R 資料分析應用：變異數(標準差)檢定

沈郁婷 副統計分析師

上期 eNews 介紹中位數檢定，本期將跟大家介紹如何使用 R 進行變異數差異檢定，分為以下三種檢定方法 1.單一樣本變異數(標準差)檢定 2. (獨立)雙樣本變異數(標準差)差異檢定 3. (獨立)多樣本變異數(標準差)差異檢定。延續上期使用的 CVD_ALL 資料呈現，資料詳細內容及定義可至(http://biostat.tmu.edu.tw/download/#data)下載，CVD 完整資料。

變異數檢定與平均數檢定類似，一樣有分左右尾檢定或雙尾檢定，唯一不同處是平均數檢定多檢定是否為常態分布，而變異數檢定是用卡方檢定，檢定其是否為 F 分布，如圖 1 所示，不同的自由度，其分布圖形也不同。

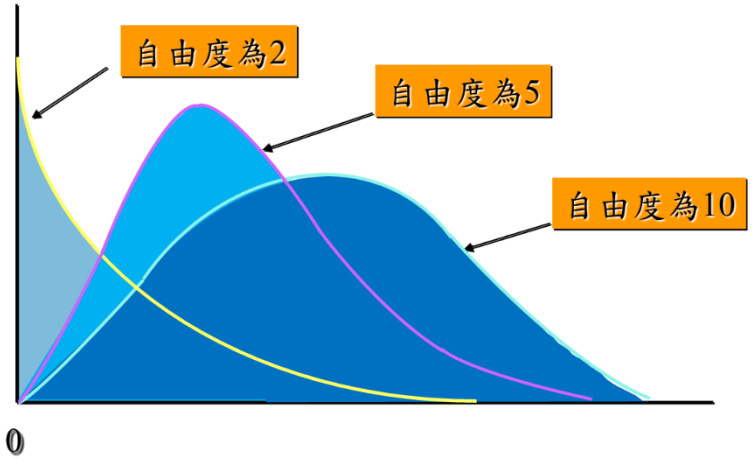

圖 1、F 分配圖

✧ 單一樣本變異數(標準差)檢定

「單一樣本變異數(標準差)檢定」此處的統計分析方法為單一樣本變異數檢定，變異數是用來觀看資料離散程度重要的參數，也就是該變數離期望值的距離。此處單一樣本變異數檢定可以幫助了解母體的變異數是否大於、小於或等於某一特定數值。R 裡面並沒有開發特別針對單樣本變異數檢定的函數，雖然沒有寫好的函數，但仍能根據統計公式輕易推算出來，以下舉個例子。

《範例》

想了解此資料庫，所有受試者年齡變化差異是否等於 10 歲。母體變異數檢定的公式是 $\chi^2 = (n-1)S^2/{\sigma_0}^2$ ，決策臨界值是採用卡方分配查表值 $\chi^2(\mathrm{n}-1)$。

- 假設檢定H_0:σ^2=10 ，H_1:$\sigma^2 \neq 10$

【程式碼】

```
#輸入資料，並命名為 cvd_all，注意 R 大小寫有區分，要注意檔案名稱。
cvd_all<-read.csv( file = "D:\\【E 報、技術會議】\\eNews201812\\CVD_ALL.csv")
#n 代表有興趣欄位之樣本數，is.na()語法為去除遺漏值(原因為有遺漏值，有些函數無法計算)
n=sum(!is.na(cvd_all$年齡))
# sigma.squ 代表母群體的變異數，名稱可自由設置，假設年齡變化等於 10
sigma.squ=10
# s.squ 代表樣本變異數，選取有興趣欄位，並去除遺漏值後計算
s.squ=var(cvd_all$年齡,na.rm=T )
#卡方檢定值(公式計算)
test.a=(n-1)*s.squ/sigma.squ
#c1 卡方分配左尾臨界值；c2 卡方分配右尾臨界值，α=0.05
c1=qchisq(0.025,n-1)
c2=qchisq(0.975,n-1)
cat("卡方檢定值=",test.a,"\n")
cat("左、右尾臨界值:",c1,",",c2,"\n")
if (test.a< c1 || test.a>c2 ){
   cat("拒絕虛無假設!\n")
}else{
   cat("不拒絕虛無假設\n")
}
#計算 P-value
if (s.squ<sigma.squ){
   p.value=2*pchisq(test.a,n-1)
}else{
   p.value=2*(1-pchisq(test.a,n-1))
```

```
}
cat("p-value=", p.value,"\n")
```

《output》單一樣本變異數(標準差)檢定

```
> n=sum(!is.na(cvd_all$年齡))
> sigma.squ=10
> s.squ=var(cvd_all$年齡,na.rm=T )
> test.a=(n-1)*s.squ/sigma.squ
> c1=qchisq(0.025,n-1)
> c2=qchisq(0.975,n-1)
> cat("卡方檢定值=",test.a,"\n")
卡方檢定值= 1245134
> cat("左、右尾臨界值:",c1,",",c2,"\n")
左、右尾臨界值: 63781.04 , 65188.75
> if (test.a< c1 || test.a>c2 ){
+   cat("拒絕虛無假設!\n")
+ }else{
+   cat("不拒絕虛無假設\n")
+ }
拒絕虛無假設!
> #計算P-value
> if (s.square<sigma.squ){
+   p.value=2*pchisq(test.a,n-1)
+ }else{
+   p.value=2*(1-pchisq(test.a,n-1))
+ }
> cat("p-value=", p.value,"\n")
p-value= 0
```

【分析結果】

由於卡方檢定值=1245134，大於右尾臨界值 65188.75，因此拒絕虛無假設。P 值因為太小了(預設顯示為小數點後 9 位)，以至於用 0 表示，下結論時，P 值可以寫 P<0.0001 來代表，結論為，年齡變化差異不等於 10 歲。

✧ 獨立雙樣本變異數(標準差)差異檢定

「獨立雙樣本變異數(標準差)差異檢定」與單樣本變異數(標準差)差異檢定最大的不同是資料中含有兩組樣本。為了解兩組樣本之母群體變異數是否有差異，可將此二組樣本的母體變異數相除得到一個比值(通常預設值為 1)，接著檢定此比值是否大於、小於或等於某一特定數值，即可了解此二組資料母體變異數的差異。一般來說最常計算雙樣本變數差異，是在檢定雙樣本平均數檢定時，當變異數沒有差異，才能使用 t test 比較兩樣本的平均數，若變異數有差異就需要使用不假設變異數相等的 t test。雙樣本變異數公式是使用兩個獨立卡方統計量

公式為：$F=\frac{{\chi_1}^2}{V_1}/\frac{{\chi_2}^2}{V_2}$ ；$F=\frac{{S_1}^2}{{S_2}^2}$；雙尾檢定臨界值公式 $F\geq F_{\alpha/2}$。

《範例》

根據剛剛使用的資料庫，假設想了解有吸菸的受試者與沒有吸菸的受試者兩族群的空腹血糖值變化是否有差異，換言之就是檢定兩組間的空腹血糖值變異數是否相等。R 語言有函數可以使用，不需要再另外帶入公式，使用 var.test 函數來呈現。

首先我們用 “?var.test” 觀看一下 help 檔

【基本語法】

var.test(x, ...)

var.test(x, y, ratio = 1,alternative = c("two.sided", "less", "greater"),conf.level = 0.95, ...)

【參數說明】

1. x,y：變數
2. ratio：x,y 變數族群變異假設比值
3. alternative :單尾，或雙尾檢定
4. conf.level：信賴區間

- 假設檢定

H_0: 兩個變異數比例等於 1(即兩個變異數相等，也可寫成${\sigma_1}^2 = {\sigma_2}^2$)

H_1: 兩個變異數比例不等於 1(即兩個變異數不相等，也可寫成${\sigma_1}^2 \neq {\sigma_2}^2$)

【程式碼】

```
#延續上面的程式碼，使用 cvd_all 這個檔案，根據上面的 help，修改程式碼如下，由於
var.test 函數限制，故先將要分組的變項挑出來，且刪除有興趣變項的遺漏值。
#x1 代表此資料庫篩選出沒有抽菸者，且排除空腹血糖值遺漏的受試者；x2 則代表有抽菸
者，一樣排除空腹血糖值遺漏的受試者。
x1<-cvd_all[cvd_all$抽菸== 0 & !is.na(cvd_all$空腹血糖),]
x2<-cvd_all[cvd_all$抽菸== 1 & !is.na(cvd_all$空腹血糖),]
```

```
#檢定兩組空腹血糖值，使用雙尾檢定

var_re<-var.test(x1$空腹血糖, x2$空腹血糖, alternative = "two.sided")

#查看 P 值

var_re$p.value
```

《output》獨立雙樣本變異數(標準差)差異檢定

```
> x2<-cvd_all[cvd_all$抽菸== 1 & !is.na(cvd_all$空腹血糖),]
> var.test(x1$空腹血糖, x2$空腹血糖, alternative = "two.sided")

        F test to compare two variances

data:  x1$空腹血糖 and x2$空腹血糖
F = 0.94619, num df = 43487, denom df = 16591, p-value = 1.664e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9224869 0.9703541
sample estimates:
ratio of variances
         0.9461865

> var_re<-var.test(x1$空腹血糖, x2$空腹血糖, alternative = "two.sided")
> var_re$p.value
[1] 1.663748e-05
```

【分析結果】

檢定結果顯示，P-value=1.663748e-05 代表 1.66*10^-5< α=0.05，即拒絕虛無假設，代表有沒有吸菸兩組變異數比例不等於 1(即兩組空腹血糖值的變異數不相等) ，95%信賴區間為 0.922-0.970。

✧ (獨立)多樣本變異數(標準差)差異檢定(Levene test)

與(獨立)雙樣本變異數(標準差)差異檢定一樣都使用同一個公式，最大差別在於是兩組樣本還是三組樣本以上，當有興趣比較族群為三組以上時即可使用(獨立)多樣本變異數(標準差)差異檢定。

《範例》

呈上例，已知有無吸菸的行為兩組空腹血糖值變異數不相等，進一步想了解，吸菸量多寡，其空腹血值的變化是否會有差異。

首先我們用 “?leveneTest” 觀看一下 help 檔

leveneTest(y, group, center=median, ...)

【參數說明】

5. y：有興趣變數
6. group：分組變數
7. alternative :單尾，或雙尾檢定
8. center：mean：指的是 Levene's test 原本提出的公式所使用的是平均值；median 是後續 Brown and Forsythe(1974)提出的另一種使用 median 的方法，有興趣者請見參考資料 (Levene Test for Equality of Variances http://atomic.phys.uni-sofia.bg/local/nist-e-handbook/e-handbook/eda/section3/eda35a.htm)，在此處不做多加說明。

➢ 假設檢定

H_0: ${\sigma_1}^2 = {\sigma_2}^2 = {\sigma_3}^2$(各組資料的變異數相等)

H_1: (各組資料的變異數不全等)

【程式碼】

```
#leveneTest 需要下載 car 這個 package
install.packages("car")
library(car)
#有興趣變項為空腹血糖值和抽菸量特別抓出這兩個變項
cvd_a<-cvd_all[,c("空腹血糖","抽菸量")]
#檢定時不能有遺漏值，排除兩個變項的遺漏值，保留空腹血糖值及抽菸量都有資料的筆數
CVD <- cvd_a[complete.cases(cvd_a),]
#0 的部分為沒有抽菸者，由於有興趣的變項是有抽菸者的抽菸量，因此只留下每日一包、每日兩包及每日三包以上
CVD_b <- CVD[CVD$抽菸量 %in% c(1,2,3),]
#在根據函數帶入資料，這邊使用原始 Levene's test 原本提出的公式，因此 center 設 mean
leveneTest(CVD_b$空腹血糖, CVD_b$抽菸量,center=mean)
```

《output》(獨立)多樣本變異數(標準差)差異檢定

```
> cvd_a<-cvd_all[,c("空腹血糖","抽菸量")]
> CVD <- cvd_a[complete.cases(cvd_a),]
> CVD_b <- CVD[CVD$抽菸量 %in% c(1,2,3),]
> leveneTest(CVD_b$空腹血糖, CVD_b$抽菸量,center=mean)
Levene's Test for Homogeneity of Variance (center = mean)
         Df F value    Pr(>F)
group     2  41.135 < 2.2e-16 ***
      15279
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

【分析結果】

檢定結果顯示，F 值=41.135，P-value=2.2e-16 代表 2.2*10^-16 < α=0.05，即拒絕虛無假設，代表抽菸量三組的空腹血糖值變異數不完全相等，即表示三組的變異數是有差異的 。

此次最主要介紹變異數(標準差)檢定，使用時機大多在檢定平均數前，要確保樣本間變異數相等才可進行特定平均數檢定方法。最後本期共介紹三種檢定方法：單一樣本變異數(標準差)檢定、(獨立)雙樣本變異數(標準差)差異檢定、(獨立)多樣本變異數(標準差)差異檢定，本期生統 eNews 的介紹到這結束。

參考資料

1. R 軟體 應用統計方法 陳景祥編著 東華書局

2. Levene Test for Equality of Variances http://atomic.phys.uni-sofia.bg/local/nist-e-handbook/e-handbook/eda/section3/eda35a.htm

3. CRAN - Package car - R Project https://cran.r-project.org/web/packages/car/index.html