Bài 9.

Kiểm định giả thuyết thống kê
(Trường hợp hai mẫu)

I. So sánh kỳ vọng giữa hai tổng thể độc lập:

Giả sử $(X_1, X_2, \ldots, X_n)$ (mẫu 1) là mẫu ngẫu nhiên chọn từ một tổng thể có phân phối chuẩn với kỳ vọng là μ_X và phương sai là σ_X^2 và $(Y_1, Y_2, \ldots, Y_m)$ (mẫu 2) là mẫu ngẫu nhiên độc lập chọn từ một tổng thể có phân phối chuẩn với kỳ vọng là μ_Y và phương sai σ_Y^2. Trong phần thực hành, ta giả sử σ_X^2 và σ_Y^2 không biết.

Các giả thuyết:

- Một phía – bên trái: $\begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases}$
- Một phía – bên phải: $\begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases}$
- Hai phía: $\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$

Trường hợp $\sigma_X^2 = \sigma_Y^2 = \sigma^2$**:** giả sử phương sai của hai tổng thể bằng nhau.

Sử dụng phương sai mẫu chung S_P^2,

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \tag{1}$$

Thống kê kiểm định,

$$T_0 = \frac{\bar{X} - \bar{Y}}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{2}$$

T_0 có phân phối Student với $n + m - 2$ bậc tự do.

Với $\bar{X}, S_X$ và Y, S_Y lần lượt là trung bình mẫu và độ lệch tiêu chuẩn mẫu tương tứng với mẫu 1 và mẫu 2.

Kết luận:

Giả thuyết H_0	Đối thuyết H_1	Miền bác bỏ với mức ý nghĩa α	P – giá trị
$\mu_X = \mu_Y$	$\mu_X \neq \mu_Y$	$\lvert T_0 \rvert > t_{1-\alpha/2}^{n+m-2}$	$2P\{T_{n+m-2} \geq \lvert T_0 \rvert\}$
$\mu_X \leq \mu_Y$	$\mu_X > \mu_Y$	$T_0 > t_{1-\alpha}^{n+m-2}$	$P\{T_{n+m-2} \geq T_0\}$
$\mu_X \geq \mu_Y$	$\mu_X < \mu_Y$	$T_0 < -t_{1-\alpha}^{n+m-2}$	$P\{T_{n+m-2} \leq T_0\}$

Bảng 1

Trường hợp $\sigma_X^2 \neq \sigma_Y^2$: trong một số trường hợp, ta không thể giả sử các phương sai không biết σ_X^2 và σ_Y^2 bằng nhau. Khi đó thống kê kiểm định:

$$T_0^* = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \tag{3}$$

T_0 có phân phối Student với bậc tự do *df* được xác định bởi

$$df = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{\left(S_X^2/n\right)^2}{n-1} + \frac{\left(S_Y^2/m\right)^2}{m-1}} \tag{4}$$

Thay thế T_0 trong bảng 1 bởi T_0^* ta thu được các kết luận tương ứng.

Chú ý, nếu biết σ_X^2 và σ_Y^2 thì công thức (3) trở thành

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \tag{5}$$

Thống kê Z ~ N(0,1), khi đó thay thế các giá trị thống kê và phân vị trong bảng 1 bằng thống kê Z và phân vị $z_{1-\alpha/2}$ hoặc $z_{1-\alpha}$ tương ứng.

Thực hành trong R:

Sử dụng hàm `t.test`

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

Trong đó

- x, y : véc-tơ dữ liệu tương ứng với hai mẫu
- alternative: đối thuyết (hai phía – "two.sided", bên trái – "less", bên phải – "greater")

- var.equal: giá trị mặc định là FALSE nghĩa là phương sai khác nhau
- conf.level: độ tin cậy của khoảng tin cậy cho $\mu_X - \mu_Y$, mặc định là 95%

Nếu như hai mẫu $(x_1,\ldots,x_n)$ và $(y_1,\ldots,y_m)$ được chứa trong cùng 1 biến, chẳng hạn `data` và được phân biệt bởi 1 biến chia nhóm, `group`. Ta sử dụng hàm `t.test` với cú pháp sau

```
t.test(data ~ group, alternative = c("two.sided", "less", "greater"),
                var.equal = FALSE, conf.level = 0.95)
```

Ví dụ 1:

File `scores.rda` chứa điểm thi giữa kỳ và cuối kỳ của môn Thống kê (Thang điểm: 100)

```
> load('scores.rda')
> scores
$midterm
 [1] 66 78 62 99 80 63 82 86 84 70 98 81 66 42 92 74 75 89 87 84 89 87 76 45
84
$final
 [1] 79 78 65 75 84 94 79 84 79 66 76 76 79 91 88 78 77 87 86 73 73 84 88 79
```

Hỏi: điểm trung bình giữa kỳ và cuối kỳ có bằng nhau hay không với α = 0.05?

```
> alpha <-0.05
> (var.equal <- var.test(scores$midterm, scores$final))
        F test to compare two variances
data:  scores$midterm and scores$final
F = 3.9807, num df = 24, denom df = 23, p-value = 0.001486
```

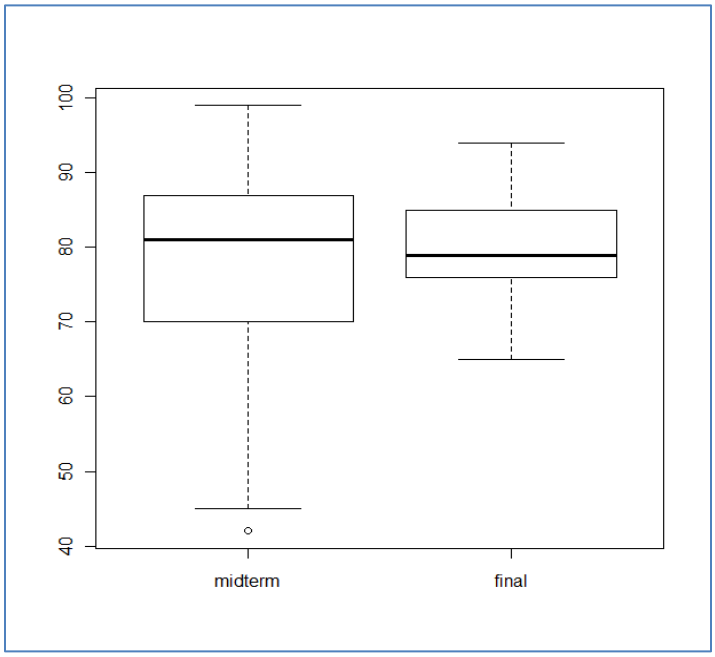
Sử dụng hàm var.test để kiểm tra xem phương sai hai tổng thể có bằng nhau hay không? p giá trị = 0.001486 suy ra phương sai khác nhau.

```
> t.test(scores$midterm, scores$final, var.equal=FALSE)
        Welch Two Sample t-test
data:  scores$midterm and scores$final
t = -0.7354, df = 35.656, p-value = 0.4669
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.858370  4.145037
sample estimates:
mean of x mean of y
 77.56000  79.91667
```

p giá trị = 0.4669 > 0.05, ta kết luận rằng điểm thi giữa kỳ không khác với điểm thi cuối kỳ với mức ý nghĩa $\alpha = 0.05$.

Biểu đồ Boxplot cho midterm và final:

```
> boxplot(scores)
```



II. So sánh theo cặp:

Giả sử $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_n)$ là hai mẫu có cùng cỡ được chọn từ cùng một tổng thể có phân phối chuẩn với kỳ vọng lần lượt là μ_X và μ_Y; như vậy, sẽ có một mối liên hệ giữa các giá trị X_i và Y_i. Do đó, các cặp giá trị (X_i, Y_i), $i = 1, \ldots, n$ sẽ không độc lập. Do đó, ta sử dụng phương pháp kiểm định theo cặp để kiểm tra các giả thuyết liên quan đến μ_X và μ_Y. Thông thường, đối với bài toán so sánh theo cặp, ta sẽ ghi nhận hai giá trị trên mỗi phần tử của mẫu tại hai thời điểm khác nhau, chẳng hạn như trước và sau khi áp dụng một phương pháp nghiên cứu trên các phần tử đó. Ví dụ: đo lượng cholesterol ở những người béo phì trước và sau khi cho họ áp dụng các bài tập thể dục và chế độ ăn kiêng; đo nhịp tim ở của những người tham gia thí nghiệm trước và sau khi cho họ uống cà-fê để so sánh tác dụng của chất caffein trong cà-fê, …

Với $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ là n cặp quan trắc, đặt

$$D_i = X_i - Y_i, \quad i = 1, \ldots, n \tag{5}$$

Biến ngẫu nhiên D_i giả sử có phân phối chuẩn với kỳ vọng là

$$\mu_D = E(D_i) = E(X_i - Y_i) = E(X_i) - E(Y_i) = \mu_X - \mu_Y$$

và phương sai σ_D^2.

Các giả thuyết:

$$\begin{cases} H_0 : \mu_D \geq 0 \\ H_1 : \mu_D < 0 \end{cases} \qquad \begin{cases} H_0 : \mu_D \leq 0 \\ H_1 : \mu_D > 0 \end{cases} \qquad \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

Thống kê kiểm định:

$$T_0 = \frac{\bar{D}}{S_D / \sqrt{n}}$$

có phân phối Student với n – 1 bậc tự do.

Trong đó: $\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$ và $S_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$

Kết luận:

Giả thuyết H_0	Đối thuyết H_1	Miền bác bỏ với mức ý nghĩa α	P – giá trị
$\mu_D = 0$	$\mu_D \neq 0$	$\lvert T_0 \rvert > t_{1-\alpha/2}^{n-1}$	$2P\{T_{n-1} \geq \lvert T_0 \rvert\}$
$\mu_D \leq 0$	$\mu_D > 0$	$T_0 > t_{1-\alpha}^{n-1}$	$P\{T_{n-1} \geq T_0\}$
$\mu_D \geq 0$	$\mu_D < 0$	$T_0 < -t_{1-\alpha}^{n-1}$	$P\{T_{n-1} \leq T_0\}$

Bảng 2

Thực hành trong R:

Sử dụng hàm `t.test`

```
t.test(x, y, alternative = c("two.sided", "less", "greater"),
         paired = TRUE, conf.level=0.95)
```

hoặc

```
t.test(x ~ group, alternative = c("two.sided", "less", "greater"),
         paired = TRUE, conf.level=0.95)
```

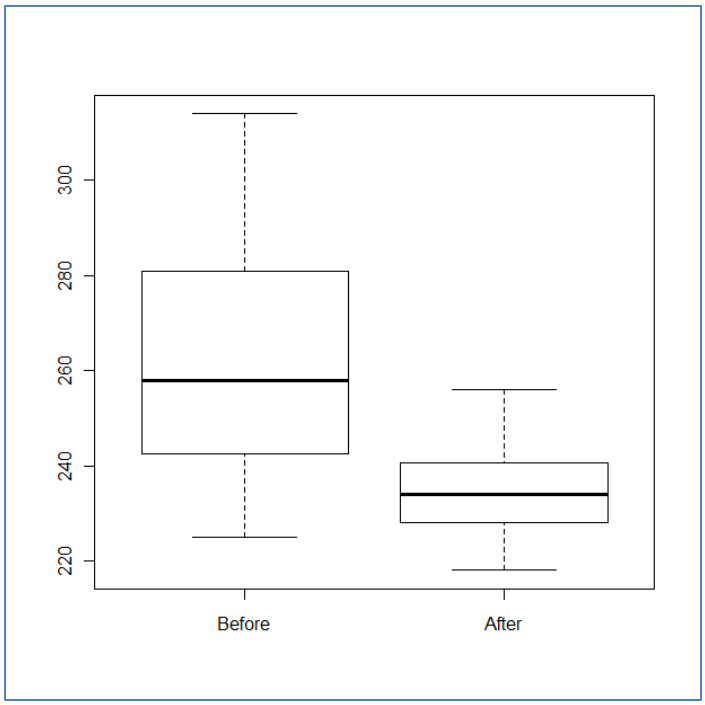
Tương tự như trường hợp hai mẫu độc lập, sử dụng `paired = TRUE`.

Ví dụ 2:

File `cholesterol.txt` chứa hàm lượng cholesterol trong máu đo trên 15 người đàn ông béo phì từ ở độ tuổi 35 – 50. Những người tham gia thử nghiệm sẽ thực hiện 1 chế độ ăn kiêng và tập thể dục trong vòng 3 tháng. Bác sĩ sẽ đo hàm lượng cholesterol trong máu mỗi người trước và

sau khi thử nghiệm. Với mức ý nghĩa 5%, chế độ ăn kiêng và tập thể dục có hiệu quả trong việc giảm hàm lượng cholesterol trong máu những người béo phì này không?

```
> cholesterol = read.table('cholesterol.txt',header=T)
> cholesterol
> boxplot(cholesterol[2:3])
```




```
> attach(cholesterol)
> v.equal <- var.test(Before,After,data=cholesterol)
        F test to compare two variances
data:  Before and After
F = 5.6691, num df = 14, denom df = 14, p-value = 0.002529
```

p giá trị bằng 0.002529, suy ra phương sai khác nhau.

```
> p.v.equal <- v.equal$p.value
> alpha <- 0.05
> v.equal <- ifelse(p.v.equal <= alpha, FALSE, TRUE)
> t.test(Before,After, paired=TRUE, var.equal = v.equal,
alternative = 'greater', data=cholesterol)
```

```
        Paired t-test


data:  Before and After
t = 5.4659, df = 14, p-value = 4.158e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 18.20922      Inf
sample estimates:
mean of the differences
               26.86667
```

p giá trị < 0.05 và dựa vào đồ thị boxplot ta kết luận rằng với mức ý nghĩa 5%, chế độ ăn kiêng và tập thể dục có hiệu quả trong việc giảm hàm lượng cholesterol trong máu của những người đàn ông béo phì.

III. So sánh 2 tỷ lệ:

Gọi p_1 và p_2 lần lượt là tỷ lệ của các phần tử thoả một tính chất quan tâm tương ứng với hai tổng thể. Từ hai tổng thể chọn ra hai mẫu độc lập với cỡ mẫu lần lượt là n_1 và n_2 (n_1 và n_2 lớn).

Ta cần kiểm định các giả thuyết sau: $\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{cases}$ $\begin{cases} H_0 : p_1 \leq p_2 \\ H_1 : p_1 > p_2 \end{cases}$ với mức ý nghĩa α

Gọi Y_1 và Y_2 là số phần tử thoả tính chất quan tâm trong hai mẫu độc lập được chọn từ hai tổng thể.

Các tỷ lệ mẫu: $\hat{p}_1 = \frac{Y_1}{n_1}; \hat{p}_2 = \frac{Y_2}{n_2}; \hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}$

Tính thống kê kiểm định: $Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Kết luận:

Giả thuyết H_0	Đối thuyết H_1	Miền bác bỏ với mức ý nghĩa α	P – giá trị
$p_1 = p_2$	$p_1 \neq p_2$	$\lvert Z_0 \rvert > z_{1-\alpha/2}$	$2P\{Z \geq \lvert Z_0 \rvert\}$

$p_1 \le p_2$	$p_1 > p_2$	$Z_0 > z_{1-\alpha}$	$P\{Z \ge Z_0\}$

Bảng 3

Thực hành trong R:

Sử dụng hàm `prop.test`

```
prop.test(y, n, alternative = c("two.sided", "less", "greater"),
                conf.level = 0.95, correct = TRUE)
```

Trong đó:
$y = (y_1, y_2)$: Số phần tử thoả tính chất quan tâm tương ứng với mẫu 1 và 2.
$n = (n_1, n_2)$: cỡ mẫu 1 và 2.

Ví dụ 3:

Trong một nhà máy sản xuất linh kiện điện tử, các kỹ sư thử nghiệm một phương pháp sản mới trong sản xuất vi mạch. Các kỹ sư cho rằng phương pháp mới hiệu quả hơn phương pháp cũ trong việc làm giảm tỷ lệ vi mạch hỏng. Để kiểm tra, 320 vi mạch được sản xuất theo phương pháp mới và 360 vi mạch được sản xuất theo phương pháp cũ. Kết quả cho thấy trong 320 vi mạch có 76 vi mạch hỏng và trong 360 vi mạch sản xuất theo phương pháp cũ có 94 vi mạch hỏng. Với mức ý nghĩa 5% và dựa vào dữ liệu mẫu, có thể cho rằng phương pháp mới làm giảm tỷ lệ vi mạch hỏng hay không?

Gọi

p_1 : tỷ lệ vi mạch hỏng khi sản xuất theo phương pháp cũ

p_2 : tỷ lệ vi mạch hỏng khi sản xuất theo phương pháp mới

Ta cần kiểm định giả thuyết $\begin{cases} H_0 : p_1 \le p_2 \\ H_1 : p_1 > p_2 \end{cases}$

```
> y = c(94,76)
> n = c(360,320)
> prop.test(y,n,alternative = 'greater')
        2-sample test for equality of proportions with
continuity correction
data:  y out of n
X-squared = 0.3856, df = 1, p-value = 0.2673
alternative hypothesis: greater
95 percent confidence interval:
```

```
 -0.03393952  1.00000000
sample estimates:
   prop 1    prop 2
0.2611111 0.2375000
```

P - giá trị = 0.2673 > 0.05 => không có đủ cơ sở để bác bỏ giả thuyết H_o, chưa đủ bằng chứng để kết luận rằng phương pháp mới có hiệu quả hơn phương pháp cũ

BÀI TẬP

1. Hai máy máy rót sữa tự động được sử dụng để đưa sữa vào hộp giấy có dung tích 1 lít trong một dây chuyền sản xuất. Lượng sữa thực tế hai máy đưa vào hộp có phân phối chuẩn với độ lệch tiêu chuẩn lần lượt là 0.002 và 0.0025 lít. Một thành viên trong số các kỹ sư giám sát dây chuyền cho rằng thể tích tích sữa trung bình được hai máy đưa vào các hộp là như nhau và không phụ thuộc vào dung tích của các hộp chứa. Một mẫu ngẫu nhiên được lấy từ hai máy này được cho trong file *volume.csv*.

a) Bạn có cho rằng phán đoán của kỹ sư trên là đúng hay không? Sử dụng $\alpha = 0.05$.
b) P-giá trị của kiểm định trên là bao nhiêu?
c) Hãy tìm khoảng tin cậy 95% cho sự khác biệt về trung bình lượng sữa hai máy đưa vào các hộp.
d) Viết hàm **test.leq.oneside(x, y, μ_0, $\sigma_1, \sigma_2, \alpha$)** để kiểm định giả thiết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$ trong đó $\mu = \mu_1 - \mu_2$ với μ_1, μ_2 lần lượt là thể tích sữa trung bình hai máy 1 và 2 đưa vào các hộp; **x, y** là hai vector dữ liệu chứa thể tích sữa do máy số 1 và máy số 2 đưa vào các hộp giấy; α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.
e) Viết hàm **test.geq.oneside(x, y, μ_0, $\sigma_1, \sigma_2 \alpha$)** để kiểm định giả thiết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$ với các ký hiệu và yêu cầu như ở câu trên.

2. Đường kính (mm) của các thanh thép được sản xuất bởi hai máy cán thép tại một nhà máy được đem ra so sánh. Hai mẫu ngẫu nhiên về đường kính của các thanh thép được sản xuất bởi hai máy này được thu thập và tổng hợp trong file *diameter.csv*. Giả sử rằng đường kính của các thanh thép được cán bởi hai máy này có phân phối chuẩn với phương sai bằng nhau.

a) Hỏi có bằng chứng ủng hộ giả thiết hai máy cho ra các thanh thép với đường kính khác nhau hay không? Hãy kiểm tra và đưa ra kết luận với mức ý nghĩa $\alpha = 0.05$.
b) Hãy tìm P-giát trị cho thống kê bạn tính ở câu vừa rồi.
c) Hãy ước lượng sự sai biệt về đường kính trung bình của các thanh thép do hai máy này sản xuất với độ tin cậy 95%.

d) Viết hàm **test.leq.oneside(x, y, μ_0, α)** để kiểm định giả thiết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$ trong đó $\mu = \mu_1 - \mu_2$ với μ_1, μ_2 lần lượt là đường kính trung bình của các thanh thép do máy số 1 và máy số 2 sản xuất ra; **x, y** là hai vector dữ liệu chứa đường kính của các thanh thép do máy số 1 và máy số 2 sản xuất ra; α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.

e) Viết hàm **test.geq.oneside(x, y, μ_0, α)** để kiểm định giả thiết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$ với các ký hiệu và yêu cầu như ở câu trên.

3. Một hãng công nghệ (G) sản xuất hệ điều hành (HĐH) cho máy tính công bố rằng tốc độ khởi động của HĐH mới của họ cho thời gian khởi động nhanh gấp đôi HĐH mới nhất của hãng đối thủ (M). Một mẫu khảo sát gồm 250 laptop cùng cấu hình, trong đó 100 laptop được cài HĐH của hãng (G) và 150 laptop còn lại được cài HĐH của hãng đối thủ (M). Kết quả được ghi lại trong file *"tg.khoidong.csv"*. Giả sử độ lệch chuẩn về thời gian khởi động HĐH của hãng (G) và (M) lần lượt là $\sigma_1 = 1.5$ (s) và $\sigma_2 = 1$ (s). Hỏi công bố về thời gian khởi động của hãng (G) có phù hợp với số liệu thu thập được? Với mức ý nghĩa 5\%.

4. Hai mươi người nam với độ tuổi từ 35 đến 50 tham gia vào một nghiên cứu để đánh giá sự ảnh hưởng của chế độ ăn uống và luyện tập thể thao lên hàm lượng cholesterol trong máu. Mỗi người tham gia được đo lượng cholesterol trước khi bắt đầu chương trình tập luyện thể dục và chuyển sang ăn uống với chế độ ít chất béo. Dữ liệu được cho trong file *cholesterol.csv*.

a) Hỏi rằng dữ liệu có ủng hộ kết luận rằng chế độ ăn kiêng và luyện tập thể dục đã có tác dụng trong việc giảm lượng cholesterol trong máu hay không? Hãy tiến hành ở mức ý nghĩa 0.05.

b) Giả sử hàm lượng cholesterol ở mỗi người tham gia nghiên cứu trước và sau khi bắt đầu chương trinh luyện tập thể dục thể thao có phân phối chuẩn với phương sai khác nhau. Viết hàm **test.leq.oneside(x, y, μ_0, α)** để kiểm định giả thiết $H_1 : \mu = \mu_0$ và đối thiết $H_1 : \mu < \mu_0$ trong đó $\mu = \mu_1 - \mu_2$ với μ_1, μ_2 lần lượt là trung bình hàm lượng cholesterol trước và sau khi tham gia chương trình luyện tập, α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.

c) Với cùng giả định như ở câu b, hãy viết hàm **test.geq.oneside(x, y, μ_0, α)** để kiểm định giả thiết $H_1 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$ với các ký hiệu và yêu cầu như ở câu trên.

5. Một sản phẩm ăn kiêng dạng lỏng được nhà sản xuất quảng cáo có tác dụng giảm cân ít nhất 1.5 kg khi sử dụng sản phẩm trong 1 tháng. Một mẫu ngẫu nhiên 50 người sử dụng sản phẩm trong một tháng, kết quả trọng lượng (kg) trước khi sử dụng sản phẩm và sau khi dùng 1 tháng được ghi lại trong file *'giamcan.csv'*. Hãy sử dụng các thủ tục kiểm định giả thuyết để trả lời các câu hỏi sau

(a) Dữ liệu có phù hợp với quảng cáo của nhà sản xuất về sản phẩm ăn kiêng hay không với mức ý nghĩa 0.05?

(b) Dữ liệu có phù hợp với quảng cáo của nhà sản xuất về sản phẩm ăn kiêng hay không với mức ý nghĩa 0.01?

(c) Để cải thiện doanh số bán hàng, nhà sản xuất đang xem xét thay đổi quảng cáo từ "ít nhất 1.5 kg" thành "ít nhất 2.5 kg". Thực hiện lại câu (a), (b) để kiểm định quảng cáo mới này.

6. Một nhà khoa học máy tính tiến hành kiểm tra sự hữu dụng của hai ngôn ngữ thiết kế khác nhau trong việc cải thiện các tác vụ lập trình. Hai mươi chuyên gia lập trình thành thạo với cả hai ngôn ngữ này được yêu cầu lập trình cho cùng một module chức năng trên hai ngôn ngữ thiết kế. Thời gian hoàn thành công việc (tính theo phút và tuân theo phân phối chuẩn) được ghi lại và tóm tắt trong file *pro.time.csv.*

a) Hãy xác định khoảng tin cậy 95% cho sự sai khác giữa thời gian lập trình trung bình trên hai ngôn ngữ. Có dấu hiệu nào cho thấy một trong hai ngôn ngữ là tốt hơn hay không?

b) Viết hàm **test.leq.oneside(x, y,** μ_0, α**)** để kiểm định giả thiết $H_1 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$ trong đó $\mu = \mu_1 - \mu_2$ với μ_1, μ_2 lần lượt là thời gian trung bình của các lập trình viên viết xong module đối với ngôn ngữ thứ nhất và thứ hai, α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.

c) Viết hàm **test.geq.oneside(x, y,** μ_0, α**)** để kiểm định giả thiết $H_1 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$ với các ký hiệu và yêu cầu như ở câu trên..

7. Hai loại máy đúc khác nhau được sử dụng để gia công các chi tiết bằng plastic. Một chi tiết được xem là phế phẩm nếu nó có sự thay đổi về hình dạng và khối lượng hoặc sai biệt về màu sắc so với chi tiết mẫu vượt quá mức cho phép. Hai mẫu ngẫu nhiên được thu thập, mỗi mẫu có cỡ 300. Trong mẫu thu được từ máy số 1 người ta nhận thấy có 15 chi tiết là phế phẩm và đối với mẫu thu được từ máy thứ 2 có 8 chi tiết là phế phẩm. Liệu có cơ sở để kết luận rằng hai loại máy có tỉ lệ phế phẩm như nhau hay không (mức ý nghĩa 0.025)? Tìm P-giá trị cho kiểm định vừa rồi.

8. Trong một mẫu ngẫu nhiên gồm 500 người sống tại thành phố Hồ Chí Minh người ta thấy có 385 người ủng hộ việc tăng tốc độ lưu thông trên các tuyến đường quốc lộ lên 65 km/h. Trong khi đó một mẫu ngẫu nhiên khác gồm 400 người sống tại Hà Nội cho thấy có 267 người ủng hộ việc tăng tốc độ. Liệu số liệu thu thập được có cho ta kết luận rằng có sự khác biệt về tỉ lệ người ủng hộ việc tăng tốc độ tại hai thành phố hay không ($\alpha = 0.05$)? P-giá trị cho kiểm định trên là bao nhiêu?

9. Trong một mẫu ngẫu nhiên 200 tài xế ở Tp.HCM có 165 trường hợp mang đai an toàn thường xuyên, trong khi một mẫu khác, 250 tài xế ở Hà Nội, cho thấy có 198 trường hợp mang đai an toàn thường xuyên.

(a) Thực hiện thủ tục kiểm định giả thuyết để xác định xem có sự khác biệt mang ý nghĩa thống kê giữa việc sử dụng đai an toàn ở Tp.HCM và Hà Nội hay không? Với mức ý nghĩa 0.05

(b) Tương tự câu (a) với mức ý nghĩa 0.1

(c) Giả sử các số liệu ở trên được nhân đôi. Nghĩa là, trong một mẫu ngẫu nhiên 400 tài xế ở Tp.HCM có 330 trường hợp mang đai an toàn thường xuyên, trong khi một mẫu khác, 500 tài xế ở Hà Nội cho thấy có 396 trường hợp mang đai an toàn thường xuyên. Thực hiện lại câu (a), (b) và cho nhận xét về ảnh hưởng của việc tăng kích thước mẫu mà không thay đổi các tỉ lệ lên kết quả kiểm định.

10. Doanh số bán hàng (triệu đồng) của hai chi nhánh trong hệ thống siêu thị Co-op Mart đặt tại quận 1 và quận 3 trong 200 ngày được cho trong file *Profit.csv.*

a) Hãy xác định khoảng tin cậy 95% cho sự sai khác về doanh số bán hàng trung bình của hai chi nhánh này.
b) Một ngày được gọi là có doanh số cao nếu doanh số bán hàng trong ngày đó trên 600 triệu đồng. Có nhận xét rằng tỉ lệ các ngày có doanh số cao tại chi nhánh quận 1 cao hơn tỉ lệ các ngày có doanh số bán hàng cao tại chi nhánh quân 3. Hãy kiểm chứng nhận xét này với $\alpha = 0.05$. Hãy cho biết P-giá trị của kiểm định vừa rồi.
c) Viết hàm **prop.test.leq(x, y,α)** để kiểm định giả thiết $H_0: p_x = p_y$ và đối thuyết $H_1: p_x < p_y$ trong đó p_x, p_y lần lượt là tỉ lệ ngày có doanh số bán cao (trên 600 triệu) tại chi nhánh quận 1 và quận 3; **x, y** lần lượt là hai vector chứa doanh số bán hàng tại hai chi nhánh của Co-op Mart đặt tại quận 1 và quận 3 trong một số ngày nào đó; α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.

11. Thu nhập của các nhân viên trong ngành công nghệ thông tin tại hai thành phố Hà Nội và thành phố Hồ Chí Minh được cho trong file *Inf.Sal.csv*.

a) Hãy xác định khoảng tin cậy 95% cho sự sai khác về mức lương trung bình của nhân viên trong ngành công nghệ thông tin tại hai thành phố.
b) Một người được gọi là có thu nhập cao nếu thu nhập của người đó trên 11,5 triệu đồng. Có nhận xét rằng tỉ lệ nhân viên công nghệ thông tin có thu nhập cao tại thành phố Hồ Chí Minh cao hơn so với tỷ lệ này tại Hà Nội. Hãy kiểm chứng nhận xét này với $\alpha = 0.025$. Hãy cho biết P-giá trị của kiểm định vừa rồi.
c) Viết hàm **prop.test.geq(x, y,α)** để kiểm định giả thiết $H_0: p_x = p_y$ và đối thuyết $H_1: p_x > p_y$ trong đó p_x, p_y lần lượt là tỉ lệ người có thu nhập cao (trên 11,5 triệu đồng) tại thành phố Hồ Chí Minh và Hà Nội; **x, y** là hai vector chứa giá trị thu nhập của một số nhân viên ngành công nghệ thông tin tại Hà Nội và thành phố Hồ Chí Minh; α là mức ý nghĩa. Hàm xuất ra kết quả chấp nhận hoặc bác bỏ và cho biết P-giá trị.
d) Tìm khoảng tin cậy 95% cho hiệu hai tỉ lệ trên.