R 軟體資料分析應用：比例檢定與相關係數

林怡諄 副統計分析師

本期 eNews 與各位討論使用 R 進行比例檢定與相關係數，比例檢定將從單一樣本比例檢定出發，並且討論(獨立)雙樣本比例差異檢定；相關係數則將介紹皮爾生相關係數(Pearson's correlation)以及斯皮爾曼等級相關係數(Spearman's rank correlation coefficient)。以下我們就逐一進行介紹並說明 R 程式步驟。

一、比例檢定

(一) 單一樣本比例檢定(One-sample test for proportion)

1. 檢定說明

單一樣本比例檢定通常應用於，當資料僅有一組樣本資料，我們想要瞭解其母體比例是否大於、小於或是等於某特定數值。舉例說明，前一年度完成學校學生規律吃早餐比例的全面調查，其比例為 65%，我們想要分析今年度規律吃早餐的學生比例是否有增加，則可以抽樣 500 位學生，紀錄是否有規律吃早餐，並使用單一樣本檢定即可得知其比例是否有統計上顯著的變化。

從以上舉例，可以得知另一個資料型態上的注意事項，就是資料型態為類別型資料 (如「是否」規律吃早餐，「是否」為男性，「是」則資料紀錄為 1，「否」則資料紀錄為 0)。如此一來，統計類別型資料時，即可獲得一個樣本的比例數值。

我們假設比例為 p，則單一樣本比例檢定的假設檢定設定如下，虛無假設(Null Hypothesis，H_0) 為 $P = P_0$，而對立假設(Alternative hypothesis，H_1) 為 $P \neq P_0$，而 P_0 為特定數值。進一步，介紹單一樣本比例檢定的檢定統計量如下：

$$\text{檢定統計量} = \frac{\hat{p} - p_0}{\sigma_p}$$

其中， $\hat{p}$ 為樣本比例，P_0 為特定數值， σ_p 為$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$，n 為樣本個數

2. R 程式說明

(1) 資料來源與說明

單一樣本比例檢定的示範資料，可從以下網址進行下載，網址為 http://rweb.tmu.edu.tw/upload/data.php?dir=2&data=%E7%AF%84%E4%BE%8BA-3 ，其資料為討論網路成癮之示範資料，資料分成男性與女性等兩群，而高度網路使用者則紀錄為 1，非高度網路使用者則紀錄為 0。

(2) 假設檢定之設定

我們欲檢定男性的高度網路使用者之比例(P)是否小於等於 30%，

$$H_0: \ P \leq 0.3 \quad \text{v.s.} \quad H_1: \ P > 0.3$$

(3) R 指令語法說明

【語法】

```
prop.test (x, n, p = NULL,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95, correct = TRUE)
```

【參數說明】

x 為所關心類別變數之次數的向量資料

n 為資料庫之總樣本數的向量資料

alternative 為設定對立假設的比例值等於、小於、大於特定值

conf.level 為設定信賴區間的信心水準

correct 為是否進行 Yates 的連續性修正

(4) R 程式碼

```
# 讀取「範例 A-3.csv」，並且建立 PROP_data 的資料檔案

PROP_data <- read.csv(file = "E:\\DEMO\\範例 A-3.csv", header=F)

# 將 PROP_data 資料檔案的兩個變數分別取名為 MALE 以及 FEMAL

names(PROP_data) <- c("MALE" , "FEMALE")

# 將男性資料獨立出來，建構 MALE_DATA 資料檔案

MALE_DATA <- c(PROP_data[,"MALE"])

# 使用加總方式，計算出高度網路使用者的次數

USE_COUNT <-sum(MALE_DATA,na.rm=T)

# 使用加總方式，並且排除缺漏值，計算出男性總人數

num <- sum(!is.na(MALE_DATA))

# 使用 proc.test 的指令，給予高度網路使用者次數、男性總人數、欲檢定特定數值等資訊，
進行單一樣本比例檢定

prop.test(USE_COUNT, num, p = 0.3,

          alternative = c("greater"),

          conf.level = 0.95, correct = TRUE)
```

(5) R 程式執行結果

```
        1-sample proportions test with continuity correction

data:  USE_COUNT out of num, null probability 0.3
X-squared = 6.2891, df = 1, p-value = 0.9939
alternative hypothesis: true p is greater than 0.3
95 percent confidence interval:
 0.2005121 1.0000000
sample estimates:
        p
0.2371429
```

由以上結果顯示，依照男性資料所計算高度網路使用者的比例為 0.2371429 (23.71429%)，95% 信賴區間為 (0.201, 1.000)，其信賴區間包含 0.3，並且 P 值為 0.9939，這代表著無法拒絕虛無假設，男性高度網路使用者比例沒有顯著大於 0.3。

(二) 雙樣本比例檢定(One-sample test for proportion)

1. 檢定說明

雙樣本比例檢定適用於欲瞭解兩個獨立樣本群體的比例是否有其差異，或是其比例差異值大於、小於或是等於某特定數值，例如想要瞭解某一行業男女性的吸菸比例是否有其差異，我們會去抽樣該行業男女性員工，並且記錄是否有吸菸習慣，這裡的資料類型也是類別型變數(有吸菸習慣為 1，無吸菸習慣為 0)。

延續上一小節的男女性的網路成癮之範例，我們假設男性的抽樣人數為 n_1，高度網路使用者比例為 P_1，女性的抽樣人數為 n_2，高度網路使用者比例為 P_2，並且設定虛無假設(Null Hypothesis，H_0) 為 $P_1 - P_2 = 0$，而對立假設(Alternative hypothesis，H_1) 為 $P_1 - P_2 \neq 0$。其檢定統計量如下：

$$\text{檢定統計量} = \frac{(\hat{P}_1 - \hat{P}_2) - P_1 - P_2}{S_{\hat{P}_1 - \hat{P}_2}}$$

$$\text{其中，} S_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

2. R 程式說明

(1) 假設檢定之設定

我們沿用上一節的示範資料，在此，我們欲檢定男女性的高度網路使用者之比例差異是否等於 0，假設檢定之設定如下：

$$H_0: \ P_1 - P_2 = 0 \quad \text{v.s.} \quad H_1: \ P_1 - P_2 \neq 0$$

(2) R 指令語法說明

【語法】

```
prop.test (x, n, p = NULL,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95, correct = TRUE)
```

【參數說明】

x 為所關心類別變數之次數的雙變數資料矩陣 (矩陣包含兩個樣本群組)

n 為資料庫之總樣本數的雙變數資料矩陣 (矩陣包含兩個樣本群組)

alternative 為設定對立假設的比例差異值等於、小於、大於特定值

conf.level 為設定信賴區間的信心水準

correct 為是否進行 Yates 的連續性修正

(3) R 程式碼

```
# 讀取「範例 A-3.csv」，並且建立 PROP_data 的資料檔案
PROP_data <- read.csv(file = "E:\\DEMO\\範例 A-3.csv", header=F)
# 將 PROP_data 資料檔案的兩個變數分別取名為 MALE 以及 FEMAL
names(PROP_data) <- c("MALE" , "FEMALE")
# 將男性與女性資料獨立出來，建構 MALE_DATA_1 與 FEMALE_DATA_1 資料檔案
MALE_DATA_1 <- c(PROP_data[,"MALE"])
FEMALE_DATA_1 <- c(PROP_data[,"FEMALE"])
# 使用加總方式，計算出男性與女性高度網路使用者的次數，分別為 outcome1 與 outcome2 的向量資料
outcome1 <- sum(MALE_DATA_1,na.rm=T)
outcome2 <- sum(FEMALE_DATA_1,na.rm=T)
# 使用加總方式，並且排除缺漏值，計算出男性總人數
n1 <- sum(!is.na(MALE_DATA_1))
n2 <- sum(!is.na(FEMALE_DATA_1))
# 建構上述 R 指令說明之中的 x 與 n 的矩陣，設定為 var1 與 var2，而 var1 代表男性與女性的高度網路使用者次數，var2 代表男性與女性的抽樣樣本數
var1 <- c(outcome1,outcome2)
```

```
var2 <- c(n1,n2)

# 使用 proc.test 的指令，給予高度網路使用者次數、男性總人數、欲檢定特定數值等資訊，進行單一樣本比例檢定

prop.test(var1, var2, alternative = c("two.sided"), conf.level = 0.95, correct = T)
```

(4) R 程式執行結果

```
> prop.test(var1,var2,alternative = c("two.sided"),
+           conf.level = 0.95, correct = T)

        2-sample test for equality of proportions with continuity
        correction

data:  var1 out of var2
X-squared = 20.242, df = 1, p-value = 6.825e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.09491155 0.21937417
sample estimates:
   prop 1    prop 2
0.2371429 0.0800000
```

由以上結果顯示，依照男性與女性資料所計算高度網路使用者的比例分別為 0.2371429 (23.71429%) 以及 0.08 (8%)，而其男女比例差異之 95% 信賴區間為 (0.0949, 0.2194)，其信賴區間不包含 0，並且 P 值為 6.825e-06 (P＝0.000006825)，這代表著拒絕虛無假設，男女性高度網路使用者比例差異值顯著不為 0，此顯示出男性高度網路使用者比例顯著異於女性。

二、相關係數

(一) 皮爾生相關係數(Pearson's correlation coefficient)與斯皮爾曼等級相關係數 (Spearman’s rank correlation coefficient)

1. 檢定說明

皮爾生相關係數(Pearson’s correlation) 與斯皮爾曼等級相關係數 (Spearman’s rank correlation coefficient)通常用於欲瞭解兩個變數之間線性關聯性程度，其相關係數值會介於 -1 與 1 之間，若是相關係數為 -1 ，則兩個變數為

完全負相關，而相關係數為 0，則兩個變數為無相關，若是相關係數為 1，則兩個變數為完全正相關。在此，我們需要注意的事項為，皮爾生相關係數之資料型態是連續型且為比例尺度數值，而比例尺度通常為重量、金額或是長度等類型資料；斯皮爾曼等級相關係數之資料型態是順序尺度資料，而順序尺度係指可依照喜好程度進行排列順序，可謂是表示偏好程度的優先順序，例如問卷選項中有從最好至最差的排序。因此，在應用上需要注意皮爾生相關係數與斯皮爾曼等級相關係數的使用時機。

我們可透過兩個變數之平均數 $(\bar{X}\ ,\bar{Y})$ 與標準差來進行計算其相關係數。

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}}$$

而想要瞭解相關係數 r 是否顯著異於 0，則可透過假設檢定方式來分析，假設母體相關係數為 ρ，其虛無假設(Null Hypothesis，H_0) 為 $\rho = 0$，而對立假設(Alternative hypothesis，H_1) 為 $\rho \neq 0$ ，其檢定統計量如下

$$\text{檢定統計量} = r\sqrt{\frac{n-2}{1-r^2}}$$，其中，r 為樣本相關係數

2. R 程式說明

(1) 資料來源與說明

皮爾生相關係數(Pearson's correlation)的示範資料，可從以下網址下載，網址為，http://www.r-web.com.tw/upload/data_test.php?dir=2&data=%E7%AF%84%E4%BE%8BB-4，

其資料為美國軌道哩數與載客量的相關資料，其美國軌道哩數與載客量變數均為連續性數值且為比例尺度資料。

斯皮爾曼等級相關係數(Spearman's rank correlation coefficient)的示範資料，可從

以下網址下載，網址為:

http://www.r-web.com.tw/upload/data_test.php?dir=2&data=%E7%AF%84%E4%BE%8BB-6

其資料為棒球的打擊率等級與跑壘速度等級的相關資料，而此兩變數為連續型且為順序尺度資料。

(2) 假設檢定之設定

不論是我們欲檢定美國軌道哩數與載客量是否顯著有關聯，或是棒球的打擊率等級與跑壘速度等級是否顯著有關聯，我們均設定其母體相關係數為 ρ，檢設檢定如下所示：

$$H_0：\ \rho = 0 \quad \text{v.s.} \quad H_1：\ \rho \neq 0$$

(3) R 指令語法說明

【語法】

cor.test (x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)

【參數說明】

x, y 為欲分析的兩個變數之向量資料

alternative 為設定對立假設的相關係數等於、小於、大於特定值

exact = 是否計算 Kendall's tau 或是 Spearman's rho 的 p 值

conf.level 為設定信賴區間的信心水準

correct 為是否進行 Kendall's tau 或是 Spearman's rho 的連續性修正

(4) R 程式碼

(1) 皮爾生相關係數(Pearson's correlation)之 R 程式碼

```
# 讀取「範例 B-4.csv」，並且建立 pearson _data 的資料檔案
pearson_data <- read.csv (file = "E:\\DEMO\\範例 B-4.csv", header=F)
# 將 pearson_data 資料檔案的兩個變數分別取名為 railway 以及 guest
names(pearson_data) <- c("railway" , "guest")
# 使用 cor.test 的指令，進行美國軌道哩數與載客量之皮爾生相關係數與檢定
pearson_r = cor.test(pearson_data[,"railway"], pearson_data[,"guest"],
                                alternative = c("two.sided"),
                                method="pearson",exact=T,
                                conf.level = 0.95, continuity = FALSE)
```

(2) 斯皮爾曼等級相關係數(Spearman's rank correlation coefficient)之 R 程式碼

```
# 讀取「範例 B-6.csv」，並且建立 spearman_data 的資料檔案
spearman_data <- read.csv(file = "E:\\DEMO\\範例 B-6.csv", header=F)
# 將 spearman_data 資料檔案的兩個變數分別取名為 hit 以及 run
names(spearman_data) <- c("hit" , "run")
# 使用 cor.test 的指令，進行棒球的打擊率等級與跑壘速度等級之斯皮爾曼等級相關係數與檢定
spearman_r = cor.test(spearman_data[,"hit"],spearman_data[,"run"],
                                alternative = c("two.sided"),
                                method="spearman",exact=T,
                                conf.level = 0.95, continuity = FALSE)
```

(3) 兩個變數的散佈圖與趨勢線之 R 程式碼

```
# 軌道哩數與載客量兩個變數之散佈圖
plot(x=pearson_data[,"railway"], y=pearson_data[,"guest"],
     main="Scatter Plot for pearson",
     xlab="railway",
     ylab="guest")
# 將軌道哩數與載客量兩個變數之散佈圖加上趨勢線，其趨勢線為線性迴歸模型所估算
abline(lm(pearson_data[,"guest"]~pearson_data[,"railway"]))
```

(5) R 程式執行結果

a. 皮爾生相關係數之執行結果

```
> pearson_r

        Pearson's product-moment correlation

data:  pearson_data[, "railway"] and pearson_data[, "guest"]
t = 3.5256, df = 5, p-value = 0.01682
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2510914 0.9765232
sample estimates:
      cor
0.8444714
```

由以上皮爾生相關係數之 R 程式結果可知，美國軌道哩數與載客量之相關係數為 0.844714，為高度正相關。進一步檢定其相關係數是否顯著異於 0，其結果顯示 95% 信賴區間為 (0.2511, 0.9765)，並其信賴區間未包含 0，而 P-Value 為 0.01682，也小於 0.05，所以當軌道哩數愈多時，載客量也會愈多。

b. 斯皮爾曼等級相關係數之執行結果

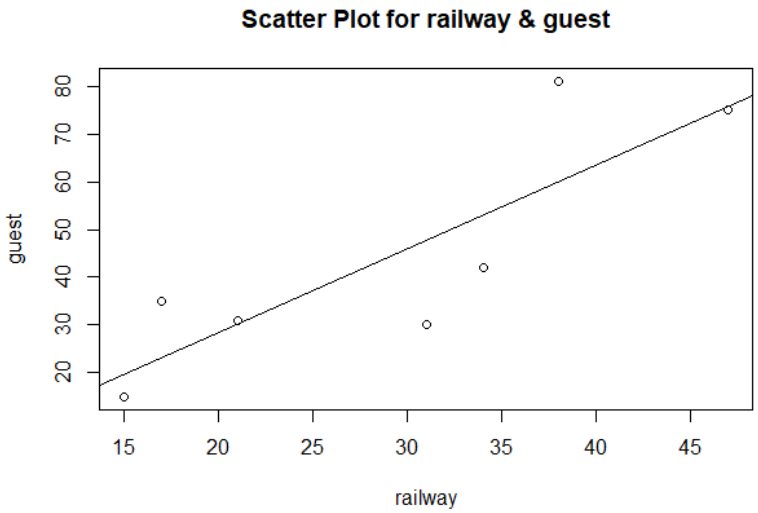
```
> spearman_r

        Spearman's rank correlation rho

data:  spearman_data[, "hit"] and spearman_data[, "run"]
S = 64, p-value = 0.00466
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7762238
```

由以上斯皮爾曼等級相關係數之 R 程式結果可知，棒球打擊率等級與跑壘速度等級之相關係數為 0.7762238，為高度正相關。進一步檢定其相關係數是否顯著異於 0，其 P-Value 為 0.00466，也小於 0.05，所以當棒球打擊率等級愈高時，跑壘速度等級也相對愈高。

c. 美國軌道哩數與載客量範例資料之散佈圖與趨勢線

當進行相關係數分析的同時，我們也可以透過兩個變數之散佈圖形，初步判斷兩個變數之關聯性，若是散佈點趨勢向上，則為正向關係；若是散佈點趨勢向下，則為負向關係；若是散佈點隨機分散，沒有趨勢，則兩者則為無關係。

我們透過以上的 R 程式碼，可以繪製出軌道哩數與載客量之散佈圖與趨勢線，來驗證兩者於平面座標圖形之關聯性，其圖形如下：



由上圖可知，美國軌道哩數與載客量的散佈趨勢，為向上趨勢，再觀察其趨

勢線，更加清楚整個分布趨勢為正斜率向上趨勢，因此，驗證了皮爾生相關係數的數值與檢定結果。

本期 eNews 介紹了比例檢定與相關係數，並且分別討論比例檢定之中單一樣本比例檢定與雙樣本比例差異檢定之 R 程式操作，而相關係數則介紹了皮爾生與斯皮爾曼等級相關係數的使用時機，以及如何使用 R 程式繪製散佈圖與趨勢線，與相關係數之計算與檢定。希望大家能夠透過本期 eNews 瞭解比例檢定與相關係數之 R 程式實作方式。

參考資料

1. 陳景祥， R 軟體應用統計方法，東華書局。
2. 維基百科，Statistical hypothesis testing，網址：https://en.wikipedia.org/wiki/Statistical_hypothesis_testing
3. 維基百科，Correlation and dependence，網址：https://en.wikipedia.org/wiki/Correlation_and_dependence