STAT 3690 Lecture Note

Part V: Comparisons of population mean vectors

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

2023/Feb/26 23:06:49

Comparisons of population mean vectors

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

- Two independent samples following two distributions with equal covariance
  - $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1} \overset{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$
  - $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2} \overset{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$
- Let $\bar{\boldsymbol{X}}_i$ and $\mathbf{S}_i$ be the sample mean and sample covariance for the ith 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

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

  - $\mathbf{S}_{\text{pool}} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1+n_2-2}$
- Level α rejection region (w.r.t. $T(\mathcal{X})$)

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

- 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)
```

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

---

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

$$\mathbf{SSP}_\mathrm{t} = \mathbf{SSP}_\mathrm{w} + \mathbf{SSP}_\mathrm{b}$$

    * Total (corrected) SSP: $\mathbf{SSP}_\mathrm{t} = \sum_{ij}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}})^\top = \mathbf{SSP}_\mathrm{w} + \mathbf{SSP}_\mathrm{b}$
    * Within-group SSP: $\mathbf{SSP}_\mathrm{w} = \sum_i(n_i - 1)\mathbf{S}_i = \sum_{ij}(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^\top$
    * Between-group SSP: $\mathbf{SSP}_\mathrm{b} = \sum_i n_i(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})^\top$

- ML estimator of $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma})$
  - Unconstrained
    * $\hat{\boldsymbol{\mu}}_i = \bar{\boldsymbol{X}}_i = n_i^{-1}\sum_j \boldsymbol{X}_{ij}$
    * $\widehat{\boldsymbol{\Sigma}} = (\sum_i n_i)^{-1}\mathbf{SSP}_\mathrm{w}$
  - Under H_0
    * $\hat{\boldsymbol{\mu}}_i = \bar{\boldsymbol{X}}$ for each i
    * $\widehat{\boldsymbol{\Sigma}} = (\sum_i n_i)^{-1}\mathbf{SSP}_\mathrm{t}$
- Likelihood ratio

$$\lambda = \left\{\frac{\det(\mathbf{SSP}_\mathrm{w})}{\det(\mathbf{SSP}_\mathrm{t})}\right\}^{\sum_i n_i/2}$$

  monotonic with respect to the Wilk's lambda test statistic

$$\Lambda = \frac{\det(\mathbf{SSP}_\mathrm{w})}{\det(\mathbf{SSP}_\mathrm{t})} = \frac{\det(\mathbf{SSP}_\mathrm{w})}{\det(\mathbf{SSP}_\mathrm{w} + \mathbf{SSP}_\mathrm{b})}$$

  - $\Lambda \sim$ Wilk's lambda distribution $\Lambda(\boldsymbol{\Sigma}, \sum_i n_i - m, m-1)$ under H_0
    * Since $\mathbf{SSP}_\mathrm{w} \sim W_p(\boldsymbol{\Sigma}, \sum_i n_i - m)$ and $\mathbf{SSP}_\mathrm{b} \sim W_p(\boldsymbol{\Sigma}, m-1)$ under H_0
    * When $\sum_i n_i - m$ is large (i.e., $(p+m)/2 - \sum_i n_i + 1 \ll 0$), apply the Bartlett's approximation

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

- Level α rejection region (with respect to Λ)

$$\begin{aligned}&\Big\{\Lambda : \{(p+m)/2 - \sum_i n_i + 1\}\ln\Lambda \geq \chi^2_{1-\alpha, p(m-1)}\Big\}\\ &= \left\{\Lambda : \Lambda \leq \exp\left\{\frac{\chi^2_{1-\alpha,p(m-1)}}{(p+m)/2 - \sum_i n_i + 1}\right\}\right\}\end{aligned}$$

- 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: factors in producing 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
- 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.

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

- Model: $\boldsymbol{X}_{ijk} = \boldsymbol{\mu}+\boldsymbol{\tau}_i+\boldsymbol{\beta}_j+\boldsymbol{\gamma}_{ij}+\mathbf{E}_{ijk}$ with $\mathbf{E}_{ijk} \overset{\text{iid}}{\sim} \text{MVN}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, m$, $j = 1, \ldots, b$, $k = 1, \ldots, n$
  - $\boldsymbol{\tau}_i$: 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

$$\mathbf{SSP}_{\text{t}} = \mathbf{SSP}_{\text{m1}} + \mathbf{SSP}_{\text{m2}} + \mathbf{SSP}_{\text{2fi}} + \mathbf{SSP}_{\text{r}}$$

  - Total (corrected) SSP

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

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

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

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

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

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

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

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

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

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

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

    * 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

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

  - 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 $H_0 : \boldsymbol{\tau}_1 = \cdots = \boldsymbol{\tau}_m = \mathbf{0}$ v.s. H_1 : otherwise
    * Wilk's lambda test statistic

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

      · 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

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

    * p-value

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

  - Testing factor 2 main effects
    * Hypotheses $H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_b = \mathbf{0}$ v.s. H_1 : otherwise
    * Wilk's lambda test statistic

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

      · 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 5.2: factors in producing plastic film (continued)
  - One more factor `ADDITIVE` (amount of an additive, 2-level, low or high) in the production test

---

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

- Model: m independent samples, where
  - $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1} \overset{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
  - $\vdots$
  - $\boldsymbol{X}_{m1}, \ldots, \boldsymbol{X}_{mn_m} \overset{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$
- Hypotheses $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_m$ v.s. H_1 : otherwise
- MLE of $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m)$
  - Under H_0
    * $\hat{\boldsymbol{\mu}}_i = \bar{\boldsymbol{X}}_i = n_i^{-1} \sum_j \boldsymbol{X}_{ij}$
    * $\widehat{\boldsymbol{\Sigma}}_i = (\sum_i n_i)^{-1} \mathbf{SSP}_{\text{w}} = (\sum_i n_i)^{-1} \sum_{ij} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^\top$ for all i
  - No restriction on $\boldsymbol{\Sigma}_i$
    * $\hat{\boldsymbol{\mu}}_i = \bar{\boldsymbol{X}}_i = n_i^{-1} \sum_j \boldsymbol{X}_{ij}$
    * $\widehat{\boldsymbol{\Sigma}}_i = n_i^{-1}(n_i - 1)\mathbf{S}_i = n_i^{-1} \sum_j (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^\top$
- Likelihood ratio

$$\lambda = \prod_i \left[\frac{\det\{n_i^{-1}(n_i - 1)\mathbf{S}_i\}}{\det\{(\sum_i n_i)^{-1}(\sum_i n_i - m)\mathbf{S}_{\text{pool}}\}} \right]^{n_i/2}$$

  - $\mathbf{S}_{\text{pool}} = (\sum_i n_i - m)^{-1} \mathbf{SSP}_{\text{w}}$

- 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

$$(1 - u)M \approx \chi^2(p(p + 1)(m - 1)/2)$$

    * $u = \{\sum_i (n_i - 1)^{-1} - (\sum_i n_i - m)^{-1}\}\{6(p + 1)(m - 1)\}^{-1}(2p^2 + 3p - 1)$
  - Level α Rejection region (w.r.t. M)

$$\left\{ M : (1 - u)M \geq \chi^2_{1-\alpha, p(p+1)(m-1)/2} \right\}$$

  - p-value

$$1 - F_{\chi^2_{1-\alpha, p(p+1)(m-1)/2}}\{(1 - u)M\}$$

- Exercise: 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`.