L18: Two-sample problem

1. Populations, parameters, samples and statistics

   (1) Populations and parameters
   Two independent normal populations: $N(\mu_x, \Sigma)$ and $N(\mu_y, \Sigma)$.
   Two unknown mean vectors $\mu_x \in R^p$ and $\mu_y \in R^p$.
   One unknown common but unknown variance-covariance matrix $\Sigma \in R^{p\times p}$.
   We are interested in the comparison of μ_x and μ_y, i.e., the inference on $\mu_x - \mu_y$.

   (2) Samples
   $\mathbf{x}_i \in R^p$, $i = 1, .., n_1$, is a random sample from $N(\mu_x, \Sigma)$. $X' = (\mathbf{x}_1, .., \mathbf{x}_{n_1}) \in R^{p\times n_1}$.
   $X \in R^{n_1\times p}$ is a data matrix from $N(\mu_x, \Sigma)$.

$$X' \sim N_{p\times n_1}(\mu_x 1'_{n-1}, \Sigma, I_{n_1}) \Longleftrightarrow X \sim N_{n_1\times p}(1_{n_1}\mu'_x, I_n, \Sigma).$$

   $\mathbf{y}_i \in R^p$, $i = 1, .., n_2$, is a random sample from $N(\mu_y, \Sigma)$. $X' = (\mathbf{y}_1, .., \mathbf{y}_{n_2}) \in R^{p\times n_2}$.
   $Y \in R^{n_2\times p}$ is data matrix from $N(\mu_y, \Sigma)$.

$$Y' \sim N_{p\times n_2}(\mu_y 1'_{n_1}, \Sigma, I_{n_2}) \Longleftrightarrow Y \sim N_{n_2\times p}(1_{n_2}\mu'_y, I_{n_2}, \Sigma).$$

   Let $Z' = (X', Y') \in R^{p\times n}$ where $n = n_1 + n_2$. Then $Z \begin{pmatrix} X \\ Y \end{pmatrix} \in R^{n\times p}$ is the combined data matrix.
   With $D = \begin{pmatrix} 1_{n_1} & 00 & 1_{n_2} \end{pmatrix} \in R^{n\times 2}$,

$$Z' \sim N_{p\times n}\left((\mu_x, \mu_y)D', \Sigma, I_n\right) \Longleftrightarrow Z \sim N_{n\times p}\left(D(\mu_x, \mu_y)', I_n, \Sigma\right).$$

   (3) Basic statistics
   $\overline{\mathbf{x}} = \frac{\sum_{i=1}^{n_1} x_i}{n_1} = \frac{X'1_{n_1}}{n_1}$ is the mean of the sample from X.
   $\overline{\mathbf{y}} = \frac{\sum_{i=1}^{n_2} y_i}{n_2} = \frac{Y'1_{n_2}}{n_2}$ is the mean of the sample from Y.
   $(\overline{x}, \overline{y}) = (X', Y') \begin{pmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{pmatrix} \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix} = Z'D(D'D)^{-1}$.

   Matrix $\sum_{i=1}^{n_1}(x_i - \overline{x})(x_i - \overline{x})' = X'\left(I_{n_1} - \frac{1_{n_1}1'_{n_1}}{n_1}\right)X = \text{CSSCP}_x$ is from sample X.
   Matrix $\sum_{i=1}^{n_2}(y_i - \overline{y})(y_i - \overline{y})' = Y'\left(I_{n_2} - \frac{1_{n_2}1'_{n_2}}{n_2}\right)Y = \text{CSSCP}_y$ is from sample Y.

$$\begin{aligned} \text{Let CSSCP} &= \sum_{i=1}^{n_1}(x_i - \overline{x})(x_i - \overline{x})' + \sum_{j=1}^{n_2}(y_j - \overline{y})(y_j - \overline{y})' = \text{CSSCP}_x + \text{CSSCP}_y \\ &= (X', Y') \begin{pmatrix} I - \frac{11'}{n_1} & 0 \\ 0 & I - \frac{11'}{n_2} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = Z'[I_n - D(D'D)^{-1}D']Z. \end{aligned}$$

   (4) Sampling distributions
   $(\overline{x}, \overline{y}) \sim N_{p\times 2}\left((\mu_x, \mu_y), \Sigma, \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix}\right)$ and $\text{CSSCP} \sim W_{p\times p}(0, \Sigma, n-2)$ are independent.

   Proof Note that $Z' \sim N_{p\times n}((\mu_x, \mu_y)D', \Sigma, I_n)$.
   So $(\overline{x}, \overline{y}) = Z'D(D'D)^{-1} \sim N_{p\times 2}\left((\mu_x, \mu_y), \Sigma, \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix}\right)$.
   Thus $\overline{x} \sim N\left(\mu_x, \frac{\Sigma}{n_1}\right)$ and $\overline{y} \sim N\left(\mu_y, \frac{\Sigma}{n_2}\right)$ are independent.

   $\text{CSSCP} = Z'[I - D(D'D)^{-1}D']Z \sim W_{p\times p}(0, \Sigma, n-2)$.

   $Z'D(D'D)^{-1}$ and $Z'[I - D(D'D)^{-1}D']$ are indep. since $[D(D'D)^{-1}]'I_n[I - D(D'D)^{-1}D'] = 0$.
   Therefore $(\overline{x}, \overline{y})$ and CSSCP are independent.

2. Point estimators a variable with T^2-distribution

(1) Point estimators for μ_x, μ_y and Σ

With $\overline{x}$, $S_x = \frac{\text{CSSCP}_x}{n_1}$ and $S_{xu} = \frac{\text{CSSCP}_x}{n_1-1}$, from sample X only,
$\overline{x}$ is MLE for μ_x, S_x is MLE for Σ and S_{xu} is an UE for Σ.

With $\overline{y}$, $S_y = \frac{\text{CSSCP}_y}{n_2}$ and $S_{yu} = \frac{\text{CSSCP}_y}{n_2-1}$, from sample Y only,
$\overline{y}$ is MLE for μ_y, S_y is MLE for Σ and S_{yu} is an UE for Σ.

With $S = \frac{\text{CSSCP}}{n}$ and $S_u = \frac{\text{CSSCP}}{n-2}$, from the combined sample Z,

$$\begin{aligned} L(\mu_x, \mu_y, \Sigma) &= L_1(\mu_x, \Sigma) \cdot L_2(\mu_y, \Sigma) \le L_1(\overline{x}, \Sigma) \cdot L_2(\overline{y}, \Sigma) \\ &= L(\overline{x}, \overline{y}, \Sigma) = \tfrac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\tfrac{1}{2}(\Sigma^{-1/2}\,\text{CSSCP}\,\Sigma^{-1/2}\right] \\ &\le L(\overline{x}, \overline{y}, S) = \left(\tfrac{n}{2\pi e}\right)^{np/2} |\text{CSSCP}|^{-n/2}. \end{aligned}$$

Thus $\overline{x}$, $\overline{y}$ and S are MLEs for μ_x, μ_y and Σ; $\overline{x}$, $\overline{y}$, and S_u are UEs for μ_x, μ_y and Σ.

(2) Point estimator for $\mu_x - \mu_y$

$\overline{x} \sim N\left(\mu_x, \frac{\Sigma}{n_1}\right)$ and $\overline{y} \sim N\left(\mu_y, \frac{\Sigma}{n_2}\right)$ are independent.

So $\overline{x} - \overline{y} \sim N\left(\mu_x - \mu_y, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$. Thus $\overline{x} - \overline{y}$ is an UE for $\mu_x - \mu_y$.

(3) $[(\overline{x} - \overline{y}) - (\mu_x - \mu_y)]' \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_u\right)^{-1} [(\overline{x} - \overline{y}) - (\mu_x - \mu_y)] \sim T^2(p, n-2)$.

Proof Note that $\frac{(\overline{x}-\overline{y})-(\mu_x-\mu_y)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0, \Sigma)$ is independent to $\text{CSSCP} \sim W_{p\times p}(\Sigma, n-2)$.

Thus $\left[\frac{(\overline{x}-\overline{y})-(\mu_x-\mu_y)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right]' \left(\frac{\text{CSSCP}}{n-2}\right)^{-1} [\frac{(\overline{x}-\overline{y})-(\mu_x-\mu_y)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}] \sim T^2(p, n-2)$.

Therefore $[(\overline{x} - \overline{y}) - (\mu_x - \mu_y)]' \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_u\right)^{-1} [(\overline{x} - \overline{y}) - (\mu_x - \mu_y)] \sim T^2(p, n-2)$.

3. Computations for $\overline{x}$, CSSCP_x, $\overline{y}$ and CSSCP_y.

(1) Two sample data

In SAS a class variable with two values identifies each observation into two samples.
The class variable could be numeric variable of character variable.

```
data a;
   infile "D:\ex.txt";
   input y1 y2 y3 Sname @@;
```

```
data a;
   infile "D:\ex.txt";
   input y1 y2 y3 Sname $ @@;
```

(2) Proc corr for each samples

```
proc sort;
   by Name;
proc corr CSSCP nocorr;
    var y1 y2 y3;
    by Name;
    run;
```

L19: Two-sample T^2-tests

1. Two-sample T^2-tests

   (1) A T^2-tests
   With two populations and two samples. For $H_0 : \mu_x - \mu_v = v_0$ versus $H_a : \mu_x - \mu_y \neq v_0$, $T^2 = (\overline{x}-\overline{y}-v_0)' \left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_u\right]^{-1} (\overline{x}-\overline{y}-v_0)$ measures the squared distance between estimated $\mu_x - \mu_y$, $\overline{x}-\overline{y}$, and hypothesized $\mu_x - \mu_y$, v_0. It is reasonable to reject H_0 when T^2 is large.

   > $H_0 : \mu_x - \mu_y = v_0$ versus $H_a : \mu_x - \mu_y \neq v_0$
   > Test statistic: $T^2 = (\overline{x} - \overline{y} - v_0)' \left[\left(\frac{1}{n_1}+\frac{1}{n_2}\right) S_u\right]^{-1} (\overline{x} - \overline{y} - v_0)$
   > Reject H_0 if $T^2 > c$.

   (2) An α-level T^2-test
   In (1) replace c by $T^2_\alpha(p, n-2)$. Then $T^2|H_0 \sim T^2(p, n-2)$. So

   $$\begin{aligned} P(\text{Type I error}) &= P(\text{Rejecting } H_0 \mid H_0 \text{ is true}) = P(T^2 > T^2_\alpha(p, n-2) \mid H_0) \\ &= P(T^2(p, n-2) > T^2_\alpha(p, n-2)) = \alpha. \end{aligned}$$

   Thus the test is an α-level test.
   Comment: $T^2_\alpha(p, n-2) = \frac{(n-2)p}{n-p-1} F_\alpha(p, n-p-1)$.

   (3) P-value
   In (1) replace the rejection rule by a formula, $\boxed{\text{P-value: } P(T^2(p, n-2) > T^2_{ob})}$. Then

   $$\begin{aligned} \alpha\text{-level test rejects } H_0 &\iff T^2_{ob} > T^2_\alpha(p, n-2) \\ &\iff P(T^2(p, n-2) > T^2_{ob}) < P(T^2(p, n-2) > T^2_\alpha(p, n-2)) \\ &\iff P(T^2(p, n-2) > T^2_{0b}) < \alpha \\ &\iff P(T^2(p, n-2) > T^2_{ob}) \text{ is the observed significance level.} \end{aligned}$$

   Comment: $P(T^2(p, n-2) > T^2_{ob}) = P\left(\frac{(n-2)p}{n-p-1} F(p, n-p-1) > T^2_{ob}\right)$
   $= P\left(F(p, n-p-1) > \frac{n-p-1}{(n-2)p} T^2_{ob}\right) = P(F(p, n-p-1) > F_{ob})$.

2. Implementation by proc reg

   (1) Regression representing two populations
   In regression $\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $\epsilon \sim N(0, \Sigma)$, suppose x_1 and x_2 are two indicators for two populations, i.e.,

   $$x_1 = \begin{cases} 1 & \text{The obs is from Population 1} \\ 0 & \text{Otherwise} \end{cases} \quad \text{and } x_2 = \begin{cases} 1 & \text{The obs is from Population 2} \\ 0 & \text{Otherwise} \end{cases}.$$

   When $x_1 = 1$, $x_2 = 0$ and $\mathbf{y} \sim (\beta_1, \Sigma)$; When $x_2 = 1$, $x_1 = 0$ and $\mathbf{y} \sim N(\beta_2, \Sigma)$.
   So the regression represents two normal populations with possible different mean vectors and a common variance-covariance matrix.

   (2) Testing on $H_0 : \mu_x - \mu_y = v_0$
   The test on $H_0 : \mu_x - \mu_y = v_0$ becomes the test in the regression on $H_0 : \beta_1 - \beta_2 = v_0$.
   Here we consider the case for $v_0 = 0$.

```
data a;
  infile "D:\ex.txt";
  input y1 y2 Sname $ @@;
  if Sname="A-1" then do;
    x1=1; x2=0; end;
  if Sname="A-2" then do;
    x1=0; x2=1; end;
```

```
proc reg;
  model y1 y2=x1 x2/noint noprint;
  mtest x1-x2=0;
  run;
```

(3) SAS output

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.4964	8.12	2	16	0.0037
Pillai's Trace	0.5036	8.12	2	16	0.0037
Hotelling-Lawley Trace	1.0147	8.12	2	16	0.0037
Roy's Greatest Root	1.0147	8.12	2	16	0.0037

$T^2 = \left(\frac{1}{\Lambda} - 1\right)(n-2) = \left(\frac{1}{0.4964} - 1\right) \times 17 = 17.25$; $\Lambda + (\text{Pillai trace}) = 1$.
$T^2 = (H - L\ trace)(n-2) = 1.0147 \times 17 = 17.25$; $(H - L\ trace) = (Roy\ Root)$.
$P(T^2(2,17) > 17.25) = P\left(F(2,16) > \frac{16}{17\times 2} \times 17.25 = 8.12\right) = 0.0037$

(4) Report

$H_0 : \mu_x - \mu_y = 0$ versus $H_a : \mu_x - \mu_y \neq 0$
Test statistic: $T^2 = (\overline{x} - \overline{y})' \left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right) S_u\right]^{-1} (\overline{x} - \overline{y})$
p-value: $P(T^2(2, n-2) > T^2_{ob})$.

p-value: $P(T^2(2,17) > 17.25) = P(F(2,\ 16) > 8.12) = 0.0037$
Reject H_0 at the level 0.0038.

(5) Test $\mu_x - \mu_y = \begin{pmatrix} 2 \\ 10 \end{pmatrix}$

$\mu_x - \mu_y = \begin{pmatrix} 2 \\ 10 \end{pmatrix} \Longleftrightarrow \mu_x = \mu_y + \begin{pmatrix} 2 \\ 10 \end{pmatrix}$. So we can change the second population to the one with mean $\mu_y + \begin{pmatrix} 2 \\ 10 \end{pmatrix}$ and test the equal means in new setting.

```
data a;
  infile "D:\ex.txt";
  input y1 y2 Sname $ @@;
  if Sname="A-1" then do;
    x1=1; x2=0; end;
  if Sname="A-2" then do;
    x1=0; x2=1;
    y1=y1+2; y2=y2+100; end;
```

```
proc reg;
  model y1 y2=x1 x2/noint noprint;
  mtest x1-x2=0;
  run;
```

3. Implementation by proc anova

(1) SAS code

```
data a;
  infile "D:\ex.txt";
  input y1 y2 Sname $ @@;
proc anova;
  class Sname;
  model y1 y2=Sname/nouni;
  manova h=Sname;
  run;
```

(2) SAS output
See (3) in 2.

(3) How to test $\mu_x - \mu_2 = v_0$?