R 資料分析應用：中位數檢定

王博賢　副統計分析師

本次 eNews 將跟大家介紹如何在 R 中進行『中位數檢定』，本次分析同樣使用 CVD_ALL 這組資料作呈現，檔案位置可在 http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv 下載，資料詳述內容定義可至 http://biostat.tmu.edu.tw/attachment/25_help.docx 文件內觀看。

『中位數檢定』為無母數的方法，它對資料沒有分配的假設，可適用於資料樣本數較小時（通常以樣本筆數<30 為區分標準）。如同平均數檢定， 中位數檢定的方法會依據研究的問題與感興趣的群體而有所不同，包含： 單一樣本中位數檢定、獨立雙樣本中位數差異檢定、成對雙樣本中位數差異檢定、獨立多樣本中位數差異檢定。本次 eNews 將會跟大家一一介紹。

➢ 讀取資料檔案，並自定義 summary 函數

在分析前，首先我們要把檔案讀進 R 中，讀取檔案方式可參考前幾期的 eNews ，我們利用”read.csv”指令來讀取檔案，並命名為 cvd_all。

```
cvd_all <- read.csv(
  file = 'http://biostat.tmu.edu.tw/attachment/94_CVD_ALL.csv '
)
```

為了方便我們接下來觀看資料的筆數，平均數，標準差及中位數，我們先撰寫一個名為 ”my.summary” 的函數，來幫助我們計算

```
my.summary <- function(x){
  x_c <- x[!is.na(x)]
  n <- length(x_c)
  median <- median(x_c)
  mean <- mean(x_c)
  sd <- sd(x_c)
  data.return <- data.frame(n,median,mean,sd)
  return(data.return)
}
```

因為在進行中位數檢定時，無法有遺漏值。所以先利用 ”is.na()” 抓出遺漏值，但因為我們是不要遺漏值，所以要在前面加上 “!” ，並將完整資料命名為 “x_c”，並利用 “length()” 計算筆數，”median” 計算中位數，”mean()” 計算平均數，”sd()” 計算標準差，並把它們合成為一個 data.frame，命名為 “data.return” ，並利用”return()” 輸出“data.return”。

- **單一樣本中位數檢定（Wilcoxon signed-rank test）**

當資料中僅討論單一樣本且樣本數較小時，我們可用『單一樣本中位數檢定』來檢定母體中位數是否大於、小於或等於某一特定數值。我們將使用範例資料來檢定資料檔中男性的 腰圍中位數是否在安全值範圍內（< 90 公分）。

首先我們用 “?wilcox.test” 觀看一下 help 檔

【基本語法】

```
wilcox.test(x, y = NULL,
            alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```

【參數說明】

1. x,y 變數
2. alternative : 單尾，或雙尾檢定
3. mu : 比較特定值時，輸入特定值，或是看兩者差是否大於特定值
4. paired : 是否為成對
5. exact : 是否要精確計算 p-value(樣本數大時，不建議使用)
6. correct : 是否要連續性校正
7. conf.int : 是否要顯示信賴區間
8. conf.level : 信賴區間範圍

了解 wilcox.test 如何使用後，我們就可以開始分析，程式碼如下:

```
#抓出男性(性別=1) 且扣除腰圍為遺漏值的筆數
cvd_m   <- cvd_all[cvd_all$性別==1 & !is.na(cvd_all$腰圍),]
#利用自定義函數觀看男性腰圍
my.summary(cvd_m$腰圍)
#中位數檢定:男性腰圍中位數是否<90 公分為左尾檢定，因筆數夠大所以 exact = F
wilcox.test(cvd_m$腰圍,
            mu=90,
            alternative="less",
            exact = F)
```

output:

```
> #抓出男性(性別=1) 且扣除腰圍為遺漏值的筆數
> cvd_m   <- cvd_all[cvd_all$性別==1 & !is.na(cvd_all$腰圍),]
>
> #利用自定義函數觀看男性腰圍
> my.summary(cvd_m$腰圍)
      n median     mean       sd
1 23417     84 84.37257 9.441471
>
> #中位數檢定:男性腰圍中位數是否<90公分為左尾檢定，因筆數夠大所以exact = F
> wilcox.test(cvd_m$腰圍,
+             mu=90,
+             alternative="less",
+             exact = F)

        Wilcoxon signed rank test with continuity correction

data:  cvd_m$腰圍
V = 47624000, p-value < 2.2e-16
alternative hypothesis: true location is less than 90
```

【分析結果】

首先我們可以由我們自定義的函數得知，篩選出來男性且腰圍沒有遺失的筆數為 23417 筆，中位數為 84，平均數為 84.37257，標準差為 9.441471。

之後由單一樣本中位數檢定我們得知，在 α=0.05 之下，我們的 p-value < 2.2e-16，故拒絕虛無假設，表示資料中男性腰圍中位數沒有超過標準值(90)。

- **獨立雙樣本中位數差異檢定（Wilcoxon rank-sum test）**

當資料為兩組獨立樣本且樣本數較小時，我們可用『獨立雙樣本中位數差異檢定』來檢定兩樣本間母體中位數的差異是否大於、小於或等於 某一特定數值。

假設我們現在感興趣的是有吸菸者與沒有吸菸者之間的腰圍差異，透過此方法可檢定兩群人腰圍的中位數差異是否為 0。

這邊先介紹一下等下會用到的兩個函數，

complete.case(…)

【參數說明】

… 可以放序列、矩陣或是 data frame

【函數說明】

這個函數會依照 row 去判斷是否有遺失值，只要那一筆資料任意變數有遺失，就會回傳 F，若那筆資料全部變數都存在則回傳 T。

tapply(X, INDEX, FUN = NULL)

【參數說明】

X : 為我們要放入 function 的變數

INDEX : 分類依據

FUN : 要執行的 function

【函數說明】

這個函數會依照我們給定的分類依據，對我們的變數作分類並執行給定的函數。

做獨立雙樣本中位數差異檢定所用的函數一樣為“wilcox.test”，但依照資料型態的不同，程式碼也有些微的差距，程式碼如下:

```
#先抓取腰圍及抽菸兩個變數，並命名為 smoke_data
smoke_data <- cvd_all[,c("腰圍","抽菸")]

#做檢定時不能有遺漏值，所以我們使用 complete.case()保留腰圍與抽菸都在的筆數
smoke_data_c <- smoke_data[complete.cases(smoke_data),]

#利用 tapply 幫助我們觀看有抽菸與沒抽菸的腰圍概況
tapply(smoke_data_c$腰圍,smoke_data_c$抽菸,my.summary)

#a.資料型態為一檢定變數及一分組變數時，使用 formula 欲檢定變數~分類變數
wilcox.test(formula=腰圍~抽菸,
            data=smoke_data_c,
            alternative="two.sided",
            exact = F)

#b. 資料型態為兩獨立樣本
#把資料分為有抽菸與沒抽菸兩個
smoke_no  <- smoke_data_c$腰圍[smoke_data_c$抽菸==0]
smoke_yes <- smoke_data_c$腰圍[smoke_data_c$抽菸==1]

wilcox.test(smoke_no,smoke_yes, "two.sided",mu=0,exact = F)
```

output:

```
> #先抓取腰圍及抽菸兩個變數，並命名為smoke_data
> smoke_data <- cvd_all[,c("腰圍","抽菸")]
> #做檢定時不能有遺漏值，所以我們使用complete.case()保留腰圍與抽菸都在的筆數
> smoke_data_c <- smoke_data[complete.cases(smoke_data),]
> #利用tapply 幫助我們觀看有抽菸與沒抽菸的腰圍概況
> tapply(smoke_data_c$腰圍,smoke_data_c$抽菸,my.summary)
$`0`
      n median     mean       sd
1 45227     76 76.79177 10.32995

$`1`
      n median     mean       sd
1 17280     83 82.40398 10.47467

>
> #a.資料型態為一檢定變數及一分組變數時，使用formula 欲檢定變數~分類變數
> wilcox.test(formula=腰圍~抽菸,
+             data=smoke_data_c,
+             alternative="two.sided",
+             exact = F)

        Wilcoxon rank sum test with continuity correction

data:  腰圍 by 抽菸
W = 268020000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

>
> #b. 資料型態為兩獨立樣本
> smoke_no  <- smoke_data_c$腰圍[smoke_data_c$抽菸==0]
> smoke_yes <- smoke_data_c$腰圍[smoke_data_c$抽菸==1]
> wilcox.test(smoke_no,smoke_yes, "two.sided",mu=0,exact = F)

        Wilcoxon rank sum test with continuity correction

data:  smoke_no and smoke_yes
W = 268020000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

【分析結果】

從 tapply()的結果我們得知沒有抽菸的筆數為 45227，中位數為 76，有抽菸的筆數為 17280，中位數為 83 ，因為我們是想知道的是是否有差異，故為雙尾檢定 alternative="two.sided"，從結果得知，在 α=0.05 之下，我們的 p-value < 2.2e-16，故拒絕虛無假設，表示有吸菸者與沒有吸菸者的腰圍中位數是有所差異的。

- **成對雙樣本中位數差異檢定（Wilcoxon signed-rank test）**

當我們想比較資料中兩兩成對樣本的差異值（如：減肥前體重與減肥 後體重相減）且樣本數小時我們可用『成對雙樣本中位數差異檢定』。 此差異值資料可視為單一組樣本，即檢定此差異值資料的母體中位數是否大於、小於或等於某一特定數值。

同一人的收縮壓與舒張壓為一成對樣本，我們可用範例資料來檢定研究對象的脈壓差（收縮壓與舒張壓相減值）中位數是否高於標準值（> 60 mmHg）。

進行成對雙樣本中位數差異檢定時一樣使用“wilcox.test”，程式碼如下:

```
#先抓取舒張壓及收縮壓兩個變數，並命名為 BP
BP <- cvd_all[,c("舒張壓","收縮壓")]

#做檢定時不能有遺漏值，所以我們使用 complete.case()保留舒張壓及收縮壓都在的筆數
BP_c <- BP[complete.cases(BP),]

#利用自定義函數觀看脈壓差
BP_diff <- BP_c$收縮壓-BP_c$舒張壓
my.summary(BP_diff)

wilcox.test(BP_c$收縮壓,
            BP_c$舒張壓,
            paired=T,
            mu=60,
            alternative="greater",
            exact = F)

#可以先自行相減，再進行單一樣本中位數檢定（Wilcoxon signed-rank test）
wilcox.test(BP_diff,
            mu=60,
            alternative="greater",
            exact = F)
```

output:

```
> #先抓取舒張壓及收縮壓兩個變數，並命名為BP
> BP <- cvd_all[,c("舒張壓","收縮壓")]
>
> #做檢定時不能有遺漏值，所以我們使用complete.case()保留舒張壓及收縮壓都在的筆數
> BP_c <- BP[complete.cases(BP),]
>
> #利用自定義函數觀看脈壓差
> BP_diff <- BP_c$收縮壓-BP_c$舒張壓
> my.summary(BP_diff)
      n median     mean       sd
1 63205     43 45.15957 14.32348
>
> wilcox.test(BP_c$收縮壓,
+             BP_c$舒張壓,
+             paired=T,
+             mu=60,
+             alternative="greater",
+             exact = F)

        Wilcoxon signed rank test with continuity correction

data:  BP_c$收縮壓 and BP_c$舒張壓
V = 157200000, p-value = 1
alternative hypothesis: true location shift is greater than 60

>
> #可以先自行相減，再進行單一樣本中位數檢定（Wilcoxon signed-rank test）
> wilcox.test(BP_diff,
+             mu=60,
+             alternative="greater",
+             exact = F)

        Wilcoxon signed rank test with continuity correction

data:  BP_diff
V = 157200000, p-value = 1
alternative hypothesis: true location is greater than 60
```

【分析結果】

從上面結果得知我們共有 63205 筆資料，且脈壓差中位數為 43，要注意的是 wilcox.test(x,y)，x 與 y 放的順序要小心，假設我們把舒張壓放前面，收縮壓放後面，則程式會拿舒張壓-收縮壓，因為是成對檢定，所以 paired=T，我們想要比較是否大於 60，故 alternative="greater"。

在 α=0.05 之下，本分析之虛無假設為母體中位數差異≦60，而 p-值為 1 表不 顯著，無法拒絕虛無假設，也就是說資料中研究對象之脈壓差中位數在標準值範圍內。

- **獨立多樣本中位數差異檢定（Kruskal-Wallis test）**

當資料中包含多組樣本（三組以上之樣本）且樣本數較小時，我們可用『獨立多樣本中位數差異檢定』來檢定多組樣本間母體中位數是否有差異。延續獨立雙樣本中位數差異檢定的分析結果，我們已知有吸菸者與沒有吸菸者的腰圍是有所差異的，接下來可再將吸菸者分為三個等級，每日一包、每日兩包及每日三包以上，進一步來檢定這三組吸菸者的腰圍中位 數是否有差異。

進行獨立多樣本中位數差異檢定時使用“kruskal.test”，程式碼如下:

```
#先抓取腰圍及抽菸量，並命名為 smoke_data
smoke_data <- cvd_all[,c("腰圍","抽菸量")]

#做檢定時不能有遺漏值，所以我們使用 complete.case()保留腰圍及抽菸量都在的筆數
smoke_data_c <- smoke_data[complete.cases(smoke_data),]

#因為我們討論的是每日一包、每日兩包及每日三包以上的人，故把抽菸量為 1,2,3 者抓出
smoke_data_no1 <- smoke_data_c[smoke_data_c$抽菸量 %in% c(1,2,3),]

#利用 tapply 幫助我們觀看有各抽菸量的腰圍概況
tapply(smoke_data_no1$腰圍,smoke_data_no1$抽菸量,my.summary)

#資料型態為一檢定變數及一分組變數 故用 formula 的形式
kruskal.test(腰圍~抽菸量,
             data = smoke_data_no1)
```

output

```
>
> #先抓取腰圍及抽菸量，並命名為smoke_data
> smoke_data <- cvd_all[,c("腰圍","抽菸量")]
>
> #做檢定時不能有遺漏值，所以我們使用complete.case()保留腰圍及抽菸量都在的筆數
> smoke_data_c <- smoke_data[complete.cases(smoke_data),]
>
> #因為我們討論的是每日一包、每日兩包及每日三包以上的人，故把抽菸量為1,2,3者抓出
> smoke_data_no1 <- smoke_data_c[smoke_data_c$抽菸量 %in% c(1,2,3),]
>
> #利用tapply 幫助我們觀看有各抽菸量的腰圍概況
> tapply(smoke_data_no1$腰圍,smoke_data_no1$抽菸量,my.summary)
$`1`
      n median     mean       sd
1 14208     82 82.14108 10.42624

$`2`
     n median     mean       sd
1 1573     86 85.63477 9.947858

$`3`
    n median    mean     sd
1 164     89 87.4878 10.305

>
> #資料型態為一檢定變數及一分組變數 故用formula 的形式
> kruskal.test(腰圍~抽菸量,
+                data = smoke_data_no1)

        Kruskal-Wallis rank sum test

data:  腰圍 by 抽菸量
Kruskal-Wallis chi-squared = 198.53, df = 2, p-value < 2.2e-16
```

【分析結果】

從 tapply 的結果我們知道抽 1, 2, 3 包的筆數分別為 14208, 1573, 164 筆，腰圍中位數分別為 82, 86, 89。而在 α=0.05 之下，本分析之虛無假設為三組之母體中位數完全相等，而 p-值 < 2.22e-16 為顯著，拒絕虛無假設，表示各組資料中位數不完全相等。

本期生統 eNews 的介紹在這邊告一段落囉！這次分別介紹了中位數檢 定的四種方法：單一樣本中位數檢定、獨立雙樣本中位數差異檢定、成對雙樣本中位數差異檢定、獨立多樣本中位數差異檢定，希望本期生統 eNews 能幫助大家更加熟悉 R 中這些檢定方法的操作方式。