Kiểm định tỉ lệ, phương sai hai tổng thể

Ngày 30 tháng 9 năm 2018

Nhóm .....Vắng: .....

Câu hỏi 1 (Ôn tập). *Dữ liệu ChiTieu2010.csv là mẫu điều tra ngẫu nhiên vài chục nghìn hộ gia đình ở nước ta. Từ đó, tại mức ý nghĩa 5% hãy thực hiện các kiểm định sau*

1. *Kiểm định khẳng định cho rằng trung bình một năm các hộ gia đình nước ta dành cho chi tiêu điều nội trú nhiều hơn chi tiêu điều trị ngoại trú. Mẫu được chọn là theo đôi hay độc lập?*
2. *Kiểm định khẳng định cho rằng chi giáo dục trung bình của các hộ ở khu thành thị (khu vực 1) là cao hơn so với nông thôn (khu vực 2). Mẫu được chọn là độc lập hay theo đôi? Trong các lời giải đó giải thích vì sao lại dùng kiểm định t hay kiểm định z.*

Câu hỏi 2. *Khảo sát cho thấy trong số 400 sinh viên học ngành kế toán ra trường có 300 sinh viên có việc làm, trong khi đó chỉ có 200 sinh viên học quản trị kinh doanh trong số 300 sinh viên được khảo sát đang có việc làm. Tại mức ý nghĩa 5%, có thể cho rằng tỉ lệ xin được việc của sinh viên học ngành kế toán là cao hơn so với tỉ lệ đó của tổng thể sinh viên học ngành quản trị kinh doanh?*

Câu hỏi 3. *Người ta muốn so sánh chỉ số IQ của những đứa trẻ hay chơi cờ với những đứa trẻ hay chơi game. Họ điều chọn được 15 cặp sinh đôi, trong mỗi cặp có 1 bé ham chơi game, 1 bé ham chơi cờ. Ta giả định rằng hai tổng thể có phân bố chuẩn. Trước khi so sánh trung bình, người ta phải xem nó có được coi là có phương sai như nhau hay không. Dựa vào mẫu sau đây, hãy trả lời câu hỏi đó ở mức ý nghĩa 5%.*

Cặp	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Chơi game	*126*	*115*	*133*	*136*	*111*	*89*	*101*	*126*	*110*	*122*	*125*	*114*	*110*	*119*	*98*
Chơi cờ	*117*	*138*	*111*	*148*	*106*	*119*	*125*	*120*	*134*	*109*	*97*	*128*	*120*	*131*	*128*

Câu hỏi 4. *Từ dữ liệu ChiTieu2010.csv, hãy kiểm định những khẳng định sau tại mức ý nghĩa 5%:*

1. *Tỉ lệ hộ nghèo ở nông thôn là cao hơn thành thị.*
2. *Phương sai của chi tiêu giáo dục của tổng thể hộ gia đình ở nông thôn và của của tổng thể các hộ gia đình ở thành thị là ngang nhau. Giả sử chi tiêu cho giáo dục của hai tổng thể đều có phân bố chuẩn.*

Câu hỏi 5. *Từ tập dữ liệu trên, hãy tự thiết kế ra ít nhất 3 bài toán kiểm định về tỉ lệ và phương sai hai tổng thể, giải nó và kết luận.*

Họ và Tên: Nguyễn Văn Sang - 1851061983 - 60TH2

BÀI LÀM

Câu 1:

1, Gọi muA,muB lần lượt là trung bình tổng thể chi tiêu điều trị nội trú và chi tiêu điều trị ngoại trú H0: muA <= muB H1: muA > muB

```
chitieu = read.csv("ChiTieu2010.csv")
> attach(chitieu)
> x = DieuTriNoiTru
> y = DieuTriNgoaiTru
t.test(x, y, mu=0, alternative="g",var.equal = FALSE)
```

Welch Two Sample t-test

data: x and y

t = 0.8283, df = 15914, p-value = 0.2038

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

 -4.674444   Inf

sample estimates:

mean of x mean of y

 93.92887 89.18773

Vì p-value = 0.2038> 0.05 nên chấp nhận giả thuyết bác bỏ đối thuyết

Vậy trung bình một năm các hộ gia đình nước ta dành cho chi tiêu nội trú nhiều hơn chi tiêu điều trị ngoại trú

Mẫu này là mẫu độc lập

2, Gọi muA,muB lần lượt là trung bình tổng thể chi tiêu thành thị (khu vực 1) và chi tiêu nông thôn (khu vực 2)

H0: muA <= muB H1: muA > muB

```
chitieu = read.csv("ChiTieu2010.csv")
> attach(chitieu)
> thanhthi=ChiTieuGiaoDucTrongNam[KhuVuc==1]
> nongthon=ChiTieuGiaoDucTrongNam[KhuVuc==2]
> t.test(thanhthi, nongthon, mu=0, alternative="g",var.equal = FALSE)

        Welch Two Sample t-test

data:  thanhthi and nongthon
t = 9.5558, df = 2871.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 186.9815      Inf
sample estimates:
mean of x mean of y
 402.3981  176.5241
```

Vì p-value < 2.2e-16< 0.05 nên bác bỏ giả thuyết chấp nhận đối thuyết

Mẫu này là mẫu độc lập

Sử dụng hàm t.test vì giả thuyết không cho biết phuong sai, mẫu chọn độc lập

Câu 2:

Goi p1,p2 lần lượt là tỉ lệ xin được việc làm của sinh viên ngành kế toán và sinh viên học ngành quản trị kinh doanh

Bài toán kiểm định giả thiết cho hiệu 2 tỉ lệ, cỡ mẫu lớn

H0:p1-p2<=0; H1:p1-p2>0 ;

Ta có:

x=c(300,200); n=c(400,300) thỏa mãn 5≤300≤200-5; 5≤250≤300-5

```
>prop.test(x,n, alternative = "g", correct = FALSE)
```

2-sample test for equality of proportions without continuity correction

data: x out of n
X-squared = 5.8333, df = 1, p-value = 0.007863
alternative hypothesis: greater

95 percent confidence interval:
0.02612905 1.00000000
sample estimates:
prop 1 prop 2
0.7500000 0.6666667

Với p-value = 0.007863< 0.05 nên ta bác bỏ giả thuyết H0. Tỉ lệ có việc làm của sinh viên kế toán cao hơn tỉ lệ có việc làm của sinh viên quản trị kinh doanh.

Câu 3:

Gọi V1, V2 lần lượt là phương sai của tổng thể 2 cặp chơi game và chơi cờ

H0 : V1/V2=1 ; H1 : V1/V2 khác 1

```
x=scan()
1: 126 115 133 136 111 89 101 126 110 122 125 114 110 119 98
16:
Read 15 items
> y=scan()
1: 117 138 111 148 106 119 125 120 134 109 97 128 120 131 128
16:
Read 15 items
> var.test(x, y, ratio = 1,alternative = "t",conf.level = 0.95)

        F test to compare two variances

data:  x and y
F = 0.9765, num df = 14, denom df = 14, p-value = 0.9651 alternative
hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3278395 2.9085870 sample
estimates: ratio of
variances
0.9764987
```

Do p-value = 0.9651 > 0.05 nên chấp nhận giả thuyết H0, có thể xem 2 phương sai của các tổng thể như nhau.

Câu 4:

1,

Goi p1,p2 lần lượt là tỉ lệ hộ nghèo ở nông thôn và thành thị

Bài toán kiểm định giả thiết cho hiệu 2 tỉ lệ, cỡ mẫu lớn

H0:p1-p2<=0; H1:p1-p2>0 ;

```
>table(HoNgheo,KhuVuc==1)


HoNgheo FALSE TRUE
          0  4830 2459
          1  1921  188
> table(HoNgheo,KhuVuc==2)

HoNgheo FALSE TRUE
          0   2459 4830
          1   188 1921
> table(KhuVuc)
KhuVuc
   1    2
2647 6751
> x=c(188,1921)
> n=c(2647,6751)
> prop.test(x,n,alt="g",conf.level = 0.95,correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

data: x out of n
X-squared = 498.1, df = 1, p-value = 1
alternative hypothesis: greater
95 percent confidence interval:
-0.2257343 1.0000000
sample estimates:
prop 1 prop 2
0.0710238 0.2845504

Do p-value= 1>0.05 nên chấp nhận gt H0.

Vậy tỉ lệ hộ nghèo ở thành thị 1 là thấp hơn nông thôn 2 .

2,

Gọi V1, V2 lần lượt là Phương sai của chi tiêu giáo dục của tổng thể hộ gia đình ở thành thị và của của tổng thể các hộ gia đình ở nông thôn.

H0 : V1/V2=1 ; H1 : V1/V2 khác 1

```
>x=ChiTieuGiaoDucTrongNam[KhuVuc==1]
>y=ChiTieuGiaoDucTrongNam[KhuVuc==2]
> var.test(x,y,ratio = 1,alternative ="t",conf.level = 0.95)


        F test to compare two variances

data:  x and y
F = 9.3001, num df = 2646, denom df = 6750, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
```

```
 8.731045 9.915083
sample estimates: ratio
of variances
9.300085
```

Do p-value < 2.2e-16<0.05 nên bác bỏ gỉa thuyết H0.

Vậy phương sai khác nhau.

Câu 5:

Câu hỏi 5. *Từ tập dữ liệu trên, hãy tự thiết kế ra ít nhất 3 bài toán kiểm định về tỉ lệ và phương sai hai tổng thể, giải nó và kết luận.*

Bài toán 1 : Từ dữ liệu ChiTieu2010.csv, hãy kiểm định khẳng định sau tại mức ý nghĩa 5%: phương sai chi tiêu cho giáo dục hàng tháng của hai tổng thể hộ nghèo và tổng thể hộ không nghèo là khác nhau.Gỉa sử hai tổng thể đều có phân bố chuẩn

Gọi V1, V2 lần lượt là phương sai chi tiêu giáo dục hàng tháng của tổng thể các hộ nghèo và tổng thể các hộ không nghèo.

H0 : V1/V2=1 ; H1 : V1/V2 khác 1

```
>x=ChiTieuGiaoDucTrongNam[HoNgheo==1]
>y=ChiTieuGiaoDucTrongNam[HoNgheo==0]
> var.test(x,y,ratio = 1,alternative ="t",conf.level = 0.95)
F test to compare two variances

data:  x and y
F = 0.0098301, num df = 2108, denom df = 7288, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.009184761 0.010534662 sample
estimates: ratio of variances
0.009830125
```

Ta có `p-value < 2.2e-16`< 0.05 nên bác bỏ gỉa thuyết H0. Vậy với xác suất sai lầm không quá 5%, ta có thể cho rằng phương sai của hai tổng thể nói trên là khác nhau.

Bài toán 2 : Từ dữ liệu ChiTieu2010.csv, hãy kiểm định khẳng định sau tại mức ý nghĩa 4%: Tỉ lệ chi tiêu cho giáo dục hàng tháng trên 1500 ở tổng thể hộ không nghèo là cao hơn so với tổng thể hộ nghèo.

Gọi p1, p2 lần lượt là tỉ lệ chi tiêu giáo dục hàng tháng trên 1500 của tổng thể các hộ không nghèo và của tổng thể các hộ nghèo.

H0 : P1 − P2 <= 0 ; H1 : P1 − P2 > 0

```
>table(HoNgheo,ChiTieuGiaoDucTrongNam>1500)

HoNgheo FALSE TRUE
0  7070  219
1  2109    0
```

```
> table(HoNgheo)
HoNgheo
   0    1
7289 2109
> x=c(0,219)
> n=c(2109,7289)
> prop.test(x, n,alt= "greater",conf.level = 0.96, correct = TRUE)
```

2-sample test for equality of proportions with continuity correction

data: x out of n
X-squared = 63.564, df = 1, p-value = 1
alternative hypothesis: greater
96 percent confidence interval:
-0.03385151 1.00000000
sample estimates:
prop 1 prop 2
0.00000000 0.03004527

Do p-value = 1 > 0.04 nên chấp nhận gỉa thuyết H0.

Vậy, với mức ý nghĩa 4%, ta không thể cho rằng tỉ lệ chi tiêu cho giáo dục trên 1500 ở tổng thể hộ không nghèo là cao hơn so với tỉ lệ đó ở hộ nghèo.

Bài toán 3: Từ tập dữ liệu ChiTieu2010.csv, với mức ý nghĩa 0.05.

Hãy kiểm định trung bình chi tiêu các hạng mục ChiTieuGiaoDucTrongNam, ChiTieuYTe, ChiTieuKhac có như nhau không?

Gọi muA,muB, muC lần lượt là trung bình tổng thể chi tiêu các hạng mục ChiTieuGiaoDucTrongNam, ChiTieuYTe, ChiTieuKhac

Ta có: H0: muA=muB= muC; H1:Tồn tại i,j thuộc {A,B,C}: mui khác muj

```
> DL=read.csv("ChiTieu2010.csv")
> attach(DL)
> x=ChiTieuGiaoDucTrongNam
> y=ChiTieuYTe
> z=ChiTieuKhac
> MauGop=c(x,y,z)
> length(x)
[1] 9398
> length(y)
[1] 9398
> length(z)
[1] 9398
> PhanLoai=factor(c(rep("A",length(x)),rep("B",length(y)),rep("C",length(z))))
> anova(lm(MauGop ~ PhanLoai))

Analysis of Variance Table

Response: MauGop
             Df     Sum Sq  Mean Sq F value    Pr(>F)
PhanLoai      2   36949797 18474898  61.082 < 2.2e-16 ***
Residuals 28191 8526650007   302460
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Do p-value < 2.2e-16 < 0.05 nên bác bỏ gt H0, có sự khác nhau giữa các trung bình tổng thể.

s