L17 Two-sample tests

1. Three matrices

   (1) Model error matrix E
   In the two-sample problem μ_x and μ_y are estimated by $\overline{X}$ and $\overline{Y}$. Thus

$$\text{CSSCP}_x + \text{CSSCP}_y = \sum_{i=1}^{n_1}(X_i - \overline{X})(X_i - \overline{X})' + \sum_{j=1}^{n_2}(Y_j - \overline{Y})(Y_j - \overline{Y})'$$

   measures the model error and is denoted by E.

   (2) Reduced model error matrix E_0
   Under $H_0: \mu_x - \mu_y = \delta_0$, $\mu_x = \mu_y + \delta_0$. Thus $(X - \delta_0 1'_{n_1}, Y) \in R^{p\times n}$ is a r. s. from $N(\mu_y, \Sigma)$ and μ_y is estimated by $\widehat{\mu}_y = \frac{(\overline{X}-\delta_0)n_1}{n} + \frac{\overline{Y}n_2}{n} = \overline{Y} + \frac{n_1}{n}(\overline{X} - \overline{Y} - \delta_0)$. Let $h = \overline{X} - \overline{Y} - \delta_0$. Then

$$\begin{cases} \widehat{\mu}_y &= \overline{Y} + \frac{n_1}{n}h \\ \widehat{\mu}_x &= \widehat{\mu}_y + \delta_0 = \overline{Y} + \frac{n-n_2}{n}(\overline{X} - \overline{Y} - \delta_0) + \delta_0 = \overline{X} - \frac{n_2}{n}h. \end{cases}$$

   Thus for the model reduced by H_0, the error matrix is

$$\begin{aligned} E_0 &= \textstyle\sum_{i=1}^{n_1}(X_i - \widehat{\mu}_x)(X_i - \widehat{\mu}_x)' + \sum_{j=1}^{n_2}(Y_j - \widehat{\mu}_y)(Y_j - \widehat{\mu}_y)' \\ &= \textstyle\sum_{i=1}^{n_1}\left[(X_i - \overline{X}) + \frac{n_2}{n}h\right]\left[(X_i - \overline{X}) + \frac{n_2}{n}h\right]' + \sum_{j=1}^{n_2}\left[(Y_j - \overline{Y}) - \frac{n_1}{n}h\right]\left[(Y_j - \overline{Y}) - \frac{n_1}{n}h\right]' \\ &= E + \frac{n_1 n_2^2}{n^2}hh' + \frac{n_2 n_1^2}{n^2}hh' = E + \frac{n_1 n_2}{n}hh'. \end{aligned}$$

   (3) Matrix H
   The difference between E and E_0 is caused by the hypothesis $H_0: \mu_x - \mu_y = \delta_0$.
   So we write $E_0 = E + H$ where

$$H = \frac{n_1 n_2}{n}(\overline{X} - \overline{Y} - \delta_0)(\overline{X} - \overline{Y} - \delta_0)'.$$

2. Likelihood ratio tests

   (1) Likelihood ratio
   It has been shown the $\max[L(\mu_x, \mu_y, \Sigma): \mu_x, \mu_y, \Sigma] = \left(\frac{n}{2\pi e}\right)^{np/2}|E|^{-n/2}$.
   So under $H_0: \mu_x - \mu_y = \delta_0$, $\max[L(\mu_x, \mu_y, \Sigma): H_0] = \left(\frac{n}{2\pi e}\right)^{np/2}|E_0|^{-n/2}$.
   Thus the likelihood ratio

$$LR = \frac{\max[L(\mu_x, \mu_y, \Sigma): H_0]}{\max[L(\mu_x, \mu_y, \Sigma): \mu_x, \mu_y, \Sigma]} = \left(\frac{|E|}{|E_0|}\right)^{n/2}$$

   is an increasing function of $\Lambda = \frac{|E|}{|E_0|}$ called Wilks Lambda.

   (2) Likelihood ratio tests
   By intuition one would reject H_0 when LR is small, equivalently when Λ is small. Therefore, the followings are likelihood ratio tests.

$H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \neq \delta_0$ Test statistic: LR= $\left(\frac{\lvert E\rvert}{\lvert E_0\rvert}\right)^{n/2}$ Reject H_0 if LR $< c_1$.

$H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \neq \delta_0$ Test statistic: $\Lambda = \frac{\lvert E\rvert}{\lvert E_0\rvert}$ Reject H_0 if $\Lambda < c_2$.

(3) Comments

To make the above tests α-level tests, c_1 and c_2 must be selected such that

$P(LR < c_1|H_0) \le \alpha$ and $P(\Lambda < \Lambda_{ob}|H_0) \le \alpha$.

For doing so we have to know the distributions of the test statistics under H_0, called the null distributions.

3. α-level LRT

(1) T^2 with known null distribution

For $H_0: \mu_x - \mu_y = \delta_0$, let $T^2 = (\overline{X} - \overline{Y} - \delta_0)' \left(\frac{n}{n_1 n_2} S_p\right)^{-1} (\overline{X} - \overline{Y} - \delta_0)$.

By 1 (2) in L16, under H_0, $T^2 \sim T^2(p, n-2)$.

(2) Λ is a decreasing function of T^2

Note that $\begin{vmatrix} 1 & -\frac{n_1 n_2}{n} h' \\ h & E \end{vmatrix} = 1 \cdot \left|E + \frac{n_1 n_2}{n} hh'\right| = |E + H| = |E_0|$. But we also

$$\begin{vmatrix} 1 & -\frac{n_1 n_2}{n} h' \\ h & E \end{vmatrix} = |E| \cdot \left(1 + \frac{n_1 n_2}{n} h' E^{-1} h\right) = |E| \left(1 + \frac{h'\left(\frac{n}{n_1 n_2} S_p\right)^{-1} h}{n-2}\right) = |E| \left(1 + \frac{T^2}{n-2}\right).$$

Thus $|E + H| = |E| \left(1 + \frac{T^2}{n-2}\right)$.

So $\Lambda = \frac{|E|}{|E+H|} = \left(1 + \frac{T^2}{n-2}\right)^{-1} \Longleftrightarrow T^2 = \left(\frac{1}{\Lambda} - 1\right)(n-2)$

are decreasing functions each other. Thus T^2 can be used as a LRT statistic.

(3) α-level LRT

> $H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \ne \delta_0$
> Test statistic: $T^2 = (\overline{X} - \overline{Y} - \delta_0)' \left(\frac{n}{n_1 n_2} S_p\right)^{-1} (\overline{X} - \overline{Y} - \delta_0)$
> Reject H_0 if $T^2 > T^2_\alpha(p, n-2)$.

If Λ_{ob} is given, then $T^2_{ob} = \left(\frac{1}{\Lambda_{ob}} - 1\right)(n-2)$.

Since under H_0, $T^2 \sim T^2(p, n-2) = \frac{(n-2)p}{n-p-1} F(p, n-p-1)$,

$$T^2_\alpha(p, n-2) = \frac{(n-2)p}{n-p-1} F_\alpha(p, n-p-1).$$

(4) Test by p-value

> $H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \ne \delta_0$
> Test statistic: $T^2 = (\overline{X} - \overline{Y} - \delta_0)' \left(\frac{n}{n_1 n_2} S_p\right)^{-1} (\overline{X} - \overline{Y} - \delta_0)$
> p-value: $P(T^2(p, n-2) > T^2_{ob})$.

Since under H_0, $T^2 \sim T^2(p, n-2) = \frac{(n-2)p}{n-p-1} F(p, n-p-1)$,

$$F_{ob} = \frac{n-p-1}{(n-2)p} T^2_{ob} \text{ and } P(T^2(p, n-2) > T^2_{ob}) = P(F(p, n-p-1) > F_{ob}).$$

L18 Two-sample test implementation

1. Two-sample tests
   $N(\mu_x, \Sigma)$ and $N(\mu_y, \Sigma)$ are two populations.

   (1) Testing on $\mu_x - \mu_y \in R^p$

   $H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \neq \delta_0$
   Test statistic: $T^2 = (\overline{X} - \overline{Y} - \delta_0)' \left(\frac{n}{n_1 n_2} S_p\right)^{-1} (\overline{X} - \overline{Y} - \delta_0)$
   Reject H_0 if $T^2 > T^2_\alpha(p, n-2)$.

   $H_0: \mu_x - \mu_y = \delta_0$ vs $H_a: \mu_x - \mu_y \neq \delta_0$
   Test statistic: $T^2 = (\overline{X} - \overline{Y} - \delta_0)' \left(\frac{n}{n_1 n_2} S_p\right)^{-1} (\overline{X} - \overline{Y} - \delta_0)$
   p-value: $P(T^2(p, n-2) > T^2_{ob})$.

   (2) Testing on $L(\mu_x - \mu_y) \in R^q$
   Transformed populations $LN(\mu_x, \Sigma) = N(L\mu_x, L\Sigma L')$ and $LN(\mu_y, \Sigma) = N(L\mu_y, L\Sigma L')$ have transformed samples $L(X, Y) = (LX, LY) \in R^{q\times n}$ with means $L\overline{X}$ and $L\overline{Y}$; and pooled estimator for $L\Sigma L'$, LS_pL'. So

   $H_0: L(\mu_x - \mu_y) = \delta_0$ vs $H_a: L(\mu_x - \mu_y) \neq \delta_0$
   Test statistic: $T^2 = [L(\overline{X} - \overline{Y}) - \delta_0]' \left(\frac{n}{n_1 n_2} LS_pL'\right)^{-1} [L(\overline{X} - \overline{Y}) - \delta_0]$
   Reject H_0 if $T^2 > T^2_\alpha(q, n-2)$.

   $H_0: L(\mu_x - \mu_y) = \delta_0$ vs $H_a: L(\mu_x - \mu_y) \neq \delta_0$
   Test statistic: $T^2 = [L(\overline{X} - \overline{Y}) - \delta_0]' \left(\frac{n}{n_1 n_2} LS_pL'\right)^{-1} [L(\overline{X} - \overline{Y}) - \delta_0]$
   p-value: $P(T^2(q, n-2) > T^2_{ob})$.

2. Data modification

   (1) proc anova
   SAS procedure proc anova with entered two samples from $N(\mu_x, \Sigma)$ and $N(\mu_y, \Sigma)$ can produce information on the testing on $H_0: \mu_x = \mu_y$.

   (2) Data modification for $H_0: \mu_x - \mu_y = \delta_0$
   Note that $H_0: \mu_x - \mu_y = \delta_0 \Longleftrightarrow \mu_x = \mu_y + \delta_0$.
   Thus we can keep the sample from $N(\mu_x, \Sigma)$, but modify the sample from $N(\mu_y, \Sigma)$ to that from $N(\mu_y + \delta_0, \Sigma)$.

   Ex1: Suppose file ex.txt contains four variables x1, x2, x3 and sname= $\begin{cases} \text{AB} & \text{First sample} \\ \text{CD} & \text{Second sample} \end{cases}$.

   For $H_0: \mu_x - \mu_y = \begin{pmatrix} -1 \\ 2 \\ -3 \end{pmatrix}$.

```
data a;
   infile "D\ex.txt";
   input x1 x2 x3 sname $ @@;
   if sname='CD' then do;
       x1=x1-1;
       x2=x2+2;
       x3=x3-3;
   end;
```

(3) Data modification for $H_0: L(\mu_x - \mu_y) = \delta_0$
First transform the two samples from $N(\mu_x, \Sigma)$ and $N(\mu_y, \Sigma)$ to that from $LN(\mu_x, \Sigma)$ and $LN(\mu_y, \Sigma)$. Then according to $H_0 : L(\mu_x - \mu_y) = \delta_0 \iff L\mu_x = L\mu_y + \delta_0$ modify the second transformed sample accordingly.

Ex2: For $H_0 : \begin{pmatrix}\mu_{x1} - \mu_{x_2} \\ \mu_{x2} + \mu_{x3}\end{pmatrix} - \begin{pmatrix}\mu_{y1} - \mu_{y2} \\ \mu_{y2} + \mu_{y3}\end{pmatrix} = \begin{pmatrix}1 \\ 2\end{pmatrix} \iff L\mu_x = L\mu_y + \delta_0$ where $L = \begin{pmatrix}1 & -1 & 0 \\ 0 & 1 & 1\end{pmatrix}$ and $\delta_0 = \begin{pmatrix}1 \\ 2\end{pmatrix}$,

```
data a;
   infile "D\ex.txt";
   input x1 x2 x3 sname $ @@;
   y1=x1-x2;
   y2=x2+x3;
   if sname="CD" then do;
   y1=y1+1;
   y2=y2+2;
```

3. Use SAS

(1) SAS procedure and output

```
proc anova;
     class sname;
     model x1 x2 x3=sname/nouni;
     manova h=sname;
     run;
```

Statistics	Value	F-value	Num DF	Den DF	Pr>F
Wilks' Lambda	0.7200	0.19	2	1	0.8485
Pillai's Trace	0.2800	0.19	2	1	0.8485
Hotellig-Lawley Trace	0.3889	0.19	2	1	0.8485
Roy's Greatest Root	0.3889	0.19	2	1	0.8485

(2) T^2_{ob} and p-value
Recall: $T^2 = \left(\frac{1}{\Lambda} - 1\right)(n-2)$. Λ is displayed in the output.
p-value: $P(T^2(q, n-2) > T^2_{ob}) = P\left(F(q, n-q-1) > \frac{n-q-1}{(n-2)q} T^2_{ob}\right) = P(F(q, n-q-1) > F_{ob})$.
SAS displays Numerator DF, Denominator DF, F_{ob} and p-value.

(3) Four statistics
The same information can be derived from other three statistics because of the relation of them in two-sample case.

Let $E^{-1/2}HE^{-1/2} = Q\Gamma Q'$ be EVD where $\Gamma = \mathrm{diag}(\gamma_1, .., \gamma_p)$ with $\gamma_1 \geq \cdots \geq \gamma_p > 0$. Then

$$\begin{aligned}
&\text{R-root} && \overset{def}{=\!=} \gamma_1 \\
&\text{H-L-trace} && \overset{def}{=\!=} \mathrm{tr}\left(HE^{-1}\right) = \mathrm{tr}\left(E^{-1/2}HE^{-1/2}\right) = \gamma_1 + \cdots + \gamma_p \overset{*}{=\!=} \gamma_1 \\
& && \quad \text{since } \mathrm{rank}\left(HE^{-1}\right) = \mathrm{rank}(H) = \mathrm{rank}\left(\tfrac{n_1 n_2}{n} hh'\right) = \mathrm{tr}(\overline{X} - \overline{Y} - \delta_0) = 1 \\
&\text{P-trace} && \overset{def}{=\!=} \mathrm{tr}[H(E+H)^{-1}] = \cdots = \tfrac{\gamma_1}{1+\gamma_1} + \cdots + \tfrac{\gamma_p}{1+\gamma_p} = \tfrac{\gamma_1}{1+\gamma_1} \\
&\Lambda && \overset{def}{=\!=} \tfrac{|E|}{|E+H|} = |E^{1/2}|\,|(E+H)^{-1}|\,|E^{-1/2}| = |E^{1/2}(E+H)^{-1}E^{1/2}| \\
& && = |(I - E^{-1/2}HE^{-1/2})^{-1}| = |[Q(I+\Gamma)Q']^{-1}| = |(I+\Gamma)^{-1}| \\
& && = \tfrac{1}{1+\gamma_1} \cdots \tfrac{1}{1+\gamma_p} = \tfrac{1}{1+\gamma_1}.
\end{aligned}$$