SAS軟體介紹3 —PROC ttest

吳淑娟

獨立樣本與相依樣本

- 不同的平均數可能計算自不同的樣本，亦有可能計算自同一個樣本的同一群人，或是具有配對關係的不同樣本。
- 獨立樣本設計
  - 不同平均數來自於獨立沒有關連的不同樣本
  - 根據機率原理，當不同的平均數來自於不同的獨立樣本，兩個樣本的抽樣機率亦相互獨立
- 相依樣本設計
  - 重複量數設計（repeated measure design）：不同的平均數來自於同一個樣本的同一群人（例如某班學生的期中考與期末考成績）重複測量的結果
  - 配對樣本設計（matched sample design）:不同的平均數來自具有配對關係的不同樣本（例如夫妻兩人的薪資多寡）樣本抽取的機率是為非獨立、相依的情況。因此必須特別考量到重複計數或相配對的機率，以供不同的公式。

雙母體平均數檢定—相依樣本

- 當雙母體平均數檢定所使用的樣本是相依樣本時，使用相依樣本平均數檢定，例如某一群受試者參加自我效能訓練方案前後的兩次得分的自我效能平均數的比較。

- $Z = \dfrac{\overline{x_d} - \mu_0}{\sigma / \sqrt{n}}$ $t = \dfrac{\overline{x_d} - \mu_0}{s / \sqrt{n}}$ （或一般式 $t = \dfrac{\overline{D}}{S_D / \sqrt{n}}$ ）

雙母體平均數檢定—獨立樣本

- 研究者關心兩個平均數的差異是否存在之時，是為雙母體平均數檢定的問題，研究假設（H_a）為母體一平均數與母體二平均數具有差異，或$\mu_{x1} \neq \mu_{x2}$。
- 當雙母體平均數檢定所使用的樣本是獨立樣本且族群標準差已知時，使用獨立樣本平均數Z檢定。

$$Z_{obt} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

- 母體標準差未知，且樣本小於30，應使用t統計量（母群標準差未知），進行獨立樣本t考驗公式如下:

$$t_{obt} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{s_p^{\ 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

獨立樣本t檢定示意圖

$$y_1 \;,\; y_2 \;,\ldots\; y_N \qquad\qquad y_1 \;,\; y_2 \;,\ldots\; y_N$$

$$\downarrow n_1 \quad \downarrow n_1 \quad \downarrow n_1 \qquad\qquad \downarrow n_2 \quad \downarrow n_2 \quad \downarrow n_2$$

$$\bar{y}_1 \quad \bar{y}_1 \quad \bar{y}_1 \qquad\qquad \bar{y}_2 \quad \bar{y}_2 \quad \bar{y}_2$$

$$S_1^2 \quad S_1^2 \quad S_1^2 \qquad\qquad S_2^2 \quad S_2^2 \quad S_2^2$$

$$d_k = \overline{Y_{1k}} - \overline{Y_{2k}}$$

$$\Downarrow$$

$$\mu_d = \mu_1 - \mu_2$$

$$t = \frac{\bar{y} - E(\bar{y})}{S_{\bar{Y}}} \quad \text{pooled後}, \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

central limit theorem

① 如果兩樣品來源族群，具有有限變方，當樣品大小，n_1及n_2增大時，所有可能樣品均值差$(\overline{y_1}-\overline{y_2})$的分布便趨於常態。

② 如果兩樣品來源族群是常態，無論樣品大小為多少，樣品均值差 $(\overline{y_1}-\overline{y_2})$ 的分布皆是常態。

$$Z=\frac{(\overline{y_1}-\overline{y_2})-E(\overline{y_1}-\overline{y_2})}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=\frac{\text{value}-\text{mean}}{\text{st. deviation}}$$

$\sigma^2 \to S^2$替代時 $$t=\frac{(\overline{y_1}-\overline{y_2})-\varepsilon}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \quad d.f=?$$

$$\sigma_1^2 = \sigma_2^2 = \sigma_p^2$$

(1) 當F檢定的顯著水準很大(p＞0.05)，即兩母群變異數相等時($S_1^2 = S_2^2$)，則採用綜合變異數t檢定(pooled-variance t test)。

and (a) $n_1 \neq n_2$,

$$t = \frac{(\overline{y_1} - \overline{y_2}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{\upsilon_1 S_1^2 + \upsilon_2 S_2^2}{\upsilon_1 + \upsilon_2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

$$\therefore d.f = \underline{\underline{n_1 + n_2 - 2}}_{\#}$$

$$\sigma_1^2 = \sigma_2^2 = \sigma_p^2$$

(b) $n_1 = n_2 = n$

$$t = \frac{\left(\overline{y_1} - \overline{y_2}\right) - \left(\mu_1 - \mu_2\right)}{S_p\sqrt{\frac{2}{n}}}$$

$$\therefore d.f = 2(n-1)$$

$$\sigma_1 \neq \sigma_2$$

(2) 當F檢定達顯著水準時($p < 0.05$)，即兩母群變異數不相等時($S_1^2 \neq S_2^2$)，則將用個別變異數的t統計量(Cochran & Cox法，1950)。

- if, $\sigma_1^2 \neq \sigma_2^2$ 則

$$t = \frac{\left(\overline{y_1} - \overline{y_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

但degrees of freedom為多少？

$\sigma_1 \neq \sigma_2$

$$(a)\, n_1 \neq n_2$$

$$(i):\quad \left.\begin{array}{l} n_1 \rightarrow t \\ n_2 \rightarrow t \end{array}\right\} \text{算出共同之critical region}$$

$$\text{if}\quad n_1 > n_2$$

$\sigma_1 \neq \sigma_2$

$$df_1 = n_1 - 1 \text{查表} \rightarrow t_{\alpha 1} \quad df_2 = n_2 - 1 \text{查表} \rightarrow t_{\alpha 2}$$

$$t_\alpha = \text{weighted average of } t_{\alpha 1} \text{ and } t_{\alpha 2}$$

$$\Rightarrow t'_\alpha = \frac{\frac{S_1^2}{n_1} t_{\alpha 1} + \frac{S_2^2}{n_2} t_{\alpha 2}}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

(ii):

$$df \longrightarrow \quad \upsilon = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$\sigma_1 \neq \sigma_2$$

(b) $n_1 = n_2 = n$

$$S^2_{(\overline{Y_1}-\overline{Y_2})} = \sqrt{\frac{S_1^2 + S_2^2}{n}} \quad \text{use approximate } t \text{ with df} = n-1$$

$$t = \frac{\left(\overline{y_1} - \overline{y_2}\right) - \left(\mu + \mu_2\right)}{\sqrt{\frac{S_1^2 + S_2^2}{n}}}$$

公式整理

- In general $$t=\frac{(\overline{y}_1-\overline{y}_2)-(\mu_1-\mu_2)}{S_{(\overline{Y}_1-\overline{Y}_2)}}$$

	$\sigma_1^2=\sigma_2^2$: $S_{(\overline{Y}_1-\overline{Y}_2)}$	$\sigma_1^2=\sigma_2^2$: υ	$\sigma_1^2\neq\sigma_2^2$: $S_{(\overline{Y}_1-\overline{Y}_2)}$	$\sigma_1^2\neq\sigma_2^2$: υ
$n_1=n_2$	$\sqrt{\frac{2S_p^2}{n}}=S_p\sqrt{\frac{2}{n}}$	$2(n-1)$	$\sqrt{\frac{S_1^2+S_2^2}{n}}$	$n-1$
$n_1\neq n_2$	$\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	n_1+n_2-2	$\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$	$t\alpha'$

採用同大樣品的理由

- 同大樣品之好處：

（1）可減少 $\sigma_1^2 \neq \sigma_2^2$ 之不良影響

（2）使樣品均值差的變方 $\left[\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]$ 縮到極小而增加測驗能力（↘ β ）

Ex: $n_1 + n_2 = 100$

$$n_1 = n_2 = 50 \quad \Rightarrow \quad \sigma^2\left(\frac{2}{50}\right) = \frac{\sigma^2}{25}$$

$$n_1 = 1 \quad n_2 = 99 \qquad \sigma^2\left(\frac{1}{1}+\frac{1}{99}\right) = \frac{100}{99}\sigma^2$$

獨立 *t* 檢定的前提

前提假設

- 相依變項 (dependent variable) 的本質必須是連續數 (continuous variable)，且是隨機樣本 (random sample)，亦即是從母群體 (population) 中隨機抽樣而的。如果不是連續數，則必須採用無母數分析 (nonparametric test)。
- 相依變項的母體必須是常態分佈 (normal distribution)。若檢測結果不是常態分佈，則不可使用獨立 t 檢定，並須改為無母數分析。
- 其樣本的量測皆為獨立事件 (independent event)，亦即獨立變項 (independent variable) 只有一組或兩組，且第一組的樣本不會影響第二組的樣本，反之亦然。例如性別 (gender)：如果樣本是男性者一定不會影響樣本是女性者的量測。如果不是獨立事件，則應該採用配對 t 檢定。
- 兩組的樣本之變異數 (variance, s) 亦為常態分佈，且為定值 (constant)。如果不是，則其統計值 t 必須調整。

獨立 *t* 檢定的檢測假說

檢測假說 (Hypothesis testing)：

獨立 t 檢定主要在於比較兩組樣本間的平均值是否存在差異，可視為變異數分析 (ANOVA) 的特例 -- 兩組檢測。

- one sample test： 檢測其樣本平均值與母體平均值 (某特定數值) 是否不同。其虛無假設為 H_0： $\mu = \mu_0$
- two sample test： 檢測兩組樣本平均值之差值 (某特定數值) 是否不同。其虛無假設為 H_0： $\mu_1 = \mu_2$

統計模型 (Statistical model)

$$y = \alpha_0 + \alpha_1 x_1$$
$$H_0 : \mu_1 = \mu_2$$

t考驗分析的SAS語法

(一)獨立樣本

1.基本語法：

- PROC TTEST選項串；
- CLASS 自變項名稱；←旨在識別觀察體所屬的組別
- VAR 依變項名稱串；←指明對那些依變項的平均數執行t檢定
- BY 自變項名稱串；←將資料擋內的觀察體加以分組

2.詳細語法：

- PROC TTEST選項串：

選項串：
DATA：輸入資料檔名稱←指名對那一個SAS資料檔執行分析
COCHRAN←使用Cochran＆Cox(1950)機率水準之近似t統計量

- CLASS自變項名稱；←旨在識別觀察體所屬的組別
- VAR依變項名稱串；←指明對那些依變項的平均數執行t檢定
- BY 自變項名稱串；←將資料檔內的觀察體加以分組

(二)相依樣本(詳見第四章第一節)

- PROC MEANS T PRT；
- VAR 依變項名稱；