

Kiểm định tỉ lệ, phương sai hai tổng thể

Ngày 22 tháng 12 năm 2020

Họ Tên: Phùng Quang Huy- 1851061400

Câu hỏi 1 (Ôn tập). Dữ liệu ChiTieu2010.csv là mẫu đi đầu 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 đi đầu nội trú nhiều hơn chi tiêu đi đầu 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.

Bai lam :

1.

#Goi mu1,mu2 lan luot la trung binh 2 tong the ngoai tru, noi tru

#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, $n_1 > 30$ va $n_2 > 30$

#H0:mu1-mu2=>0; H1:mu1-mu2<0

```
> x=DieuTriNgoaiTru  
> y=DieuTriNoiTru  
> t.test(x, y, mu=0, alternative="l",var.equal = FALSE, conf.level=0.95)
```

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 less than 0
95 percent confidence interval:
-Inf 4.674444
sample estimates:
mean of x mean of y
89.18773 93.92887

Ta co : p-value= 0.2038>0.05 nen chap nhan H0

– mau doc chon duoc lay Theo doi

2.

```
> x=ChiTieuGiaoDucTrongNam[KhuVuc==1]  
> y=ChiTieuGiaoDucTrongNam[KhuVuc==2]  
> t.test(x, y, mu=0, alternative="greater",var.equal = FALSE, conf.level=0.95)
```

Welch Two Sample t-test

data: x and y
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

Ta co: p-value<2.2e-16<0.05 nen bac bo H0

-Mau duoc lay doc lap

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?

Bai lam

Goi p_1 , p_2 lan luot la ti le sinh vien ke toan va QTKD co vice lam

$H_0: p_1 \leq p_2$ $H_1: p_1 > p_2$

`> prop.test(c(300,200), c(400,300), alternative = "g", conf.level = 0.95,
correct = FALSE)`

2-sample test for equality of proportions without continuity
correction

data: c(300, 200) out of c(400, 300)

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 $p\text{-value} = 0.007863 < 0.05$ nên ta bác bỏ H_0 .

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 đi đầu 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

Bai Lam :

Gọi V_1, V_2 lần lượt là phương sai của tổng thể 2 cặp chơi game và chơi cờ

$H_0 : V_1/V_2 = 1$; $H_1 : V_1/V_2 \text{ khác } 1$

> 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 $p\text{-value} = 0.9651 > 0.05$ nên chấp nhận gt H_0 , có thể xem 2 PS của các tổng thể như nhau.

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

1. Tỷ lệ hộ nghèo ở nông thôn là cao hơn thành thị.

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

```
HoNgheo FALSE TRUE
0  4830 2459
1  1921  188
```

#Ta tìm đc x1=188 hộ nghèo ở KV1 TT

```
> table(HoNgheo,KhuVuc==2)
```

```
HoNgheo FALSE TRUE
0  2459 4830
1   188 1921
```

#Ta tìm đc x2=1921 hộ nghèo ở KV2 NT

```
> table(KhuVuc)
KhuVuc
 1    2
2647 6751
```

#Ta tìm đc $n_1=2647$, $n_2=6751$

```
> x=c(188,1921)
> n=c(2647,6751)
> prop.test(x,n,alt="less",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 < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
 -1.000000 -0.201319
sample estimates:
 prop 1    prop 2 
0.0710238 0.2845504
```

Do $p\text{-value} < 2.2e-16 < 0.05$ nên bác bỏ H_0 .

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

2. Phương sai của chỉ 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 chỉ tiêu cho giáo dục của hai tổng thể đều có phân bố chuẩn.

```
# Gọi V1, V2 lần lượt là Phương sai của chỉ 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 $p\text{-value} < 2.2e-16 < 0.05$ nên bác bỏ H_0 .
Vậy phương sai khác nhau.

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.

Dữ liệu ChiTieu2010.csv là mẫu đi 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 CTAnUongDipLeTrongNam nhiều hơn CTSinhHoatNgoaiAnUongTrongThang. Mẫu được chọn là theo đôi hay độc lập?

$H_0: x \leq y$ $H_1: x > y$

$> x = \text{CTAnUongDipLeTrongNam}$

$> y = \text{CTSinhHoatNgoaiAnUongTrongThang}$

$> \text{t.test}(x, y, \text{mu}=0, \text{alternative}="g", \text{var.equal} = \text{FALSE}, \text{conf.level}=0.95)$

Welch Two Sample t-test

data: x and y

$t = -70.982$, $df = 10163$, $p\text{-value} = 1$

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-477.469 Inf

sample estimates:

mean of x mean of y

164.1279 630.7821

Do $p\text{-value} = 1 > 0.05$ nên chấp nhận H_0

2. Kiểm định khẳng định cho rằng SoNguoiTrongHo 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?

$H_0 : x \leq y$ $H_1: x > y$

```
> x=SoNguoiTrongHo[KhuVuc==1]
```

```
> y=SoNguoiTrongHo[KhuVuc==2]
```

```
> t.test(x, y, mu=0, alternative="greater", var.equal = FALSE,  
conf.level=0.95)
```

Welch Two Sample t-test

data: x and y

$t = -4.7036$, $df = 5260.4$, $p\text{-value} = 1$

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

-0.2190206 Inf

sample estimates:

mean of x mean of y

3.820552 3.982817

Do p-value = 1 > 0.05 nên chấp nhận H_0

3. Tỷ lệ Giới Tính nam ở nông thôn là cao hơn thành thị.

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

| | GioiTinh FALSE | GioiTinh TRUE |
|---|----------------|---------------|
| 0 | 1406 | 920 |
| 1 | 5345 | 1727 |

```
> table(GioiTinh,KhuVuc==2)
```

| | GioiTinh FALSE | GioiTinh TRUE |
|---|----------------|---------------|
| 0 | 920 | 1406 |
| 1 | 1727 | 5345 |

```
> prop.test(c(1727,5345),c(2326,7072),alt="g",conf.level =  
0.95,correct=FALSE)
```

2-sample test for equality of proportions without continuity correction

data: c(1727, 5345) out of c(2326, 7072)

X-squared = 1.6677, df = 1, p-value = 0.9017

alternative hypothesis: greater

95 percent confidence interval:

-0.03043885 1.00000000

sample estimates:

prop 1 prop 2

0.7424764 0.7557975

Do co p-value = 0.9017 > 0.05 nen chap nhan H0