BÀI TẬP ƯỚC LƯỢNG + KIỂM ĐỊNH CHO 2 TỔNG THỂ, PHÂN TÍCH PHƯƠNG SAI

Bài toán 1: Kiểm định, tìm KTC cho hiệu 2 trung bình, khi σ đã biết: sử dụng hàm zsum.test hoặc z.test

Chú ý: **+ Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu sơ cấp): Sử dụng hàm `z.test`**

+ Dữ liệu dạng thứ cấp (thu gọn, đã qua xử lí): Sử dụng hàm`zsum.test`

Usage

```
zsum.test(mean.x, sigma.x = NULL, n.x = NULL, mean.y = NULL,
  sigma.y = NULL, n.y = NULL, alt="t", mu = 0,
  conf.level = 0.95)
```

trong đó:

alt="t" (two-side): kiểm định 2 phía và cho ước lượng khoảng

alt="g" (greater): kiểm định lớn hơn

alt="l" (less): kiểm định nhỏ hơn

Ví dụ : Một mẫu ngẫu nhiên $n_1 = 25$ lấy từ tổng thể có phân phối chuẩn với độ lệch chuẩn là $\sigma_1 = 5$, có giá trị trung bình $\overline{x}_1 = 80$. Một mẫu ngẫu nhiên thứ hai $n_2 = 6$ lấy từ tổng thể có phân phối chuẩn với độ lệch chuẩn là $\sigma_2 = 3$, có giá trị trung bình $\overline{x}_2 = 75$.

a)Kiểm định giả thiết rằng không có sự sai khác về chất lượng giữa hai tổng thể, với mức ý nghĩa 0,05.

b)Tìm khoảng tin cậy 95% cho hiệu 2 trung bình của tổng thể.

GIẢI:

- Gọi μ_1, μ_2 là giá trị trung bình của hai tổng thể 1, 2 tương ứng
- Từ giả thiết:

$$\begin{cases} n_1 = 25 \\ \overline{x}_1 = 80 \\ \sigma_1 = 5 \end{cases}; \qquad \begin{cases} n_2 = 6 \\ \overline{x}_2 = 75 \\ \sigma_2 = 3 \end{cases}; \qquad \alpha = 0,05$$

a) Đây là bài toán kiểm định hiệu hai giá trị trung bình khi σ_1^2, σ_2^2 đã biết.

Đặt bài toán: $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$

+Sử dụng hàm zsum.test trong R

```
> zsum.test(mean.x=80, sigma.x=5, n.x=25, mean.y=75, sigma.y=3,
n.y=6,mu=0,conf.level=0.95)

        Two-sample z-Test

data:  Summarized x and y
z = 3.1623, p-value = 0.001565
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.901025 8.098975
sample estimates:
mean of x mean of y
       80        75
```

p-value = 0.001565 < mức ý nghĩa alpha=0.05 nên bác bỏ giả thiết H_0

Vậy có thể xem có sự sai khác về chất lượng giữa hai tổng thể, với mức ý nghĩa 0,05.

b) Khoảng tin cậy 95% cho hiệu 2 trung bình của tổng thể:

```
>zsum.test(mean.x=80,sigma.x=5,n.x=25,mean.y=75, sigma.y=3, n.y=6,conf.level=0.95)
```

Ta được:

```
95 percent confidence interval:
 1.901025 8.098975
```

Example

Một trong những chỉ tiêu so sánh chất lượng phục vụ bay là thời gian delay bay trung bình của hãng đó. Giả sử ta có dữ liệu thời gian (giờ) delay của một số chuyến bay của hãng A (20 chuyến), B(25 chuyến) như sau:

A 3.5 2.8 2.3 2.8 1.4 1.7 2.8 3.9 3.6 4.0 0.4 1.5 1.1 2.3 4.3 2.9 2.0 4.3 2.1 2.6

B 2.7 4.0 5.5 2.6 5.6 0.9 4.3 3.8 4.0 4.6 3.5 4.8 2.5 2.6 3.7 5.3 4.0 2.9 2.9 3.8 5.6 2.7 3.1 5.7 5.0

Giả sử biết độ lệch chuẩn của thời gian hoãn bay của hai hãng A, B lần lượt là 1 và 1.5 (giờ) và cả hai tổng thể đều có phân phối chuẩn.
Hỏi, tại mức ý nghĩa 5% thời gian hoãn bay trung bình của hai hãng có như nhau không?

A: 3.5 2.8 2.3 2.8 1.4 1.7 2.8 3.9 3.6 4.0 0.4 1.5 1.1 2.3 4.3 2.9 2.0 4.3 2.1 2.6

B: 2.7 4.0 5.5 2.6 5.6 0.9 4.3 3.8 4.0 4.6 3.5 4.8 2.5 2.6 3.7 5.3 4.0 2.9 2.9 3.8 5.6 2.7 3.1 5.7 5.0

```
#Goi muA,muB lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai da biet
#H0:muA-muB=0; H1:muA-muB#0

> A=scan()
1: 3.5 2.8 2.3 2.8 1.4 1.7 2.8 3.9 3.6 4.0 0.4 1.5 1.1 2.3 4.3 2.9 2.0 4.3
19: 2.1 2.6
21:
```

```
Read 20 items
> B=scan()
1: 2.7 4.0 5.5 2.6 5.6 0.9 4.3 3.8 4.0 4.6 3.5 4.8 2.5 2.6 3.7 5.3 4.0 2.9
19: 2.9 3.8 5.6 2.7 3.1 5.7 5.0
26:
Read 25 items
> library(BSDA)
Loading required package: lattice

Attaching package: 'BSDA'

The following object is masked from 'package:datasets':

    Orange
> z.test(A,B,alt="t",mu=0,sigma.x=1,sigma.y=1.5,conf.level = 0.95)

        Two-sample z-Test

data:  A and B
z = -3.2846, p-value = 0.001021
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9623514 -0.4956486
sample estimates:
mean of x mean of y
    2.615     3.844
```

Do `p-value = 0.001021<alpha=0.05,` ta có thể bác bỏ gt `H0.`

Vậy, thời gian hoãn bay trung bình của hai hãng là khác nhau.

Bài toán 2: Kiểm định, tìm KTC về hiệu 2 trung bình, khi σ chưa biết, cỡ mẫu lớn ($n \geq 30$):

+ Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu sơ cấp): sử dụng hàm *t.test*, dữ liệu cho dạng thứ cấp: sử dụng hàm *tsum.test*.

Chú ý: Vì cỡ mẫu lớn ($n \geq 30$) nên ta có thể xấp xỉ chuẩn coi $s \approx \sigma$ và đưa về như trường hợp σ đã biết(Trường hợp 1), do đó ta sử dụng các hàm *zsum.test*, *z.test* thì kết quả gần như xấp xỉ, do bậc tự do lớn nên kiểm định z hay sử dụng nếu không bằng phần mềm. Ta ưu tiên sử dụng hàm *t.test, tsum.test trên R hơn vì không gặp trở ngại khi bậc tự do lớn.*

Example

Nhiều ý kiến cho rằng lương của phụ nữ thấp hơn lương nam giới. Để kiểm định điều này, người ta tiến hành điều tra 100 nam giới thì thấy lương trung bình là 7 (triệu/tháng) với độ lệch chuẩn là 2, điều tra 90 phụ nữ thấy lương trung bình là 6.3, độ lệch chuẩn là 1.5. Ở mức ý nghĩa $\alpha = 5\%$ hãy hãy kiểm định ý kiến trên.

```
#Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, n1>30
va n2>30
#H0:mu1-mu2=0; H1:mu1-mu2>0
> tsum.test(mean.x=7, s.x = 2, n.x = 100, mean.y = 6.3,s.y = 1.5, n.y = 90,
alternative = "g", mu = 0,conf.level = 0.95)

        Welch Modified Two-Sample t-Test

data:  Summarized x and y
t = 2.7456, df = 182.24, p-value = 0.003322
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.2785003        NA
sample estimates:
mean of x mean of y
      7.0       6.3
```

Do `p-value = 0.00302 < alpha=0.05,` ta có thể bác bỏ gt `H0.`

Vậy, ở mức ý nghĩa 5%, lương trung bình của nam giới cao hơn lương trung bình của nữ giới.

Nhận xét: Nếu sử dụng hàm zsum.test cho kết quả tương tự

```
> zsum.test(mean.x=7, sigma.x = 2, n.x = 100, mean.y = 6.3,
+           sigma.y = 1.5, n.y = 90, alternative = "g", mu = 0,
+           conf.level = 0.95)

        Two-sample z-Test

data:  Summarized x and y
z = 2.7456, p-value = 0.00302
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.280643       NA
sample estimates:
mean of x mean of y
      7.0       6.3
```

Do `p-value = 0.00302 < alpha=0.05,` ta có thể bác bỏ gt `H0.`

Vậy, ở mức ý nghĩa 5%, lương trung bình của nam giới cao hơn lương trung bình của nữ giới.

Example

Trong file "ChiTieu2010.csv" là một số thông tin về chi tiêu của một mẫu những hộ gia đình ở miền bắc, đơn vị tiền (chi tiêu) ở đây là nghìn đồng. Tại mức ý nghĩa 5%:

1. Kiểm định khẳng định cho rằng trung bình các hộ diện nghèo tiêu cho y tế ít hơn các hộ không nghèo.
2. Kiểm định nhận định rằng các hộ gia đình dùng nhiều tiền hơn cho giáo dục so với cho chăm sóc y tế.
3. có thể khẳng định được rằng trung bình chi tiêu cho ăn uống nhiều hơn 2000 (nghìn) so với chi tiêu ngoài ăn uống mỗi tháng không?
4. có thể khẳng định chi tiêu ngoại trú và nội trú là như nhau không?

```
1) #Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, n1>30
va n2>30
#H0:mu1-mu2=0; H1:mu1-mu2<0
> DL=read.csv("ChiTieu2010.csv")
> attach(DL)
The following objects are masked from DL (pos = 3):

    ChiTieuGiaoDucTrongNam, ChiTieuKhac, ChiTieuTaiSanKhongHaoMonTrong10Nam,
    ChiTieuYTe, CTAnUongDipLeTrongNam, CTAnUongTrongThang,
    CTSinhHoatNgoaiAnUongTrongThang, CuaCaiGiaTriTrongNam, CuaCaiTrongNam,
    DieuTriNgoaiTru, DieuTriNoiTru, GioiTinh, HoNgheo, Huyen, ï..Tinh, KhuVuc,
    SoNguoiTrongHo, ThietBiYTe, ThueNhaDienNuocTrongNam, Thuoc, TongChiTieu,
    Tuoi, Xa

> x=ChiTieuYTe[HoNgheo==1]
> y=ChiTieuYTe[HoNgheo==0]

#Cách 1: SD hàm t.test
> t.test(x, y, mu=0, alternative="less",var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  x and y
t = -13.563, df = 6744.1, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
    -Inf -129.13
sample estimates:
mean of x mean of y
 127.7036  274.6582
```

Do p-value < 2.2e-16<0.05 nên bác bỏ gt H0

```
#Cách2: SD hàm z.test
> z.test(x, y, alternative = "less", mu = 0, sigma.x=sd(x),sigma.y = sd(y),
conf.level = 0.95)

        Two-sample z-Test

data:  x and y
z = -13.563, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        NA -129.1324
sample estimates:
mean of x mean of y
 127.7036  274.6582
```

Do p-value < 2.2e-16<0.05 nên bác bỏ gt H0

```
2) #Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, n1>30
va n2>30
#H0:mu1-mu2=0; H1:mu1-mu2>0
> DL=read.csv("ChiTieu2010.csv")
> attach(DL)
> a=ChiTieuGiaoDucTrongNam
> b=ChiTieuYTe

#Cách 1: SD hàm t.test
> t.test(a, b, mu=0, alternative="greater",var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  a and b
t = -0.15786, df = 18292, p-value = 0.5627
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -17.55802       Inf
sample estimates:
mean of x mean of y
 240.1428  241.6802
```

Do p-value = 0.5627 > 0.05 nên chấp nhận gt H0

```
#Cách 2: SD hàm z.test
> z.test(a, b, alternative = "greater", mu = 0, sigma.x=sd(a),sigma.y = sd(b),
conf.level = 0.95)

        Two-sample z-Test

data:  a and b
z = -0.15786, p-value = 0.5627
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -17.55721        NA
sample estimates:
mean of x mean of y
 240.1428  241.6802
```

Do p-value = 0.5627 > 0.05 nên chấp nhận gt H0

```
3) #Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, n1>30
va n2>30
#H0:mu1-mu2=2000; H1:mu1-mu2>2000
> c=CTAnUongTrongThang
> d=CTSinhHoatNgoaiAnUongTrongThang

#Cách 1: SD hàm t.test
> t.test(c, d, mu=2000, alternative="greater",var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  c and d
t = -32.664, df = 12667, p-value = 1
alternative hypothesis: true difference in means is greater than 2000
95 percent confidence interval:
 1433.41     Inf
sample estimates:
mean of x mean of y
2091.3574  630.7821

#Cách 2: SD hàm z.test
> z.test(c, d, alternative = "greater", mu = 2000, sigma.x=sd(c),sigma.y = sd(d),
conf.level = 0.95)

        Two-sample z-Test

data:  c and d
z = -32.664, p-value = 1
alternative hypothesis: true difference in means is greater than 2000
95 percent confidence interval:
 1433.412       NA
sample estimates:
mean of x mean of y
2091.3574  630.7821


4) #Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, n1>30
va n2>30
#H0:mu1-mu2=0; H1:mu1-mu2#0

> e=DieuTriNgoaiTru
> f=DieuTriNoiTru

#Cách 1: SD hàm t.test
> t.test(e, f, mu=0, alternative="t",var.equal = FALSE, conf.level=0.95)

        Welch Two Sample t-test

data:  e and f
t = -0.8283, df = 15914, p-value = 0.4075
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.960684   6.478419
sample estimates:
```

```
mean of x mean of y
 89.18773  93.92887

#Cách 2: SD hàm z.test
> z.test(e, f, alternative = "t", mu = 0, sigma.x=sd(e),sigma.y = sd(f),
conf.level = 0.95)

        Two-sample z-Test

data:  e and f
z = -0.8283, p-value = 0.4075
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.959831   6.477565
sample estimates:
mean of x mean of y
 89.18773  93.92887
```

Example (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?

HD:

1)Làm như VD trên ý 4).

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
```

p-value < 2.2e-16<0.05 nên bác bỏ gt H0.

Bài toán 3.Kiểm định, tìm KTC về hiệu 2 trung bình, khi σ chưa biết, cỡ mẫu nhỏ (n < 30): sử dụng hàm t.test hoặc tsum.test (giả thiết phân phối là chuẩn)

Usage

```
tsum.test(mean.x, s.x = NULL, n.x = NULL, mean.y = NULL, s.y = NULL,
  n.y = NULL, alternative = "two.sided", mu = 0, var.equal = FALSE,
  conf.level = 0.95)
```

trong đó:

alt="t" : kiểm định 2 phía và cho KTC

alt="g" : kiểm định lớn hơn

alt="l" : kiểm định nhỏ hơn

Chú ý: + Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu sơ cấp): sử dụng hàm `t.test`

+ Nếu dữ liệu cho ở dạng số liệu quan sát (dữ liệu thứ cấp): sử dụng hàm `tsum.test`

+ Nếu phương sai bằng nhau chưa biết ta chỉ thay đổi một thông số trong hàm t với var.equal=TRUE

+ Nếu phương sai khác nhau ta để mặc định (var.equal=FALSE)

Ví dụ 1(Phương sai bằng nhau chưa biết): Để so sánh mức độ mài mòn của hai loại kim loại khác nhau, người ta lấy 12 miếng loại 1 và 10 miếng loại 2. Mẫu ứng với kim loại 1 có trung bình mài mòn là 85 đơn vị, với độ lệch chuẩn mẫu bằng 4; trong khi mẫu ứng với kim loại 2 có trung bình là 81 và độ lệch chuẩn mẫu là 5. Có thể kết luận, với mức ý nghĩa 0,05, rằng hiệu số trung bình mức độ mài mòn của kim loại 1 và kim loại 2 là hơn 2 đơn vị được không? Giả sử các mật độ đều xấp xỉ chuẩn với **phương sai bằng nhau.**

Nhận xét: Do phương sai bằng nhau ta chỉ thay đổi một thông số trong hàm t với var.equal=TRUE

Giải:

- Đặt μ_1, μ_2 là kỳ vọng cho độ mài mòn của hai kim loại 1 và 2

- Từ giả thiết:

$$\begin{cases} n_1 = 12, \\ \overline{x}_1 = 85, \\ s_1 = 4, \end{cases} \qquad \begin{cases} n_2 = 10, \\ \overline{x}_1 = 81, \\ s_1 = 5, \end{cases} \quad \alpha = 0,05$$

- Đây là bài toán kiểm định hiệu hai giá trị trung bình, với phương sai bằng nhau chưa biết

Đặt bài toán: $\begin{cases} H_0 : \mu_1 - \mu_2 = 2 \\ H_1 : \mu_1 - \mu_2 > 2 \end{cases}$.

```
> tsum.test(mean.x=85, s.x=4, n.x=12, mean.y=81, s.y=5, n.y=10, mu=2,
alternative="greater",var.equal = TRUE,conf.level=0.95)

        Standard Two-Sample t-Test
```

```
data:  Summarized x and y
t = 1.0432, df = 20, p-value = 0.1547
alternative hypothesis: true difference in means is greater than 2
95 percent confidence interval:
 0.6932903        NA
sample estimates:
mean of x mean of y
       85        81
```

Kết luận: Do `p-value = 0.1547 > alpha = 0.05` nên chấp nhận H_0. Ta không thể kết luận rằng mức độ mài mòn của kim loại 1 hơn kim loại 2 là 2 đơn vị.

Ví dụ 2(Phương sai khác nhau chưa biết): Để so sánh mức độ mài mòn của hai loại kim loại khác nhau, người ta lấy 12 miếng loại 1 và 10 miếng loại 2. Mẫu ứng với kim loại 1 có trung bình mài mòn là 85 đơn vị, với độ lệch chuẩn mẫu bằng 4; trong khi mẫu ứng với kim loại 2 có trung bình là 81 và độ lệch chuẩn mẫu là 5. Có thể kết luận, với mức ý nghĩa 0,05, rằng hiệu số trung bình mức độ mài mòn của kim loại 1 và kim loại 2 là hơn 2 đơn vị được không? Giả sử các mật độ đều xấp xỉ chuẩn.

Nhận xét: Do phương sai khác nhau ta để mặc định (var.equal=FALSE)

Giải:

- Đặt μ_1, μ_2 là kỳ vọng cho độ mài mòn của hai kim loại 1 và 2

- Từ giả thiết:

$$\begin{cases} n_1 = 12, \\ \overline{x}_1 = 85, \\ s_1 = 4, \end{cases} \qquad \begin{cases} n_2 = 10, \\ \overline{x}_1 = 81, \\ s_1 = 5, \end{cases} \quad \alpha = 0,05$$

- Đây là bài toán kiểm định hiệu hai giá trị trung bình, với phương sai chưa biết

Đặt bài toán: $\begin{cases} H_0 : \mu_1 - \mu_2 = 2 \\ H_1 : \mu_1 - \mu_2 > 2 \end{cases}$.

```
> tsum.test(mean.x=85, s.x=4, n.x=12, mean.y=81, s.y=5, n.y=10, mu=2, alternative="greater", conf.level=0.95)
```

Hoặc

```
>tsum.test(mean.x=85, s.x=4, n.x=12, mean.y=81, s.y=5, n.y=10, mu=2, alternative="greater",var.equal = FALSE,conf.level=0.95)


        Welch Modified Two-Sample t-Test


data:  Summarized x and y
t = 1.0215, df = 17.165, p-value = 0.1606
alternative hypothesis: true difference in means is greater than 2
95 percent confidence interval:
 0.5959261        NA
sample estimates:
```

```
mean of x mean of y
       85        81
```

Kết luận: Do `p-value = 0.1606 > alpha = 0.05 nên` chấp nhận H_0. Ta không thể kết luận rằng mức độ mài mòn của kim loại 1 hơn kim loại 2 là 2 đơn vị.

Ví dụ 3(dữ liệu sơ cấp): Một nghiên cứu được thực hiện bởi Trung tâm Thủy lợi và được phân tích bởi Trung tâm Thống kê, thuộc Đại học Virginia, nhằm so sánh hai thiết bị xử lý nước thải. Thiết bị A được đặt ở vùng dân cư có thu nhập trung bình thấp. Thiết bị B được đặt ở vùng dân cư có thu nhập trung bình cao. Lượng nước thải được xử lý bởi mỗi thiết bị (tính theo nghìn ga-lông/ ngày) được đo trong 10 ngày như sau:

Thiết bị A: 21, 19, 20, 23, 22, 28, 32, 19, 13, 18

Thiết bị B: 20, 39, 24, 33, 30, 28, 30, 22, 33, 24

Với mức ý nghĩa 5%, có thể kết luận rằng có sự khác nhau giữa lượng nước thải trung bình được xử lý ở vùng có thu nhập thấp và vùng có thu nhập cao không. Giả sử các mật độ đều xấp xỉ chuẩn với **phương sai bằng nhau.**

Giải:

+Nhập dữ liệu:

```
> x=c(21,   19,   20,   23,   22,   28,   32,   19,   13,   18)
> y=c(20,   39,   24,   33,   30,   28,   30,   22,   33,   24)
> t.test(x, y, mu=0, alternative="t",var.equal = TRUE, conf.level=0.95)

        Two Sample t-test

data:  x and y
t = -2.7149, df = 18, p-value = 0.01419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.0621  -1.5379
sample estimates:
mean of x mean of y
     21.5      28.3
```

KL: Do `p-value = 0.01419 < 0.05 nên bác bỏ` H_0. Với mức ý nghĩa 5%, có thể kết luận rằng có sự khác nhau giữa lượng nước thải trung bình được xử lý ở vùng có thu nhập thấp và vùng có thu nhập cao.

Example

Một chủ chuỗi cửa hàng thời trang thử nghiệm để so sánh hiệu quả hai hình thức khuyến mãi tại 20 cửa hàng của mình. Nhóm 10 cửa hàng thứ nhất chạy khuyến mãi theo hình thức mua 1 tặng 1. Nhóm thứ hai theo hình thức giảm giá 50%. Sau một tuần, lợi nhuận (triệu đồng) tại 20 cửa hàng trên như sau:

- **Nhóm thứ nhất:**

  7 10 9 8 6 12 10 7 10 7

- **Nhóm thứ hai:**

  9 13 11 7 10 12 8 10 11 8

Giả sử rằng hai tổng thể tuân theo phân phối chuẩn với phương sai như nhau. Kiểm định sự khác biệt về hiệu quả của hai hình thức khuyến mãi trên. Chọn mức ý nghĩa 5%.

Nhóm 1: 7 10 9 8 6 12 10 7 10 7

Nhóm 2: 9 13 11 7 10 12 8 10 11 8

```
#Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai bang nhau chua
biet (co mau nho)
#H0:mu1-mu2=0; H1:mu1-mu2#0
#Chon var.equal=TRUE
```

```
> T=scan()
1: 7 10  9  8  6 12 10  7 10  7
11:
Read 10 items
> G=scan()
1: 9 13 11  7 10 12  8 10 11  8
11:
Read 10 items
> t.test(T,G,alt="t",mu=0,var.equal = TRUE)

        Two Sample t-test

data:  T and G
t = -1.5262, df = 18, p-value = 0.1443
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0895559  0.4895559
sample estimates:
mean of x mean of y
      8.6       9.9
```

Ta được P - giá trị $= 0.1443 > 0.05$ nên chấp nhận H_0.

Example

Một chủ chuỗi cửa hàng thời trang thử nghiệm để so sánh hiệu quả hai hình thức khuyến mãi tại 20 cửa hàng của mình. Nhóm 10 cửa hàng thứ nhất chạy khuyến mãi theo hình thức mua 1 tặng 1. Nhóm thứ hai theo hình thức giảm giá 50%. Sau một tuần, lợi nhuận (triệu đồng) tại 20 cửa hàng trên như sau:

- **Nhóm thứ nhất:**

  7 10 9 8 6 12 10 7 10 7

- **Nhóm thứ hai:**

  9 13 11 7 10 12 8 10 11 8

Giả sử rằng hai tổng thể tuân theo phân phối chuẩn. Kiểm định sự khác biệt về hiệu quả của hai hình thức khuyến mãi trên trong hai trường hợp:

1. Phương sai hai tổng thể là như nhau.
2. Chưa có thông tin gì về phương sai hai tổng thể.

Chọn mức ý nghĩa 5%.

Trường hợp 2: Chưa có thông tin gì về phương sai hai tổng thể.

```
#Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho hieu 2 trung binh, phuong sai chua biet, co mau
nho (<30)
```

```
#H0:mu1-mu2=0; H1:mu1-mu2#0
#Chon var.equal=FALSE
```

```
> t.test(T,G,alt="t",mu=0,var.equal = FALSE)

        Welch Two Sample t-test

data:  T and G
t = -1.5262, df = 17.999, p-value = 0.1443
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0895634  0.4895634
sample estimates:
mean of x mean of y
      8.6       9.9
```

Ta được P - giá trị $= 0.1443 > 0.05$ nên chấp nhận H_0.

Bài toán 4: Kiểm định giả thiết và tìm Khoảng tin cậy cho mẫu theo đôi(quan sát cặp đôi, 2 mẫu không độc lập):

- phân phối chuẩn, mẫu chọn theo đôi:

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt, paired = TRUE)$$

Example

Để đánh giá một chương trình xóa đói giảm nghèo ở một vùng miền núi người ta chọn ngẫu nhiên 20 xã và thống kê tỉ lệ hộ nghèo thời điểm trước khi tiến hành chương trình, sau vài năm hoàn thành chương trình họ lại đến 20 xã trên và thống kê lại tỉ lệ hộ nghèo (%), số liệu cho bởi bảng dưới đây.
Giả sử tỉ lệ hộ nghèo của tổng thể các xã tuân theo phân phối chuẩn.
Kiểm định ở mức ý nghĩa 5% rằng trung bình tỉ lệ hộ nghèo giảm ít nhất 3% sau khi thực hiên chương trình trên.

T	8	10	11	10	10	12	8	11	10	14	9	9	8	9	9	7	8	17	8	12
S	8	8	5	6	7	6	8	3	6	8	7	8	10	8	5	2	6	5	9	8

T: 8 10 11 10 10 12 8 11 10 14 9 9 8 9 9 7 8 17 8 12
S: 8 8 5 6 7 6 8 3 6 8 7 8 10 8 5 2 6 5 9 8

```
#Goi mu1,mu2 lan luot la trung binh 2 tong the
#Bai toan kiem dinh gia thiet cho mẫu theo đôi(quan sát cặp đôi, 2 mẫu ko độc lập)
```

```
#H0:mu1-mu2=3; H1:mu1-mu2<3
#Chon paird=TRUE
> T=scan()
1: 8 10 11 10 10 12 8 11 10 14 9 9 8 9 9 7 8 17 8 12
21:
Read 20 items
> S=scan()
1: 8 8 5 6 7 6 8 3 6 8 7 8 10 8 5 2 6 5 9 8
21:
Read 20 items
> t.test(T,S,alt = "less", mu = 3,paired = TRUE, conf.level = 0.95)

        Paired t-test

data:  T and S
t = 0.47245, df = 19, p-value = 0.679
alternative hypothesis: true difference in means is less than 3
95 percent confidence interval:
     -Inf 4.630978
sample estimates:
mean of the differences
                  3.35
```

Do p-value = 0.679 > 0.05 nên chấp nhận gt H0.

Bài toán 5: Kiểm định giả thiết và tìm Khoảng tin cậy cho hiệu 2 tỷ lệ, cỡ mẫu lớn

Usage

```
+prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
+prop.test(c(x1,x2), c(n1,n2), p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

Trong đó:

alt="t" : kiểm định 2 phía và cho ước lượng KTC

alt="g" : kiểm định lớn hơn

alt="l" : kiểm định nhỏ hơn

correct: tham số dạng logic chỉ xem có hay không sự điều chỉnh liên tục Yate, mặc định là correct = TRUE.

Nếu $5 \le x_1 \le n_1 - 5$ và $5 \le x_2 \le n_2 - 5$, chọn correct = FALSE.

(Hoặc là: Nếu $n\hat{p} \ge 5$ và $n(1-\hat{p}) \ge 5$, chọn correct = FALSE).

Ví dụ : Một cuộc bỏ phiếu được đưa ta để xác định vị trí xây dựng một nhà máy hóa chất ở trong thành phố hay ở ngoại ô. Có 120 trên 200 cử tri trong thành phố đồng ý cho xây dựng trong thành phố và 240 trên 500 cử tri ở ngoại ô đồng ý với đề xuất này. Liệu có thể cho rằng tỷ lệ cử tri trong thành phố và ngoại ô đồng ý với đề xuất này là như nhau không? Sử dụng mức ý nghĩa 0,025?

GIẢI:

- Gọi p là tỷ lệ cử tri trong thành phố và ngoại ô đồng ý.
- Từ giả thiết: $n_1 = 200, x_1 = 120, n_2 = 500, x_2 = 240, \alpha = 0.025$
- Đây là bài toán kiểm định về sự bằng nhau giữa hai tỷ lệ với cỡ mẫu lớn

Do $n_1\hat{p}_1 = 120 \geq 5$ và $n_1(1-\hat{p}_1) = 80 \geq 5$ và $n_2\hat{p}_2 = 240 \geq 5$ và $n_2(1-\hat{p}_2) = 260 \geq 5$, chọn correct = FALSE

Đặt bài toán: $\begin{cases} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \end{cases}$

```
> prop.test(c(120,240), c(200,500), alternative = "t",conf.level = 0.975, correct = F)

        2-sample test for equality of proportions without continuity correction

data:  c(120, 240) out of c(200, 500)
X-squared = 8.2353, df = 1, p-value = 0.004108
alternative hypothesis: two.sided
97.5 percent confidence interval:
 0.02760635 0.21239365
sample estimates:
prop 1 prop 2
  0.60   0.48
```

Kết luận: Do p-value = 0.004108 < 0.025 nên bác bỏ H_0. Với mức ý nghĩa 0,025, ta bác bỏ giả thiết tức là có thể cho rằng tỷ lệ cử tri trong và ngoài thị trấn đồng ý là không bằng nhau.

Example

Để so sánh tỉ lệ sinh viên ra trường kiếm được việc đúng ngành đào tạo của hai trường A và B, người ta chọn ngẫu nhiên 250 sinh viên tốt nghiệp trường A thấy có 130 người đang làm việc đúng chuyên ngành được đào tạo; đối với trường B người ta chọn 300 sinh viên đã tốt nghiệp và thấy có 145 sinh viên đang có công việc đúng chuyên ngành được đào tạo. Hỏi với mức ý nghĩa 5% liệu có thể nói rằng tỉ lệ sinh viên ra trường có việc làm đúng chuyên ngành được đào tạo của trường A có thấp hơn trường B không?

```
#Goi p1,p2 lan luot la tỉ lệ sinh viên ra trường kiếm được việc đúng ngành đào tạo
Của hai trường A và B
#Bai toan kiem dinh gia thiet cho hieu 2 ti le, cỡ mẫu lớn
#H0:p1-p2=0; H1:p1-p2<0
#(Do 5 ≤ x1 ≤ n1 – 5, 5 ≤ x2 ≤ n2 – 5
# hay 5 ≤ 130 ≤ 250 – 5, 5 ≤ 145 ≤ 300 – 5
#thỏa mãn nên ko cần hiệu chỉnh liên tục, chọn correct=FALSE)

> x=c(130,145)
```

```
> n=c(250,300)
> prop.test(x, n, alternative = "less",conf.level = 0.95, correct = FALSE)

        2-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 0.73333, df = 1, p-value = 0.8041
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000  0.1070466
sample estimates:
   prop 1    prop 2
0.5200000 0.4833333
```

Với p-value = 0.8041 > 0.05 nên ta chấp nhận gt H0.

Example

Từ dữ liệu ChiTieu2010.csv, hãy kiểm định tại mức ý nghĩa 5% cho khẳng định tỉ lệ chi tiêu cho ăn uống hàng tháng trên 1000 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 ăn uống hàng tháng trên 1000 của tổng thể các hộ không nghèo và của tổng thể các hộ nghèo.

```
#Bai toan kiem dinh gia thiet cho hieu 2 ti le, cỡ mẫu lớn
```

#H0 : P1 − P2 = 0 ; H1 : P1 − P2 > 0

```
> DL=read.csv("ChiTieu2010.csv")
> attach(DL)
The following objects are masked from DL (pos = 3):

    ChiTieuGiaoDucTrongNam, ChiTieuKhac, ChiTieuTaiSanKhongHaoMonTrong10Nam,
    ChiTieuYTe, CTAnUongDipLeTrongNam, CTAnUongTrongThang,
    CTSinhHoatNgoaiAnUongTrongThang, CuaCaiGiaTriTrongNam, CuaCaiTrongNam,
    DieuTriNgoaiTru, DieuTriNoiTru, GioiTinh, HoNgheo, Huyen, ï..Tinh, KhuVuc,
    SoNguoiTrongHo, ThietBiYTe, ThueNhaDienNuocTrongNam, Thuoc, TongChiTieu,
    Tuoi, Xa

> table(HoNgheo,CTAnUongTrongThang>1000)

HoNgheo FALSE TRUE
      0   411 6878
      1  1162  947
> #Ta tim duoc trong cac ho khong ngheo co x1=6878 ho CTAnUongTrongThang>1000
> #va trong cac ho ngheo co x2=947 ho CTAnUongTrongThang>1000
> table(HoNgheo)
HoNgheo
   0    1
7289 2109
> #Ta tìm đc có tất cả n1=7289 hộ không nghèo và n2=2109 hộ nghèo
> #Ta có 5 ≤ 6878 ≤ 7289 − 5, 5 ≤ 947 ≤ 2109 thỏa mãn ko cần hiệu chỉnh lt > >
#nên chọn correct=FALSE
> x=c(6878,947)
```

```
> n=c(7289,2109)
> prop.test(x, n,alt= "greater",conf.level = 0.95, correct = FALSE)

        2-sample test for equality of proportions without continuity correction

data:  x out of n
X-squared = 2871.1, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.4762246 1.0000000
sample estimates:
   prop 1    prop 2
0.9436137 0.4490280
```

$\Rightarrow p-value < 2.2x10^{-16} < 0.05$ nên bác bỏ gt H0.

Vậy, với mức ý nghĩa 5%, ta có thể cho rằng tỉ lệ chi tiêu cho ăn uống trên 1000 ở tổng thể hộ không nghèo là cao hơn so với tỉ lệ đó ở hộ nghèo.

Bài toán 6: Kiểm định giả thiết và tìm Khoảng tin cậy so sánh phương sai 2 tổng thể có phân phối chuẩn

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

Câu hỏi

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
Chơi game	126	115	133	136	111	89	101	126	110	122
Chơi cờ	117	138	111	148	106	119	125	120	134	109

x: 126 115 133 136 111 89 101 126 110 122

y: 117 138 111 148 106 119 125 120 134 109

#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
11:
Read 10 items
```

```
> y=scan()
1: 117 138 111 148 106 119 125 120 134 109
11:
Read 10 items
> var.test(x, y, ratio = 1,alternative = "t",conf.level = 0.95)

        F test to compare two variances

data:  x and y
F = 1.1615, num df = 9, denom df = 9, p-value = 0.8271
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2885073 4.6763080
sample estimates:
ratio of variances
          1.161529
```

Do p-value = 0.8271 > 0.05 nên chấp nhận gt H0, có thể xem 2 PS của các tổng thể như nhau.

Example

Từ dữ liệu trong file "ChiTieu2010.csv", kiểm định khẳng định cho rằng phương sai trong chi tiêu cho ăn uống hàng tháng của hai tổng thể hộ nghèo và tổng thể hộ không nghèo là như nhau. Giả sử rằng hai tổng thể trên đều có phân bố chuẩn.

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

H0 : V1/V2=1 ; H1 : V1/V2 khác 1

```
> DL=read.csv("ChiTieu2010.csv")
> attach(DL)
The following objects are masked from DL (pos = 3):

    ChiTieuGiaoDucTrongNam, ChiTieuKhac, ChiTieuTaiSanKhongHaoMonTrong10Nam,
    ChiTieuYTe, CTAnUongDipLeTrongNam, CTAnUongTrongThang,
    CTSinhHoatNgoaiAnUongTrongThang, CuaCaiGiaTriTrongNam, CuaCaiTrongNam,
    DieuTriNgoaiTru, DieuTriNoiTru, GioiTinh, HoNgheo, Huyen, ï..Tinh, KhuVuc,
    SoNguoiTrongHo, ThietBiYTe, ThueNhaDienNuocTrongNam, Thuoc, TongChiTieu,
    Tuoi, Xa

> x=CTAnUongTrongThang[HoNgheo==1]
> y=CTAnUongTrongThang[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 = 0.071749, 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:
 0.06703860 0.07689138
sample estimates:
ratio of variances
         0.07174904
```

Ta có `p-value < 2.2x10^(-16)`< 0.05 nên bác bỏ gt 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.

Câu hỏi

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.*

1) Tỉ lệ hộ nghèo ở nông thôn 2 là cao hơn thành thị 1 hay Tỉ lệ hộ nghèo ở thành thị 1 là thấp hơn nông thôn 2?

```
> 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 n1=2647, n2=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-value < 2.2e-16<0.05` nên bác bỏ gt H0.

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

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?

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 gt H0.

Vậy phương sai khác nhau.

Bài toán 7: Phân tích phương sai

Sử dụng R trong phân tích phương sai

Mẫu gộp = c(mẫu 1, mẫu 2, . . . , mẫu k)

Phân loại = factor(rep(c(1:k,c($n_1, n_2 \ldots, n_k$)))

Và tính toán kiểm định:

anova(lm(Mẫu gộp ∼ Phân loại))

Kết quả sẽ cho ta bảng phân tích phương sai, kèm theo P - giá trị của bài toán.

Khi bác bỏ H_0, thực hiện phân tích sâu Tukey nhờ hàm:

TukeyHSD(aov(Mẫu gộp ∼ Phân loại))

Kết quả sẽ cho ta bảng các P - giá trị của từng cặp dấu hiệu.

Example

Để xem liệu điều kiện kinh tế khác nhau có ảnh hưởng đến số con trong một gia đình hay không, người ta thu phân loại ra 3 mức về điều kiện kinh tế: Trên mức khá giả, khá giả, dưới mức khá giả. Sau đó chọn ngẫu nhiên ở mỗi loại 4 gia đình và ghi lại số con của các gia đình như sau:

Trên mức khá giả	Khá giả	Dưới mức khá giả
2	1	3
3	2	4
3	1	2
2	2	3

- Liệu các con số đó có cho ta thấy số con trung bình của các hộ thuộc diện kinh tế khác nhau là như nhau không?
- Điều kiện kinh tế ảnh hưởng như thế nào tới số con trong gia đình?

(Giả thiết phân phối là chuẩn với phương sai đồng nhất)

#Gọi mu1,mu2, mu3 lần lượt là số con trung bình của các hộ thuộc 3 nhóm

H0: mu1=mu2= mu3; H1:Tồn tại i,j thuộc {1,2,3}: mui khác muj

```
> MauGop=c(2,3,3,2,1,2,1,2,3,4,2,3)
> PhanLoai=factor(rep(1:3,each=4))
> anova(lm(MauGop~PhanLoai))
Analysis of Variance Table

Response: MauGop
          Df Sum Sq Mean Sq F value  Pr(>F)
PhanLoai   2 4.6667 2.33333    5.25 0.03083 *
Residuals  9 4.0000 0.44444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ta được p-value=0.03083<0.05 nên bác bỏ gt H0

Có sự khác nhau giữa các trung bình tổng thể, điều kiện kinh tế ảnh hưởng đến số con trong các gia đình.

Muốn biết các cặp trung bình nào khác nhau?

```
> TukeyHSD(aov(MauGop~PhanLoai),conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = MauGop ~ PhanLoai)

$PhanLoai
```

```
    diff        lwr       upr     p adj
2-1 -1.0 -2.3161641 0.3161641 0.1402115
3-1  0.5 -0.8161641 1.8161641 0.5600698
3-2  1.5  0.1838359 2.8161641 0.0272373
```

Với mức ý nghĩa 0.05, chỉ có cặp 3-2 cho p-value=0.0272373<0.05 nên mu3 khác mu2 (TB dưới mức khác giả khác mức khá giả), cụ thể hơn diff(3-2)=1.5>0 nên mu3>mu2; ta chấp nhận mu2=mu1, mu3=mu1.

Chú ý: Để hiểu chi tiết cụ thể ta làm như sau

```
> MauGop=c(2,3,3,2,1,2,1,2,3,4,2,3)
>PhanLoai=factor(c(rep("TrenMucKhaGia",4),rep("KhaGia",4),rep("DuoiMucKhaGia",4)))
> #Kiểm tra lại đề bài
> DL=data.frame(MauGop,PhanLoai)
> DL
   MauGop      PhanLoai
1       2 TrenMucKhaGia
2       3 TrenMucKhaGia
3       3 TrenMucKhaGia
4       2 TrenMucKhaGia
5       1        KhaGia
6       2        KhaGia
7       1        KhaGia
8       2        KhaGia
9       3 DuoiMucKhaGia
10      4 DuoiMucKhaGia
11      2 DuoiMucKhaGia
12      3 DuoiMucKhaGia
> anova(lm(MauGop~PhanLoai))
Analysis of Variance Table

Response: MauGop
          Df Sum Sq Mean Sq F value  Pr(>F)
PhanLoai   2 4.6667 2.33333    5.25 0.03083 *
Residuals  9 4.0000 0.44444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(aov(MauGop~PhanLoai),conf.level = 0.95)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = MauGop ~ PhanLoai)

$PhanLoai
                             diff        lwr        upr     p adj
KhaGia-DuoiMucKhaGia        -1.5 -2.8161641 -0.1838359 0.0272373
TrenMucKhaGia-DuoiMucKhaGia -0.5 -1.8161641  0.8161641 0.5600698
TrenMucKhaGia-KhaGia         1.0 -0.3161641  2.3161641 0.1402115
```

Cặp KhaGia-DuoiMucKhaGia có p-value=0.0272373<0.05 nên có trung bình khác nhau, cụ thể hơn diff=-1.5<0 nên TB số con ở mức `KhaGia<` TB số con ở các hộ `DuoiMucKhaGia`.

Ví dụ 2: So sánh 3 loại thuốc bổ A, B, C trên 3 nhóm, người ta được kết quả tăng trọng(kg) như sau:

A: 1.0 1.1 1.2 1.4 0.7 0.9

B: 1.0 1.7 1.8 2.1 1.4 1.2 1.6 1.9

C: 0.5 1.3 0.7 0.5 0.3 0.6 0.5

a) Hãy so sánh kết quả tăng trọng của 3 loại thuốc bổ trên với mức ý nghĩa 0.05.

b) Nếu kết quả tăng trọng của 3 loại thuốc bổ trên khác nhau có ý nghĩa, hãy so sánh từng cặp với mức ý nghĩa 0.04.

Giả thiết phân phối là chuẩn với phương sai đồng nhất.

#Gọi muA,muB, muC lần lượt là số con trung bình của các hộ thuộc 3 nhóm A, B, C

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

```
> MauGop=scan()
1: 1.0  1.1  1.2  1.4  0.7  0.9
7: 1.0  1.7  1.8  2.1  1.4  1.2  1.6  1.9
15: 0.5  1.3  0.7  0.5  0.3  0.6  0.5
22:
Read 21 items
> PhanLoai=factor(c(rep("A",6),rep("B",8),rep("C",7)))
> #Kiểm tra lạidữ liệu
> DL=data.frame(MauGop,PhanLoai)
> DL
   MauGop PhanLoai
1     1.0        A
2     1.1        A
3     1.2        A
4     1.4        A
5     0.7        A
6     0.9        A
7     1.0        B
8     1.7        B
9     1.8        B
10    2.1        B
11    1.4        B
12    1.2        B
13    1.6        B
14    1.9        B
15    0.5        C
16    1.3        C
17    0.7        C
18    0.5        C
19    0.3        C
20    0.6        C
21    0.5        C
> anova(lm(MauGop~PhanLoai))
Analysis of Variance Table

Response: MauGop
          Df Sum Sq Mean Sq F value    Pr(>F)
PhanLoai   2 3.4677 1.73384  16.797 7.657e-05 ***
Residuals 18 1.8580 0.10322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Do p-value= 7.657e-05= 7.657x10^(-5) <0.04 nên bác bỏ gt H0, có sự khác nhau giữa các trung bình tổng thể.

b)

```
> TukeyHSD(aov(MauGop~PhanLoai),conf.level = 0.96)
```

```
  Tukey multiple comparisons of means
    96% family-wise confidence level

Fit: aov(formula = MauGop ~ PhanLoai)

$PhanLoai
           diff         lwr         upr     p adj
B-A  0.5375000  0.07531263  0.99968737 0.0162741
C-A -0.4214286 -0.89755436  0.05469722 0.0730510
C-B -0.9589286 -1.40184959 -0.51600755 0.0000519
```

Chỉ có nhóm C-A có p-value=0.0730510>0.04 nên muC=muA, các cặp còn lại có trung bình khác nhau: muB>muA, muC<muB. Hay muB lớn nhất, muC=muA, với mức ý nghĩa 0.04.

Ví dụ 3: Từ tập dữ liệu ChiTieu2010.csv, với mức ý nghĩa 0.05.

a)Hãy kiểm định trung bình chi tiêu các hạng mục `ChiTieuGiaoDucTrongNam, ChiTieuYTe, CTAnUongDipLeTrongNam` có như nhau không?

b) Yếu tố khu vực có ảnh hưởng đến chi tiêu giáo dục trong năm không?

HD:

a) Gọi muA,muB, muC lần lượt là trung bình tổng thể chi tiêu các hạng mục `ChiTieuGiaoDucTrongNam, ChiTieuYTe, CTAnUongDipLeTrongNam`

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=CTAnUongDipLeTrongNam
> 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.

b) Gọi muA,muB lần lượt là trung bình tổng thể chi tiêu giáo dục tại các khu vực 1 và 2.

H0: muA=muB (Yếu tố khu vực không ảnh hưởng đến chi tiêu giáo dục trong năm)

H1: muA khác muB (Yếu tố khu vực có ảnh hưởng đến chi tiêu giáo dục trong năm)

Cách 1:

```
> MauGop=ChiTieuGiaoDucTrongNam
```

```
> PhanLoai=as.factor(KhuVuc)
> anova(lm(MauGop ~ PhanLoai))
Analysis of Variance Table

Response: MauGop
            Df     Sum Sq  Mean Sq F value    Pr(>F)
PhanLoai     1   97010548 97010548   190.5 < 2.2e-16 ***
Residuals 9396 4784959375   509255
---
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 2 trung bình tổng thể.

Yếu tố khu vực có ảnh hưởng đến chi tiêu giáo dục trong năm.

Cách 2:

```
> x=ChiTieuGiaoDucTrongNam[KhuVuc==1]
> y=ChiTieuGiaoDucTrongNam[KhuVuc==2]
> MauGop=c(x,y)
> PhanLoai=factor(c(rep("1",length(x)),rep("2",length(y))))
> anova(lm(MauGop ~ PhanLoai))
Analysis of Variance Table

Response: MauGop
            Df     Sum Sq  Mean Sq F value    Pr(>F)
PhanLoai     1   97010548 97010548   190.5 < 2.2e-16 ***
Residuals 9396 4784959375   509255
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Cách 3: Do chỉ có 2 khu vực nên ta đưa về bài toán kiểm định 2 giá trị trung bình, dùng hàm t.test. Tuy nhiên cách này không sử dụng được nếu có nhiều hơn 2 khu vực, bài toán trở thành so sánh nhiều giá trị trung bình (phân tích phương sai anova).

Ví dụ(Về phân tích phương sai 2 nhân tố-tham khảo)

Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa đến năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau:

Giống lúa Loại phân bón	1	2	3
A1	60	55	65
B2	50	50	48
C3	85	46	74
D4	65	58	66

Với mức ý nghĩa 0.05, hãy đánh giá sự ảnh hưởng giống lúa, loại phân bón trên năng suất lúa.

Giả thiết phân phối là chuẩn với phương sai đồng nhất.

#H0: Trung bình năng suất lúa như nhau với 3 giống lúa

Trung bình năng suất lúa như nhau ứng với 4 loại phân bón

#H1: Có sự ảnh hưởng giống lúa, loại phân bón trên năng suất lúa

```
> MauGop=scan()
1: 60   55      65
4: 50   50      48
7: 85   46      74
10: 65  58      66
13:
Read 12 items
> MauGop
 [1] 60 55 65 50 50 48 85 46 74 65 58 66
> LoaiPhanBon=gl(4,3,12)
> GiongLua=gl(3,1,12)
> DuLieu=data.frame(LoaiPhanBon,GiongLua,MauGop)
> DuLieu
   LoaiPhanBon GiongLua MauGop
1            1        1     60
2            1        2     55
3            1        3     65
4            2        1     50
5            2        2     50
6            2        3     48
7            3        1     85
8            3        2     46
9            3        3     74
10           4        1     65
11           4        2     58
12           4        3     66
> anova(lm(MauGop~LoaiPhanBon+GiongLua))
Analysis of Variance Table

Response: MauGop
            Df Sum Sq Mean Sq F value Pr(>F)
LoaiPhanBon  3 576.33 192.111  2.2288 0.1854
GiongLua     2 382.17 191.083  2.2169 0.1902
Residuals    6 517.17  86.194
```

Loại phân bón: p-value=0.1854>0.05 nên 4 loại phân bón này như nhau không ảnh hưởng đến năng suất.

Giống lúa: p-value=0.1902>0.05 nên 3 loại giống lúa này như nhau không ảnh hưởng đến năng suất.

Bài tập Luyện tập

Khoảng tin cậy:

1. Một mẫu ngẫu nhiên kích thước n_1=25 được lấy từ một tổng thể có phân phối chuẩn với độ lệch chuẩn 5 và cho giá trị trung bình $\bar{x}_1 = 80$. Một mẫu ngẫu nhiên thứ hai kích thước n_2 = 6 được lấy từ một tổng thể khác cũng có phân phối chuẩn với độ lệch chuẩn 3, và cho giá trị trung bình $\bar{x}_2 = 75$. Xác định khoảng tin cậy 94% cho $\mu_1 - \mu_2$.

2. Trong một phản ứng hóa học, hai chất xúc tác được so sánh về tác động lên hiệu suất của quá trình phản ứng. Một mẫu 12 phản ứng được sử dụng chất xúc tác 1 và một mẫu 10 phản ứng được sử dụng chất xúc tác 2. 12 mẫu sử dụng chất xúc tác 1 cho khối lượng bình quân 85 với độ lệch

chuẩn mẫu 4 và khối lượng bình quân cho mẫu thứ hai là 81 với độ lệch chuẩn 5. Xác định khoảng tin cậy 90% cho hiệu số giữa các trung bình tổng thể, giả thiết các tổng thể có phân phối xấp xỉ chuẩn với các giá trị phương sai bằng nhau.

3. Dữ liệu sau đây, được ghi nhận theo ngày, thể hiện khoảng thời gian hồi phục đối với các bệnh nhân được điều trị ngẫu nhiên bằng một trong hai loại thuốc để điều trị nhiễm trùng bang quang nặng:

Loại thuốc 1	Loại thuốc 2
$n_1 = 14$	$n_2 = 16$
$\overline{x}_1 = 17$	$\overline{x}_2 = 19$
$s_1^2 = 1,5$	$s_2^2 = 1,8$

Xác định khoảng tin cậy 99% cho hiệu thời gian hồi phục trung bình cho hai loại thuốc, giả thiết các tổng thể có phân phối chuẩn với phương sai bằng nhau.

4. Trong một nghiên cứu được tiến hành tại Học viện bách khoa Virginia và Đại học tổng hợp bang về phát triển của ectomycorrhizal, một mối quan hệ cộng sinh giữa các rễ cây và nấm trong đó các khoáng chất được chuyển từ nấm sang cây và đường từ cây sang nấm, 20 cây giống sồi đỏ miền Bắc bị nấm *Pisolithus tinctorus* được trồng trong nhà kính. Tất cả các cây giống đều được trồng trong cùng một loại đất và có mức chiếu sáng và nước như nhau. Một nửa không nhận được nitơ trong thời gian trồng để làm cây đối chứng và số còn lại nhận được 368 phần triệu nitơ dưới dạng $NaNO_3$. Các trọng lượng gốc được xác định bằng gam trong ngày cuối cùng của 140 ngày như sau:

Không có Nitơ	Có Nitơ
0,32	0,26
0,53	0,43
0,28	0,47
0,37	0,49
0,47	0,52
0,43	0,75
0,36	0,79
0,42	0,86
0,38	0,62
0,43	0,46

Xác định khoảng tin cậy 95% cho hiệu số trong các trọng lượng gốc trung bình giữa các cây giống không nhận nitơ và các cây có nhận được 368 phần triệu nitơ. Giả thiết rằng các tổng thể đó có phân bố chuẩn với các phương sai bằng nhau.

5. Một nhà nghiên cứu gien quan tâm đến tỷ lệ nam giới và nữ giới trong tổng thể bị rối loạn tiểu cầu. Trong một ngẫu nhiên gồm 1000 nam giới, 250 được xác định bị rối loạn tiểu cầu, trái lại có 275 người bị rối loạn tiểu cầu trong 1000 nữ giới được kiểm tra. Tính khoảng tin cậy 95% cho sự khác nhau giữa tỷ lệ giữa nam giới và nữ giới bị bệnh này.

6. Một thử nghiệm lâm sàng được tiến hành để xác định liệu một loại thuốc tiêm chủng có ảnh hưởng lên tỷ lệ lây lan của một bệnh hay không. Một mẫu gồm 1000 con chuột được nuôi trong môi trường đối chứng trong thời gian 1 năm và 500 con chuột được tiêm chủng. Trong nhóm không được tiêm thuốc có 120 bị mắc bệnh, trong khi đó 98 trong số được tiêm chủng nhiễm bệnh . Nếu chúng ta gọi p_1 là xác suất bị nhiễm bệnh trong số chuột không được tiêm chủng và p_2 là xác suất nhiễm bệnh sau khi tiêm thuốc, tính khoảng tin cậy 90% cho $p_1 - p_2$.

7. Một nghiên cứu khảo sát trên 1000 sinh viên kết luận rằng có 274 sinh viên chọn đội bóng chày chuyên nghiệp A là đội yêu thích của mình. Trong năm 1991, nghiên cứu tương tự cũng được tiến hành trên 760 sinh viên. Kết luận rằng 240 sinh viên trong số đó cũng chọn đội bóng chày chuyên nghiệp A là đội bóng yêu thích của họ. Tìm khoảng tin cậy 95% cho sự khác nhau của hai tỷ lệ trên trong hai năm đó. Sự khác nhau đó có đáng kể không?

Kiểm định giả thiết

1. Một mẫu ngẫu nhiên cỡ $n_1 = 25$, lấy từ phân phối chuẩn với $\sigma_1 = 5,2$ có trung bình mẫu $\bar{x}_1 = 81$. Một mẫu khác cỡ $n_2 = 36$, lấy từ phân phối chuẩn với $\sigma_2 = 3,4$ có trung bình mẫu $\bar{x}_2 = 76$. Kiểm định giả thuyết $\mu_1 = \mu_2$ với đối thuyết $\mu_1 \neq \mu_2$. Mức ý nghĩa 0,05.

2. Một hãng sản xuất xe hơi muốn xác định xem, nên dùng loại lốp A hay B cho loại xe mới của họ. Họ thực hiện thí nghiệm với 12 chiếc lốp mỗi loại, và ghi lại số km đi được đến khi phải thay lốp. Kết quả như sau:

Loại A: $\bar{x}_1 = 37900$ km ; $s_1 = 5100$ km.

Loại B: $\bar{x}_2 = 39800$ km ; $s_1 = 5900$ km.

Hãy kiểm định giả thuyết rằng không có sự khác biệt giữa hai loại lốp, với mức ý nghĩa 0,05. Giả sử các phân phối đều chuẩn, với phương sai bằng nhau.

3. Một mẫu gồm 32 phụ nữ đang có thai vào giai đoạn 3 tháng cuối của thai kỳ, có độ tuổi từ 15 đến 32, được chia làm hai nhóm hút thuốc và không hút thuốc. Người ta đo nồng độ axit huyết tương ascorbic (mg/ml) trong máu của họ, khi họ chưa ăn sáng hay các đồ ăn chứa axit này, được số liệu sau:

Hút thuốc:	0,48	0,71	0,98	0,68	1,18	1,36	0,78	1,64				
Không hút:												
	0,97	0,72	1,00	0,81	0,62	1,32	1,24	0,99	0,90	0,74	1,24	1,18
	0,88	0,94	1,16	0,86	0,85	0,58	0,57	0,64	0,98	1,09	0,92	0,78

Giả sử các số liệu tuân theo phân phối chuẩn với phương sai bằng nhau. Kiểm định xem có sự sai khác đáng kể nào giữa nồng độ ascorbic trung bình của hai nhóm hút thuốc và không hút thuốc không? Mức ý nghĩa 0,005.

4. Năm mẫu quặng sắt, mỗi mẫu được chia thành hai phần, rồi lần lượt được xác định hàm lượng sắt bằng hai cách là dùng tia X và dùng phân tích hóa học, kết quả thu được là

	Số thứ tự mẫu				
Cách phân tích	**1**	**2**	**3**	**4**	**5**
Tia X	2,0	2,0	2,3	2,1	2,4
Phân tích hóa học	2,2	1,9	2,5	2,3	2,4

Giả sử các số liệu ở mỗi cách phân tích tuân theo phân phối chuẩn. Hãy kiểm định rằng hai phương pháp cho kết quả giống nhau, với mức ý nghĩa 0,05?

---

5. Trong một mẫu ngẫu nhiên gồm 200 phụ nữ trưởng thành sống ở thành thị, có 20 người mắc ung thư vú. Con số này là 10 trên 150 phụ nữ sống ở nông thôn được chọn ngẫu nhiên. Liệu có thể kết luận, với mức ý nghĩa 0,06 rằng, bệnh ung thư vú là thường gặp hơn ở thành thị không?

6. Một nhà di truyền học quan tâm tới tỷ lệ nam và nữ trong dân số bị mắc chứng rối loạn máu. Trong mẫu ngẫu nhiên gồm 100 nam giới, có 31 người mắc chứng này; và trong 100 nữ giới có 24 người mắc. Có thể kết luận với mức ý nghĩa 0,01 rằng, tỷ lệ nam giới mắc chứng rối loạn máu lớn hơn so với tỷ lệ nữ giới mắc chứng này không?

7. Một nghiên cứu được thực hiện để xem có phải nhiều người Ý hơn người Mỹ thích sâm-panh trắng hơn sâm-panh đỏ trong ngày cưới không. Chọn ngẫu nhiên 300 người Ý, thấy có 72 người

thích sâm-panh trắng; và chọn 400 người Mỹ, thì 70 người thích sâm-panh trắng hơn sâm-panh đỏ. Vậy có thể kết luận tỷ lệ người Ý thích sâm-panh trắng trong ngày cưới là cao hơn so với người Mỹ không? Dùng mức ý nghĩa 0,05.

8. Một nghiên cứu của Khoa Giáo dục thể chất, trường Đại học Virginia nhằm xác định xem sau 8 tuần luyện tập, lượng cholesterol của những người tham gia luyện tập có thực sự giảm không. Một nhóm 15 người tham gia luyện tập 2 lần một tuần. Một nhóm khác gồm 18 người với độ tuổi tương tự, không tham gia luyện tập. Sau 8 tuần, lượng cholesterol được ghi lại như sau:

Nhóm luyện tập: 129 131 154 172 115 126 175 191 122 238 159
156 176 175 126

Nhóm không luyện tập: 151 132 196 195 188 198 187 168 115
165 137 208 133 217 191 193 140 146

Ta có thể kết luận, với mức ý nghĩa 5% rằng, lượng cholesterol thực sự sẽ giảm sau khi thực hiện chương trình luyện tập không?

9. Một nghiên cứu được thực hiện bởi Trung tâm Thủy lợi và được phân tích bởi Trung tâm Thống kê, thuộc Đại học Virginia, nhằm so sánh hai thiết bị xử lý nước thải. Thiết bị A được đặt ở vùng dân cư có thu nhập trung bình dưới 22000$/năm. Thiết bị B được đặt ở vùng dân cư có thu nhập trung bình trên 60000$/năm. Lượng nước thải được xử lý bởi mỗi thiết bị (tính theo nghìn ga-lông/ ngày) được đo trong 10 ngày như sau:

Thiết bị A: 21 19 20 23 22 28 32 19 13 18

Thiết bị B: 20 39 24 33 30 28 30 22 33 24

Với mức ý nghĩa 5%, có thể kết luận rằng lượng nước thải trung bình được xử lý ở vùng có thu nhập thấp là nhỏ hơn vùng có thu nhập cao không.

10. Theo dõi hoạt động sản xuất của một xí nghiệp cho thấy độ lệch chuẩn của năng suất của thiết bị A trong 30 ngày làm việc là 45 sản phẩm/ngày, trong khi đó số liệu này của thiết bị B trong 25 ngày làm việc là 36 sản phẩm/ngày. Với mức ý nghĩa 5%, có thể kết luận rằng phương sai 2 tổng thể như nhau không, giả thiết phân phối là chuẩn.

Phân tích phương sai

Example

Một người bán hàng sử dụng một trang facebook để đăng quảng cáo. Anh ta muốn biết xem khung giờ nào là tốt nhất để đăng: sáng, trưa, chiều hay tối. Thử nghiệm của anh ta trong vài lần (cùng vào ngày thứ 7 với các loại hàng hóa đẹp tương đương nhau) cho thấy số lượt thích (đơn vị trăm) như sau:

Sáng	Trưa	Chiều	Tối
7	13	11	13
9	7	12	12
4	8	12	13
6	6	9	9

Giả sử số lượt thích của mỗi khoảng thời điểm trên là có phân bố chuẩn, có phương sai như nhau. Tại mức ý nghĩa 5%, đăng quảng cáo vào khoảng thời gian khác nhau trong ngày có dẫn tới lượt thích khác nhau không?

Example

Giả sử chúng ta muốn đánh giá tuổi tác có tác động đến việc đầu tư hay không, một mẫu ngẫu nhiên được chọn từ những tổng thể các nhà đầu theo độ tuổi: trẻ (<35 tuổi), trung niên (35 - 49), xế trung niên (50 - 65), già (trên 65) và nghi lại tỉ lệ tài sản của người đó dùng để đầu tư như sau:

Trẻ	Trung niên	Xế trung niên	Già
24.8	28.9	81.5	66.8
35.5	7.3	0.0	77.4
68.7	61.8	61.3	32.9
42.2	53.6	0.0	74.0

Giả sử tỉ lệ tài sản dành cho đầu tư của tổng thể có phân bố chuẩn.
Hãy kiểm tra điều kiện về phương sai (theo quy tắc nhanh đã nêu) cho bài toán phân tích phương sai.
Nếu điều kiện về phương sai thỏa mãn, tại mức ý nghĩa 5%, tuổi tác có tác động đến tỉ lệ tài sản mà người đó dành cho đầu tư không? Nếu có, độ tuổi nào người ta đầu tư với tỉ lệ tài sản lớn nhất?

Example (Ví dụ giả tưởng)

Một thành viên cái bang muốn xem trang phục của anh ta có ảnh hưởng đến lượng tiền xin được trong ngày không. Anh ta mặc ngẫu nhiên những áo với màu chủ đạo lần lượt là: xanh, đỏ, tím, vàng vào các ngày đi hành nghề ở cùng một khu vực. Kết quả được bảng sau đây (đơn vị nghìn đồng):

Xanh	Đỏ	Tím	Vàng
200	250	150	180
150	300	160	150
110	100	140	220
100	190	110	160
180	150	130	150

Giả sử tổng thể tiền xin được trong ngày của mỗi màu áo chủ đạo là tuân theo phân phối chuẩn, có cùng phương sai. Tại mức ý nghĩa 5%, hãy cho biết trang phục của người ăn xin có ảnh hưởng đến lượng tiền xin được không? Nếu có, trang phục nào giúp xin được nhiều nhất?