L15 Two-sample estimations

1. Sampling distributions

   (1) Populations, samples and basic statistics

$$
\begin{array}{ll}
\text{Populations:} & N(\mu_x,\,\Sigma) \text{ and } N(\mu_y,\,\Sigma). \\
\text{Samples:} & (X,\,Y) \sim N_{p\times n}(\mu J',\,\Sigma,\,I_n) \text{ where } n = n_1+n_2,\ \mu = (\mu_x,\,\mu_y) \in R^{p\times 2} \\
 & J = \begin{pmatrix} 1_{n_1} & 0 \\ 0 & 1_{n_2} \end{pmatrix} \in R^{n\times 2} \text{ so } \mu J' = (\mu_x 1'_{n_1},\,\mu_y 1'_{n_2}). \\
\text{Sample sizes:} & n = n_1 + n_2 \\
\text{Sample means:} & (\overline{X},\,\overline{Y}) = (X,\,Y)\begin{pmatrix} 1_{n_1}/n_1 & 0 \\ 0 & 1_{n_2}/n_2 \end{pmatrix} = (X,\,Y)J(J'J)^{-1} \\
\text{CSSCP} & \text{CSSCP}_x + \text{CSCP}_y = X(I_{n_1} - 11^+)X' + Y(I_{n_2} - 11^+)Y' \\
 & = (X,\,Y)(I_n - JJ^+)(X,\,Y)'.
\end{array}
$$

   (2) Sampling distribution: $\begin{pmatrix}\overline{X}\\ \overline{Y}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\,\begin{pmatrix}\frac{\Sigma}{n_1} & 0\\ 0 & \frac{\Sigma}{n_2}\end{pmatrix}\right)$

   Proof. Recall: $X \sim N_{p\times n}(M,\,\Sigma,\,\Psi)$ and $A \in R^{n\times m} \Longrightarrow XA \sim N_{p\times m}(MA,\,\Sigma,\,A'\Psi A)$.
   With $(X,\,Y) \sim N_{p\times n}(\mu J',\,\Sigma,\,I_n)$, $J(J'J)^{-1} \in R^{n\times 2}$,

$$
\begin{aligned}
(\overline{X},\,\overline{Y}) = (X,\,Y)J(J'J)^{-1} &\sim N_{p\times 2}\left(\mu J'J(J'J)^{-1},\,\Sigma,\,(J'J)^{-1}J'J(J'J)^{-1}\right) \\
&= N_{p\times 2}\left(\mu,\,\Sigma,\,\begin{pmatrix}\frac{1}{n_1} & 0\\ 0 & \frac{1}{n_2}\end{pmatrix}\right).
\end{aligned}
$$

   So $\begin{pmatrix}\overline{X}\\ \overline{Y}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\,\begin{pmatrix}\frac{\Sigma}{n_1} & 0\\ 0 & \frac{\Sigma}{n_2}\end{pmatrix}\right)$.

   (3) Sampling distribution: CSSCP $\sim W_{p\times p}(n-2,\,\Sigma)$.

   Proof. Recall: $X \sim N_{p\times n}(M,\,\Sigma,\,I), A' = A^2 = A$ has rank $r \Longrightarrow XAX' \sim W_{p\times p}(MAM',\,r,\Sigma)$.
   With $(X,\,Y) \sim N_{p\times n}(\mu J',\,\Sigma,\,I_n)$ and CSSCP $= (X,\,Y)(I_n - JJ^+)(X,\,Y)'$, where $I - JJ^+$ is symmetric and idempotent with $\text{rank}(I - JJ^+) = \text{tr}(I_n - JJ^+) = n-2$. Note that $\mu J'(I - JJ^+)(\mu J')' = 0$. So ,

$$
\text{CSSCP} = (X,\,Y)(I - JJ^+)(X,\,Y) \sim W_{p\times p}(0,\,n-1,\,\Sigma) = W_{p\times p}(n-2,\,\Sigma).
$$

   Ex1: $(\overline{X},\,\overline{Y})$ and CSSCP are independent.
   Recall: $X \sim N_{p\times n}(M,\,\Sigma,\,\Psi)$, $B' = B$ and $A'\Psi B = 0 \Longrightarrow XA$ and XBX' are independent.
   With $(X,\,Y) \sim N_{p\times n}(\mu J',\,\Sigma,\,I_n)$, $(\overline{X},\,\overline{Y}) = (X,\,Y)A$ and CSSCP $= (X,\,Y)B(X,\,Y)'$ where $A = J(J'J)^{-1}$ and $B = I - JJ^+$,

$$
[J(J'J)^{-1}]'(I - JJ^+) = [(J'J)^{-1}J'](I - JJ^+) = J^+(I - JJ^+) = 0.
$$

   Thus $(\overline{X},\,\overline{Y})$ and CSSCP are independent.

2. Point estimators

   (1) Unbiased estimators

$$
\begin{aligned}
&\overline{X} \sim N\left(\mu_x,\,\tfrac{\Sigma}{n_1}\right) \Longrightarrow \overline{X} \text{ is an UE for } \mu_x \\
&\overline{Y} \sim N\left(\mu_y,\,\tfrac{\Sigma}{n_2}\right) \Longrightarrow \overline{Y} \text{ is an UE for } \mu_y \\
&a\overline{X} + b\overline{Y} = (aI,\,bI)\begin{pmatrix}\overline{X}\\ \overline{Y}\end{pmatrix} \sim N\left((aI,\,bI)\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix},\,(aI,bI)\begin{pmatrix}\frac{\Sigma}{n_1} & 0\\ 0 & \frac{\Sigma}{n_2}\end{pmatrix}\begin{pmatrix}aI\\ bI\end{pmatrix}\right) \\
&\qquad\qquad = N\left(a\mu_x + b\mu_y,\,\left(\tfrac{a^2}{n_1} + \tfrac{b^2}{n_2}\right)\Sigma\right) \\
&\Longrightarrow a\overline{X} + b\overline{Y} \text{ is an UE for } a\mu_x + b\mu_y.
\end{aligned}
$$

$\text{CSSCP}_x \sim W_{p\times p}(n_1-1,\,\Sigma) \Longrightarrow E\left(\frac{\text{CSSCP}_x}{n_1-1}\right) = \Sigma \Longrightarrow S_x = \frac{\text{CSSCP}_x}{n_1-1}$ is an UE for Σ.
$\text{CSSCP}_y \sim W_{p\times p}(n_2-1,\,\Sigma) \Longrightarrow E\left(\frac{\text{CSSCP}_y}{n_2-1}\right) = \Sigma \Longrightarrow S_y = \frac{\text{CSSCP}_y}{n_2-1}$ is an UE for Σ.
$\text{CSSCP} \sim W_{p\times p}(n-2,\,\Sigma) \Longrightarrow E\left(\frac{\text{CSSCP}}{n-2}\right) = \Sigma \Longrightarrow S_p = \frac{\text{CSSCP}}{n-2}$ is an UE for Σ.

Ex2: $S_p = \frac{\text{CSSCP}}{n-2} = \frac{\text{CSSCP}_x+\text{CSSCP}_y}{n-2} = \frac{(n_1-1)S_x+(n_2-1)S_y}{n-2} = \frac{n_1-1}{n_1+n_2-2}S_x + \frac{n_2-1}{n_1+n_2-2}S_y$ is a weighted average of S_x and S_y with weights n_1-1 and n_2-1.

(2) Maximum likelihood estimators
$\overline{X}$ is MLE for μ_x, $\overline{Y}$ is MLE for μ_y, $\frac{\text{CSSCP}}{n}$ is MLE for Σ.
$L\left(\overline{X},\,\overline{Y},\,\frac{\text{CSSCP}}{n}\right) = \left(\frac{n}{2\pi e}\right)^{np/2}\,|\text{CSSCP}|^{-n/2}$.

Proof. Let $L(\mu_x,\,\mu_y,\,\Sigma)$ be the likelihood function. Then

$$\begin{aligned}
L(\mu_x,\,\mu_y,\,\Sigma) &= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\left[\frac{1}{-2}\text{tr}\left(\Sigma^{-1/2}\ \text{CSSCP}\ \Sigma^{-1/2}\right)\right] \\
&\quad \exp\left[\frac{n_1}{-2}(\overline{X}-\mu_x)'\Sigma^{-1}(\overline{X}-\mu_x)\right]\exp\left[\frac{n_2}{-2}(\overline{Y}-\mu_y)'\Sigma^{-1}(\overline{Y}-\mu_y)\right] \\
&\le L(\overline{X},\,\overline{Y},\,\Sigma) \\
&= \frac{|\Sigma^{-1/2}\ \text{CSSCP}\ \Sigma^{-1/2}|^{n/2}}{(2\pi)^{np/2}|\text{CSSCP}|^{n/2}}\exp\left[\frac{1}{-2}\text{tr}\left(\Sigma^{-1/2}\ \text{CSSCP}\ \Sigma^{-1/2}\right)\right] \\
&\le L\left(\overline{X},\,\overline{Y},\,\frac{\text{CSSCP}}{n}\right) = \left(\frac{n}{2\pi e}\right)^{np/2}|\text{CSSCP}|^{-n/2}.
\end{aligned}$$

Conclusion follows.

Ex3: $a\overline{X}+b\overline{Y}$ is MLE for $a\mu_x+b\mu_y$.

3. SAS
for $\overline{X}$, $\overline{Y}$, CSSCP_x and CSSCP_y

(1) Data
File ex.txt contains x1, x2, x3, sid and sname where
sid= $\begin{cases} 6 & \text{Sample 1} \\ -7 & \text{Sample 2} \end{cases}$ and sname= $\begin{cases} \text{good} & \text{Sample 1} \\ \text{bad} & \text{Sample 2} \end{cases}$.

```
data a;
    infile "D\ex.txt";
    input x1 x2 x3 sid sname $ @@;
```

(2) Procedures

```
proc sort;
     by sid;
     run;
proc corr nocorr CSSCP;
     var x1 x2 x3;
     by sid;
     run;
```

```
proc sort;
     by sname;
     run;
proc corr nocorr CSSCP;
     var x1 x2 x3;
     by sname;
     run;
```

L16 Two-sample confidence regions

1. Pivotal quantities

   (1) Analysis

   $$\begin{pmatrix}\overline{X}\\ \overline{Y}\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \begin{pmatrix}\frac{\Sigma}{n_1} & 0\\ 0 & \frac{\Sigma}{n_2}\end{pmatrix}\right) \implies \overline{X}-\overline{Y} \sim N\left(\mu_x-\mu_y,\, \left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)\Sigma\right)$$

   $$\implies \sqrt{\tfrac{n_1 n_2}{n}}H\left[(\overline{X}-\overline{Y})-(\mu_x-\mu_y)\right] \sim N(0, H\Sigma H'),$$

   $$\text{CSSCP} \sim W_{p\times p}(n-2,\, \Sigma) \implies H\ \text{CSSCP}\ H' \sim W_{q\times q}(n-2,\, H\Sigma H').$$

   But $\overline{X}-\overline{Y}$ and CSSCP are independent. So random variable below has distribution $T^2(q,\, n-2)$

   $$\sqrt{\tfrac{n_1 n_2}{n}}\ [H(\overline{X}-\overline{Y})-H(\mu_x-\mu_y)]'\left(\tfrac{H\ \text{CSSCP}\ H'}{n-2}\right)^{-1}\sqrt{\tfrac{n_1 n_2}{n}}\ [H(\overline{X}-\overline{Y})-H(\mu_x-\mu_y)]$$

   (2) Conslusions

   Let $\theta = H(\mu_x-\mu_y) \in R^q$. Then

   $$[\theta - H(\overline{X}-\overline{Y})]'\left(\tfrac{n}{n_1 n_2}HS_pH'\right)^{-1}[\theta - H(\overline{X}-\overline{Y})] \sim T^2(q,\, n-2).$$

   With $H = I_p$ and $\delta = \mu_x-\mu_y$,

   $$[\delta - (\overline{X}-\overline{Y})]'\left(\tfrac{n}{n_1 n_2}S_p\right)^{-1}[\delta - (\overline{X}-\overline{Y})] \sim T^2(p,\, n-2).$$

   (3) $\frac{l'(\overline{X}-\overline{Y})-l'(\mu_x-\mu_y)}{S_{l'(\overline{X}-\overline{Y})}} \sim t(n-2)$.

   Proof. $\overline{X}-\overline{Y} \sim N(\mu_x-\mu_y,\, \frac{n}{n_1 n_2}\Sigma) \Longrightarrow l'(\overline{X}-\overline{Y}) \sim N\left(l'(\mu_x-\mu_y),\, \frac{n}{n_1 n_2}\, l'\Sigma l\right)$.

   So $l'(\overline{X}-\overline{Y})$ has variance estimated by $S^2_{l'(\overline{X}-\overline{Y})} = \frac{n}{n_1 n_2}l'S_pl$.

   In (1) with $H = l'$ such that $l'(\overline{X}-\overline{Y}) \in R$,

   $$\frac{[l'(\overline{X}-\overline{Y})-l'(\mu_x-\mu_y)]^2}{\frac{n}{n_1 n_2}l'S_pl} = \left[\frac{l'(\overline{X}-\overline{Y})-l'(\mu_x-\mu_y)}{S_{l'(\overline{X}-\overline{Y})}}\right]^2 \sim T^2(1,\, n-2) = F(1,\, n-2) = [t(n-2)]^2.$$

   Thus $\frac{l'(\overline{X}-\overline{Y})-l'(\mu_x-\mu_y)}{S_{l'(\overline{X}-\overline{Y})}} \sim t(n-2)$.

   Comments: CSSCP $= \sum_i(X_i-\overline{X})(X_i-\overline{X})' + \sum_j(Y_j-\overline{Y})(Y_j-\overline{Y})'$ is also called the error matrix denoted by E.

2. Confidence regions

   (1) Confidence region for $\theta = H(\mu_x-\mu_y) \in R^q$

   The collection of all $\theta \in R^q$ satisfying

   $$[\theta - H(\overline{X}-\overline{Y}]'\left(\frac{n}{n_1 n_2}HS_pH'\right)^{-1}[\theta - H(\overline{X}-\overline{Y}] \le T_\alpha(q,\, n-2)$$

   is a $1-\alpha$ confidence region for $\theta = H(\mu_x-\mu_y)$.

   Proof. $P\left([\theta - H(\overline{X}-\overline{Y}]'\left(\tfrac{n}{n_1 n_2}HS_pH'\right)^{-1}[\theta - H(\overline{X}-\overline{Y}] \le T_\alpha(q,\, n-2)\right)$

   $= P(T^2(q,\, n-2) < T^2_\alpha(q,\, n-2)) = 1-\alpha$.

   (2) Confidence region for $\delta = \mu_x-\mu_y$

   The collection of all $\delta \in R^p$ satisfying

   $$[\delta - (\overline{X}-\overline{Y}]'\left(\frac{n}{n_1 n_2}S_p\right)^{-1}[\delta - (\overline{X}-\overline{Y}] \le T_\alpha(p,\, n-2)$$

   is a $1-\alpha$ confidence region for $\delta = \mu_x-\mu_y$.

   Proof. Conclusion follows from (1) with $H = I_p$.

(3) Confidence interval for $\theta = l'(\mu_x - \mu_y) \in R$

$l'(\mu_x - \mu_y) \in l'(\overline{X} - \overline{Y}) \pm t_{\alpha/2}(n-2)S_{l'(\overline{X}-\overline{Y})}$ is a $1-\alpha$ confidence interval for $l'(\mu_x - \mu_y)$

Proof.
$$\begin{aligned} 1-\alpha &= P\left(-t_{\alpha/2}(n-2) < t(n-2) < t_{\alpha/2}(n-2)\right) \\ &= P\left(-t_{\alpha/2}(n-2) < \frac{l'(\mu_x-\mu_y) - l'(\overline{X}-\overline{Y})}{S_{l'(\overline{X}-\overline{Y})}} < t_{\alpha/2}(n-2)\right) \\ &= P\left(l'(\overline{X}-\overline{Y}) - t_{\alpha/2}(n-2) < l'(\mu_x-\mu_y) < l'(\overline{X}-\overline{Y}) + t_{\alpha/2}(n-2)S_{l'(\overline{X}-\overline{Y})}\right). \end{aligned}$$

Comments: $T^2(p,\, k) = \frac{kp}{k-p+1}F(p,\, k-p+1) \Longrightarrow T^2_\alpha(p,\, k) = \frac{kp}{k-p+1}F_\alpha(p,\, k-p+1)$.

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

3. Simultaneous confidence regions

(1) Bonferroni intervals for $l_i'(\mu_x - \mu_y)$, $i = 1, ..., k$

$$l_i'(\mu_x - \mu_y) \in l_i'(\overline{X} - \overline{Y}) \pm t_{\alpha/(2k)}(n-2)S_{l_i'(\overline{X}-\overline{Y})},\ i = 1, ..., k$$

are simultaneous confidence intervals for $l_i'(\mu_x - \mu_y)$, $i = 1, ..., k$, with overall confidence coefficient $1-\alpha$.

(2) Scheffe's intervals for $l_i'(\mu_x - \mu_y)$, $i = 1, 2, \cdots$

$$l_i'(\mu_x - \mu_y) \in l_i'(\overline{X} - \overline{Y}) \pm \sqrt{T^2_q(p,\, n-2)}\, S_{l_i'(\overline{X}-\overline{Y})},\ i = 1, 2, \cdots$$

are simultaneous confidence intervals for $l_i'(\mu_x - \mu_y)$, $i = 1, 2, \cdots$, with overall confidence coefficient $1-\alpha$.

Ex: Example 6.4 on page 289

Consider $\begin{pmatrix}\text{on-peak electric consumption}\\ \text{off-peak electric consumption}\end{pmatrix}$ in July in Wisconsin for homes with and without air conditioning. For mean vectors $\mu_x = \begin{pmatrix}\mu_{x1}\\ \mu_{x2}\end{pmatrix}$ and $\mu_y = \begin{pmatrix}\mu_{y1}\\ \mu_{y2}\end{pmatrix}$, construct confidence intervals for $\mu_{xi} - \mu_{yi}$, $i = 1, 2$, with overall confidence coefficient 95% by Scheffe's method.

```
proc sort;
    by AC;
    run;
proc corr nocorr COV;
    var x1 x2;
    by AC;
    run;
```

$\Longrightarrow$

$$n_1 = 45,\ \overline{X} = \begin{pmatrix}204.4\\ 556.6\end{pmatrix},\ S_1 = \begin{pmatrix}13825.3 & 23823.4\\ 23823.4 & 73107.4\end{pmatrix}$$
$$n_2 = 55,\ \overline{Y} = \begin{pmatrix}130.0\\ 355.0\end{pmatrix},\ S_2 = \begin{pmatrix}8632.0 & 19616.7\\ 19616.7 & 55964.5\end{pmatrix}$$

Formula: $\mu_{xi} - \mu_{yi} \in \overline{X}_i - \overline{Y}_i \pm \sqrt{T^2_\alpha(p,\, n-2)}\, S_{\overline{X}_i - \overline{Y}_i}$

$T^2_{0.05}(2,\, 98) = \frac{2\times 98}{97}F_{0.05}(2,\, 97) = 2.02 \times 3.1 = 6.26$

$S_p = \frac{n_1-1}{n-2}S_1 + \frac{n_2-1}{n-2}S_2 = \frac{44}{98}S_1 + \frac{54}{98}S_2 = \begin{pmatrix}10963.7 & 21505.5\\ 21505.5 & 63661.3\end{pmatrix}$

$S^2_{\overline{X}_1 - \overline{Y}_1} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(S_p)_{11} = \frac{100}{45\times 55} \times 10963.7 = 21.047^2$

$S^2_{\overline{X}_2 - \overline{Y}_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(S_p)_{22} = \frac{100}{45\times 55} \times 63661.3 = 50.717^2$

$$\mu_{x1} - \mu_{y1} \in \overline{X}_1 - \overline{Y}_1 \pm \sqrt{6.26}\,21.047 = 74.4 \pm 52.66 = (21.7,\ 127.1)$$
$$\mu_{x2} - \mu_{y2} \in \overline{X}_2 - \overline{Y}_2 \pm \sqrt{6.26}\,50.717 = 201.6 \pm 126.89 = (74.7,\ 328.5)$$

are Sheffe's CIs with overall CC 95%.