * Rao's approximation (default for R functions `manova` and `car::Manova`)

- Reject H_0 when Λ is small (equiv. \mathbf{SSP}_b is large); specifically, H_0 is rejected when

$$\{(p + m)/2 - \sum_i n_i + 1\} \ln \Lambda \geq \chi_{1-\alpha, p(m-1)}^2 \quad \text{OR} \quad \Lambda \leq \exp \left\{ \frac{\chi_{1-\alpha, p(m-1)}^2}{(p + m)/2 - \sum_i n_i + 1} \right\}$$

- p -value

$$1 - F_{\chi^2(p(m-1))} \left[\{(p + m)/2 - \sum_i n_i + 1\} \ln \Lambda \right]$$

- $F_{\chi^2(p(m-1))}$: the cdf of $\chi^2(p(m - 1))$
-

- Exercise 5.1: investigating the factors in manufacturing plastic film (see W. J. Krzanowski (1988) *Principles of Multivariate Analysis. A User's Perspective.* Oxford UP, pp. 381.)

- Three response variables (tear, gloss and opacity) describing measured characteristics of the resultant film
- A total of 20 runs
- One factor RATE (rate of extrusion, 2-level, low or high) in the production test

```
options(digits = 4)
install.packages('car')
tear <- c(
  6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1, 6.3,
  6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6
)
gloss <- c(
  9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9, 9.5, 9.4,
  9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7, 10.1, 9.2
)
opacity <- c(
  4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9, 1.9, 5.7,
  2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9, 2.7, 1.9
)
(X <- cbind(tear, gloss, opacity))
(rate <- factor(gl(2,10,length=nrow(X)), labels=c("Low", "High")))

# Bartlett's approximation to Wilks lambda distribution
X_low <- X[rate == 'Low',]
X_high <- X[rate == 'High',]
n <- nrow(X); p <- ncol(X); m <- length(levels(rate))
SSPcor = (n-1)*cov(X)
SSPw <- (nrow(X_low) - 1)*cov(X_low) + (nrow(X_high) - 1)*cov(X_high)
(Lambda <- det(SSPw)/det(SSPcor))
(cri.point = exp(qchisq(0.95, p*(m-1))/((p+m)/2-n+1)))
Lambda <= cri.point
(p.val = 1-pchisq(((p+m)/2-n+1)*log(Lambda), p*(m-1)))

# Rao's approximation to Wilks lambda distribution
summary(manova(X ~ rate), test = 'Wilks')
summary(car::Manova(lm(X ~ rate)), test.statistic='Wilks')
```

- Report: Testing hypotheses H_0 : no RATE effect on film characteristics v.s. H_1 : otherwise, we carried on the Wilk's lambda test and obtained 0.4136 as the value of test statistic and $(-\infty, 0.6227]$ as the corresponding level .05 rejection region. In addition, the p -value was 0.002227. So, at the .05 level, there was statistical evidence against H_0 , i.e., we rejected H_0 and believed that there was an effect from RATE on film characteristics.

Testing for equality of covariance matrices (J&W Sec. 6.6)

- Model: m independent samples, where
 - $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
 - \vdots
 - $\mathbf{X}_{m1}, \dots, \mathbf{X}_{mn_m} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$
- Hypotheses $H_0 : \boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_m$ v.s. H_1 : otherwise
- MLE of $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_m)$

- Under H_0
 - * $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{X}}_{i\cdot} = n_i^{-1} \sum_j \mathbf{X}_{ij}$
 - * $\hat{\boldsymbol{\Sigma}}_i = (\sum_i n_i)^{-1} \mathbf{SSP}_w = (\sum_i n_i)^{-1} \sum_{ij} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})^\top$ for all i
 - No restriction on $\boldsymbol{\Sigma}_i$
 - * $\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{X}}_{i\cdot} = n_i^{-1} \sum_j \mathbf{X}_{ij}$
 - * $\hat{\boldsymbol{\Sigma}}_i = n_i^{-1}(n_i - 1)\mathbf{S}_i = n_i^{-1} \sum_j (\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_{i\cdot})^\top$
 - Likelihood ratio
$$\lambda = \prod_i \left[\frac{\det\{n_i^{-1}(n_i - 1)\mathbf{S}_i\}}{\det\{(\sum_i n_i)^{-1}(\sum_i n_i - m)\mathbf{S}_{\text{pool}}\}} \right]^{n_i/2}$$
 - $\mathbf{S}_{\text{pool}} = (\sum_i n_i - m)^{-1} \mathbf{SSP}_w$
-

- Box's M test statistic (a modification of LRT)

$$M = -2 \ln \prod_i \left(\frac{\det \mathbf{S}_i}{\det \mathbf{S}_{\text{pool}}} \right)^{(n_i-1)/2}$$

- Under H_0

$$(1 - u)M \approx \chi^2(p(p+1)(m-1)/2)$$
 - * $u = \{\sum_i (n_i - 1)^{-1} - (\sum_i n_i - m)^{-1}\} \{6(p+1)(m-1)\}^{-1} (2p^2 + 3p - 1)$
 - Reject H_0 at level α when
$$(1 - u)M \geq \chi_{1-\alpha, p(p+1)(m-1)/2}^2$$
 - p -value
$$1 - F_{\chi_{1-\alpha, p(p+1)(m-1)/2}^2} \{(1 - u)M\}$$
-

- Exercise 5.2: factors in producing plastic film (continued)
 - Check the equality of covariance matrices for **RATE="Low"** and **RATE="High"**
 - Report: Testing hypotheses H_0 : the covariance matrix does not vary with the level of **RATE** v.s. H_1 : otherwise, we carried on the Box's M test and obtained 4.017 as the value of test statistic. The corresponding p -value was .6743. So, at the .05 level, there was no strong statistical evidence against H_0 , i.e., we did not reject H_0 and believed that the covariance matrix does not vary with the level of **RATE**.
-

Two-way MANOVA (J&W Sec. 6.7)

- Model: $\mathbf{X}_{ijk} = \boldsymbol{\mu} + \boldsymbol{\tau}_i + \boldsymbol{\beta}_j + \boldsymbol{\gamma}_{ij} + \mathbf{E}_{ijk}$ with $\mathbf{E}_{ijk} \stackrel{\text{iid}}{\sim} \text{MVN}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \dots, m$, $j = 1, \dots, b$, $k = 1, \dots, n$
 - $\boldsymbol{\tau}_i$: the main effect of factor 1 at level i
 - $\boldsymbol{\beta}_j$: the main effect of factor 2 at level j
 - $\boldsymbol{\gamma}_{ij}$: the interaction effect of factors 1 and 2 when their levels are i and j , respectively
 - Constraints for identifiability: $\sum_i \boldsymbol{\tau}_i = \sum_j \boldsymbol{\beta}_j = \sum_i \boldsymbol{\gamma}_{ij} = \sum_j \boldsymbol{\gamma}_{ij} = \mathbf{0}$
-

- Decomposition of total (corrected) SSP

$$\mathbf{SSP}_t = \mathbf{SSP}_{m1} + \mathbf{SSP}_{m2} + \mathbf{SSP}_{2fi} + \mathbf{SSP}_r$$

- Total (corrected) SSP

$$\mathbf{SSP}_t = \sum_{i=1}^m \sum_{j=1}^b \sum_{k=1}^n (\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{...})(\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{...})^\top$$

- * $\bar{\mathbf{X}}_{...} = (mbn)^{-1} \sum_{i,j,k} \mathbf{X}_{ijk}$
- SSP for main effect of factor 1

$$\mathbf{SSP}_{m1} = \sum_{i=1}^m bn(\bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{...})(\bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{...})^\top$$

- * $\bar{\mathbf{X}}_{i..} = (bn)^{-1} \sum_{j,k} \mathbf{X}_{ijk}$
- SSP for main effect of factor 2

$$\mathbf{SSP}_{m2} = \sum_{j=1}^b mn(\bar{\mathbf{X}}_{.j.} - \bar{\mathbf{X}}_{...})(\bar{\mathbf{X}}_{.j.} - \bar{\mathbf{X}}_{...})^\top$$

- * $\bar{\mathbf{X}}_{.j.} = (mn)^{-1} \sum_{i,k} \mathbf{X}_{ijk}$
- SSP for 2-factor-interaction (2fi)

$$\mathbf{SSP}_{2fi} = \sum_{i=1}^m \sum_{j=1}^b n(\bar{\mathbf{X}}_{ij.} - \bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{.j.} + \bar{\mathbf{X}}_{...})(\bar{\mathbf{X}}_{ij.} - \bar{\mathbf{X}}_{i..} - \bar{\mathbf{X}}_{.j.} + \bar{\mathbf{X}}_{...})^\top$$

- * $\bar{\mathbf{X}}_{ij.} = n^{-1} \sum_k \mathbf{X}_{ijk}$
- SSP for residual

$$\mathbf{SSP}_r = \sum_{i=1}^m \sum_{j=1}^b \sum_{k=1}^n (\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{ij.})(\mathbf{X}_{ijk} - \bar{\mathbf{X}}_{ij.})^\top$$

- Testing interaction
 - Hypotheses $H_0 : \gamma_{11} = \dots = \gamma_{mb} = \mathbf{0}$ v.s. H_1 : otherwise
 - Wilk's lambda test statistic

$$\Lambda = \frac{\det \mathbf{SSP}_r}{\det(\mathbf{SSP}_r + \mathbf{SSP}_{2fi})}$$

- * Under H_0 , by Bartlett's approximation

$$[\{p+1 - (m-1)(b-1)\}/2 - mb(n-1)] \ln \Lambda \approx \chi^2((m-1)(b-1))$$

- Reject H_0 at level α when

$$[\{p+1 - (m-1)(b-1)\}/2 - mb(n-1)] \ln \Lambda \geq \chi_{1-\alpha, (m-1)(b-1)}^2$$

- p -value

$$1 - F_{\chi^2((m-1)(b-1))}([\{p+1 - (m-1)(b-1)\}/2 - mb(n-1)] \ln \Lambda)$$

- Testing main effects
 - Testing factor 1 main effects
 - * Hypotheses $H_0 : \tau_1 = \dots = \tau_m = \mathbf{0}$ v.s. H_1 : otherwise
 - * Wilk's lambda test statistic

$$\Lambda = \frac{\det \mathbf{SSP}_r}{\det(\mathbf{SSP}_r + \mathbf{SSP}_{m1})}$$

- Under H_0 , by Bartlett's approximation

$$[\{p+1 - (m-1)\}/2 - mb(n-1)] \ln \Lambda \approx \chi^2(m-1)$$

- * Reject H_0 at level α when

$$[\{p+1 - (m-1)\}/2 - mb(n-1)] \ln \Lambda \geq \chi_{1-\alpha, m-1}^2$$

* p -value

$$1 - F_{\chi^2(m-1)}([\{p+1-(m-1)\}/2 - mb(n-1)] \ln \Lambda)$$

– Testing factor 2 main effects

* Hypotheses $H_0 : \beta_1 = \dots = \beta_b = \mathbf{0}$ v.s. H_1 : otherwise

* Wilk's lambda test statistic

$$\Lambda = \frac{\det \mathbf{SSP}_r}{\det(\mathbf{SSP}_r + \mathbf{SSP}_{2fi})}$$

· Under H_0 , by Bartlett's approximation

$$[\{p+1-(b-1)\}/2 - mb(n-1)] \ln \Lambda \approx \chi^2(b-1)$$

* Reject H_0 at level α when

$$[\{p+1-(b-1)\}/2 - mb(n-1)] \ln \Lambda \geq \chi^2_{1-\alpha, b-1}$$

* p -value

$$1 - F_{\chi^2(b-1)}([\{p+1-(b-1)\}/2 - mb(n-1)] \ln \Lambda)$$

• Exercise 5.3: factors in producing plastic film (continued)

– One more factor ADDITIVE (amount of an additive, 2-level, low or high) in the production test