Bài 8.

KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

I. Giới thiệu

1. Định nghĩa một giả thuyết thống kê (statistical hypothesis) là một phát biểu (tuyên bố) về các tham số của một hay nhiều tổng thể.

Ví dụ: Giả sử ta cần nghiên cứu về thu nhập của người dân thành phố, gọi μ là thu nhập trung bình, ta cần xác định xem thu nhập trung bình của một người dân có bằng 7 triệu đồng/tháng hay không? Giả thuyết được phát biểu như sau

H_0: μ = 7 triệu/tháng

H_1: $\mu \neq$ 7 triệu/tháng

Phát biểu H_0: μ = 7 gọi là giả thuyết không (null hypothesis), phát biểu H_1: $\mu \neq 7$ gọi là đối thuyết (alternative hypothesis). Bài toán kiểm định giả thuyết như trên gọi là kiểm định giả thuyết hai phía. Trong một số trường hợp, ta có bài toán kiểm định giả thuyết một phía, chẳng hạn như

$$\begin{cases} H_0 : \mu \geq 7 \\ H_1 : \mu < 7 \end{cases} \quad \text{hoặc} \quad \begin{cases} H_0 : \mu \leq 7 \\ H_1 : \mu > 7 \end{cases}$$

Tổng quát, xét biến ngẫu nhiên X có phân phối $F(x;\theta)$, tham số θ chưa biết. Với một giá trị θ_0 cho trước, bài toán kiểm định giả thuyết cho tham số θ gồm các dạng sau

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases} \qquad \begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \qquad \begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

2. Các loại sai lầm trong kiểm định giả thuyết

Sai lầm loại I: Nếu ta bác bỏ H_0 khi H_0 đúng thì sai lầm đó gọi là sai lầm loại I, ký hiệu α. α còn được gọi là *mức ý nghĩa* (significance level) của kiểm định.

Sai lầm loại II: Nếu H_0 sai mà ta không bác bỏ H_0 thì sai lầm đó gọi là sai lầm loại II, ký hiệu β.

Quyết định	H_0 đúng	H_0 sai
Không bác bỏ H_0	Đúng	Sai lầm loại II
Bác bỏ H_0	Sai lầm loại I	Đúng

3. *P* – giá trị (*P* – value) Với một giả thuyết không H_0 và mẫu cỡ n cho trước, P – giá trị là mức ý nghĩa nhỏ nhất dẫn đến việc bác bỏ giả thuyết H_0, P – giá trị được tính dựa theo giá trị thống kê kiểm định.

II. Kiểm định giả thuyết cho kỳ vọng

(Trong thực hành, ta chỉ xét trường hợp không biết phương sai)

Giả thuyết: mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ được chọn từ tổng thể có phân phối chuẩn (hoặc xấp xỉ chuẩn tức phân phối có dạng đối xứng) với kỳ vọng μ và phương sai σ^2.

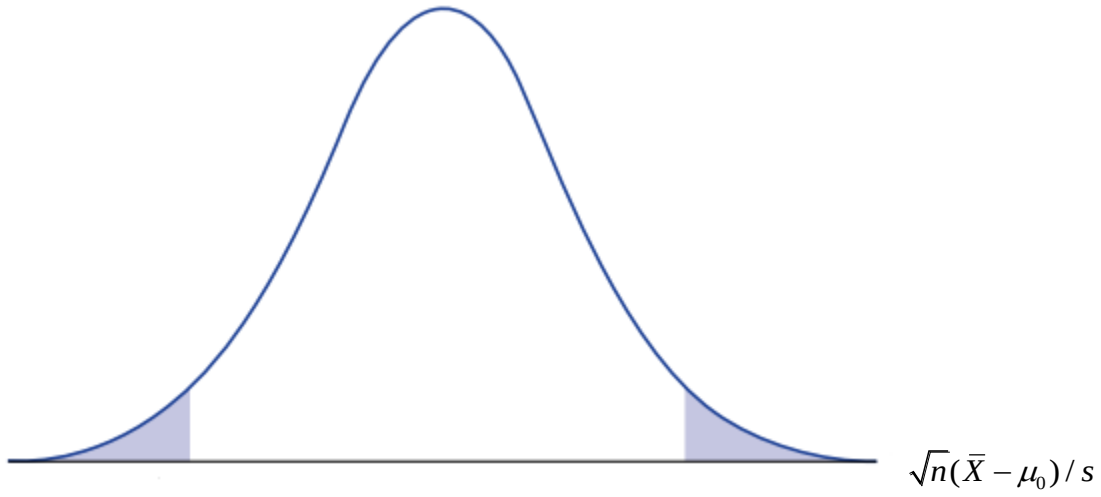
Giả thuyết: $H_0 : \mu = \mu_0$ Đối thuyết: $H_1 : \begin{cases} \mu \neq \mu_0 \\ \mu < \mu_0 \\ \mu > \mu_0 \end{cases}$ (Một trong 3 trường hợp)

Tính thống kê kiểm định:

$$T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Miền bác bỏ:

Với $H_1 : \mu \neq \mu_0$	bác bỏ H_0 nếu $T_0 < -t_{1-\alpha/2}^{n-1}$ hoặc $T_0 > t_{1-\alpha/2}^{n-1}$
Với $H_1 : \mu < \mu_0$	bác bỏ H_0 nếu $T_0 < -t_{1-\alpha}^{n-1}$
Với $H_1 : \mu > \mu_0$	bác bỏ H_0 nếu $T_0 > t_{1-\alpha}^{n-1}$



Miền bác bỏ trong trường hợp kiểm định giả thuyết hai phía

Trong **R**, để tìm phân vị $t_{1-\alpha/2}^{n-1}$ sử dụng hàm `qt(1-alpha/2,n-1)`.

Trong kết quả do **R** xuất ra, ta xác định có bác bỏ H_0 hay không thông qua P – giá trị:

***Quy tắc:* Khi *P* – giá trị bé hơn α, bác bỏ H_0.**

Khi cỡ mẫu *n* lớn, phân phối của thống kê T_0 sẽ trở thành phân phối chuẩn hoá N(0,1), khi đó giá trị tiêu chuẩn dùng để so sánh là $z_{1-\alpha/2}$ (dùng `qnorm(1-alpha/2)`).

Bảng tính p – giá trị:

Giả thuyết H_0	Đối thuyết H_1	Thống kê T_0	P – giá trị
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\bar{X} - \mu_0}{s}\sqrt{n}$	$2P\{T_{n-1} \geq \lvert T_0 \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$		$P\{T_{n-1} \geq T_0\}$
$\mu \geq \mu_0$	$\mu < \mu_0$		$P\{T_{n-1} \leq T_0\}$

(T_{n-1} là biến ngẫu nhiên có phân phối Student với n – 1 bậc tự do)

Trong **R**, để tính $P\{T_{n-1} \leq t_0\}$ sử dụng hàm `pt(t0, n - 1)`; nếu muốn tính $P\{T_{n-1} \geq t_0\}$, sử dụng `pt(t0, n - 1, lower.tail = FALSE)`.

Sử dụng hàm t.test để kiểm định

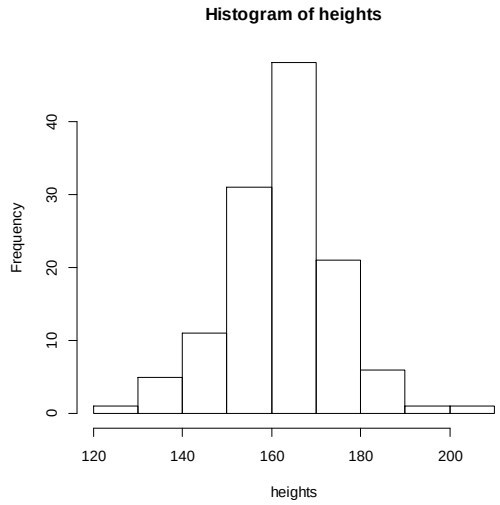
```
t.test(x,alternative = c("two.sided", "less", "greater"),
          mu = mu_0, conf.level = 0.95)
```

Trong đó:

- `x` : véc-tơ dữ liệu
- `alternative` : xác định kiểm định là hai phía ("`two.sided`"), bên trái ("`less`") hay bên phải ("`greater`"), mặc định là `two.sided`.
- `mu = mu_0` : giá trị cần kiểm định
- `conf.level` : xuất ra khoảng tin cậy với độ tin cậy tương ứng

Ví dụ: Biến heights chứa chiều cao của 125 thanh niên trong một khu vực (để mở heights, load tập tin "heights.rda"). Hãy kiểm định chiều cao của thanh niên trong khu vực có bằng 160 cm hay không, với mức ý nghĩa 5%? Đồng thời xác định khoảng tin cậy 95% cho chiều cao trung bình thanh niên trong khu vực này.

```
> load('heights.rda')
> summary(heights)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  129.3   156.2   163.4   162.8   169.6   209.3
> hist(heights)
```




```
> t.test(heights,mu = 160, conf.level=0.95)
        One Sample t-test


data:  heights
t = 2.6175, df = 124, p-value = 0.009959
alternative hypothesis: true mean is not equal to 160
95 percent confidence interval:
 160.6941 164.9990
sample estimates:
mean of x
 162.8465
```

Dùng lệnh `hist` để vẽ đồ thị histogram cho dữ liệu để kiểm tra phân phối của dữ liệu, cho nhận xét?

Lệnh `t.test` cho các kết quả sau:

Thống kê kiểm định t = 2.6175, bậc tự do n – 1 = 124, p – value = 0.00996.

Khoảng tin cậy 95%: $160.6941 \leq \mu \leq 164.9990$

Với mức ý nghĩa 5%, ta thấy p – value < 0.05, do đó bác bỏ H_0 tức là chiều cao trung bình của thanh niên trong khu vực khác 160 cm.

Nếu sử dụng giá trị thống kê t = 2.6175, ta so sánh với $t_{1-\alpha/2}^{n-1} = t_{0.975}^{124} \approx z_{0.975} = 1.96$ (dùng lệnh `qt(0.975,124)` hoặc `qnorm(0.975)`), ta cũng có kết luận tương tự.

Cho nhận xét: có thể sử dụng khoảng tin cậy 95% để xác định có bác bỏ giả thuyết H_0 hay không tại mức ý nghĩa 5%?

Chú ý:

Trong ví dụ trên, nếu ta đặt:

```
> result <- t.test(heights,mu = 162, conf.level=0.95)
```

thì ta có thể rút ra thống kê kiểm định t bằng lệnh `result$statistic` (hoặc `result[['statistic']]` hoặc `unname(result[['statistic']]))`:

Danh sách đầy đủ tên các đối tượng trả về:

```
> names(result)
[1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
[6] "null.value"  "alternative" "method"      "data.name"
```

III. Kiểm định giả thuyết cho tỷ lệ

Giả sử cần kiểm định tỷ lệ phầ ntử thoả tính chất A trong tổng thể. Khảo sát một mẫu cỡ *n* gồm n biến ngẫu nhiên $Y_1, Y_2, \ldots, Y_n$ với

$$Y_i = \begin{cases} 1 & \text{, nếu thoả A} \\ 0 & \text{, nếu không thoả A} \end{cases}$$

Gọi: $Y = \sum_{i=1}^{n} Y_i$ là tổng số phần tử thoả tính chất A trong *n* phần tử khảo sát, suy ra tỷ lệ mẫu

$$\hat{P} = \frac{Y}{n}$$

Giả thuyết: cỡ mẫu khảo sát *n* phải tương đối lớn.

Giả thuyết: $H_0 : p = p_0$ Đối thuyết: $H_1 : \begin{cases} p \neq p_0 \\ p < p_0 \\ p > p_0 \end{cases}$ (Một trong 3 trường hợp)

Tính thống kê kiểm định:

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Miền bác bỏ:

Với $H_1 : p \neq p_0$	bác bỏ H_0 nếu $Z_0 < -z_{1-\alpha/2}$ hoặc $Z_0 > z_{1-\alpha/2}$
Với $H_1 : p < p_0$	bác bỏ H_0 nếu $Z_0 < -z_{1-\alpha}$
Với $H_1 : p > p_0$	bác bỏ H_0 nếu $Z_0 > z_{1-\alpha}$

Để tìm $z_{1-\alpha/2}$, sử dụng hàm `qnorm(1-alpha/2)`.

Sử dụng hàm `prop.test` để kiểm định:

```
prop.test(y, n, p = p0,
          alternative = c("two.sided", "less", "greater"),
```

```
                    conf.level = 0.95)
```

Các tham số:

- `y` : số phần tử thoả tính chất A trong n phần tử khảo sát
- `n` : cỡ mẫu
- `alternative` : xác định kiểm định là hai phía ("`two.sided`"), bên trái ("`less`") hay bên phải ("`greater`")
- `p = p0` : giá trị cần kiểm định
- `conf.level` : xuất ra khoảng tin cậy với độ tin cậy tương ứng

Ví dụ: Trong một cuộc bầu cử thị trưởng tại một thành phố, ứng cử viên A tin rằng có trên 50% người dân thành phố ủng hộ ông ta. Để kiểm định điều này, các chuyên gia thống kê chọn ngẫu nhiên 800 người dân trong thành phố, thấy có 448 người dân cho ý kiến ủng hộ ông A. Hãy xét xem tuyên bố của ông A về tỷ lệ cử tri có đúng không với mức ý nghĩa 1%.

Ta có:

- Cỡ mẫu kháo sát n = 800.
- Số người dân ủng hộ ông A y = 448.
- Giả thuyết cần kiểm tra $\begin{cases} H_0 : p = 0.5 \\ H_1 : p > 0.5 \end{cases}$

Trong đó: p là tỷ lệ người dân thành phố ủng hộ ông A.

Sử dụng hàm `prop.test`:

```
> n = 800; y = 448
> prop.test(y,n,p=0.5,alternative="greater",conf.level=0.99)
        1-sample proportions test with continuity correction

data:  y out of n, null probability 0.5
X-squared = 11.2812, df = 1, p-value = 0.0003915
alternative hypothesis: true p is greater than 0.5
99 percent confidence interval:
 0.5182781 1.0000000
sample estimates:
   p
0.56
```

- Kết quả cho biết p – giá trị = 0.0003915 < 1% dẫn đến bác bỏ giả thuyết H_0, ta kết luận rằng tỷ lệ người dân ủng hộ ông A trong thành phố trên 50%.
- Khoảng tin cậy 99% cho tỷ lệ p là: 0.5182 ≤ p ≤ 1.0000.

Trong trường hợp kiểm định 1 mẫu, hàm `binom.test` cũng cho kết quả tương tự

```
> binom.test(y,n,p=0.5,alternative="greater",conf.level=0.99)
```

```
                        Exact binomial test

data:  y and n
number of successes = 448, number of trials = 800, p-value = 0.0003864
alternative hypothesis: true probability of success is greater than
0.5
99 percent confidence interval:
 0.5183309 1.0000000
sample estimates:
probability of success
                 0.56
```

Bảng tính p – giá trị:

Giả thuyết H_0	Đối thuyết H_1	Thống kê Z_0	P – giá trị
$p = p_0$	$p \neq p_0$	$\dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$	$2\min\{\Phi(Z_0), 1-\Phi(Z_0)\}$
$p \leq p_0$	$p > p_0$		$1-\Phi(Z_0)$
$p \geq p_0$	$p < p_0$		$\Phi(Z_0)$

Với $\Phi(Z_0) = P\{Z \leq Z_0\}$: hàm Laplace – hàm phân phối của biến ngẫu nhiên chuẩn hoá N(0,1).

Trong **R**, $\Phi(Z_0)$ = `pnorm(`Z_0`)` .

Bài tập

1. Số liệu thống kê về doanh số bán hàng của một siêu thị cho ở file ***profit.csv***:

a) Vẽ đồ thị histogram cho dữ liệu, có nhận xét gì về phân phối của dữ liệu.

b) Những ngày có doanh số bán trên 65 triệu đồng là những ngày bán đắt hàng. Hãy ước lượng doanh số bán trung bình của một ngày "bán đắt hàng" ở siêu thị này với độ tin cậy 99% (giả thuyết doanh số bán của những ngày bán đắt hàng là đại lượng ngẫu nhiên phân phối theo quy luật chuẩn).

c) Trước đây doanh số bán trung bình của siêu thị là 60 triệu đồng/ngày. Số liệu của bảng trên được thu thập sau khi siêu thị áp dụng một phương thức bán hàng mới. Hãy cho nhận xét về phương thức bán hàng mới với mức ý nghĩa là 1%.

2. Sau một đợt bồi dưỡng sư phạm, người ta kiểm tra ngẫu nhiên 70 học viên. Kết quả cho bởi bảng sau (thang điểm là 10):

Điểm (x_i)	5	6	7	8	9	10
Tần số (n_i)	5	10	15	20	12	8

Giả sử điểm số của các học viên tuân theo phân phối chuẩn. Có ý kiến cho rằng điểm số trung bình là 8. Hãy kiểm tra ý kiến trên ở mức α = 5%.

a) Chỉ ra cách biến đổi số liệu của X để sử dụng được hàm t.test. Vẽ biểu đồ stem & leaf cho số liệu đã biến đổi.

b) Viết hàm test.geq.oneside(x, μ_0, alpha) để kiểm định giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$. Xuất ra kết luận và *p* - giá trị. Áp dụng để kiểm định $H_1 : \mu > 8$. (Lưu ý: không gọi hàm t.test)

c) Viết hàm test.leq.oneside(x, μ_0, alpha) để kiểm định giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$. Xuất ra thông báo và *p* - giá trị. Áp dụng để kiểm định $H_1 : \mu < 8$. (Lưu ý: không gọi hàm t.test)

3. Sử dụng số liệu trong file teen-birth-rate-2002.txt để tạo bảng sau (xuất ra màn hình)

Bang: Ty le sinh tre em duoi gia thiet phan phoi chuan

	$\bar{X}$	**S**	***n***	***Z***	***p***
Black	76.14	15.60	44	4.79	0
Hispanic	88.96	23.66	48	7.05	0
White	32.51	11.71	51	-19.74	0

Trong đó: $\bar{X}, s, n$ lần lượt là trung bình mẫu, độ lệch tiêu chuẩn của mẫu và cỡ mẫu của các biến tương ứng; $Z = \left(\bar{X} - \mu_0\right)\sqrt{n}/s$ với μ_0 là trung bình tổng cộng của ba biến Black, Hispanic và White; *p* là *p* - giá trị tương tứng với từng thống kê Z.

4. Hội đồng khoa học Trường ĐH Khoa học Tự nhiên muốn thay đổi đánh giá học lực của sinh viên từ thang điểm 10 sang thang điểm 4. HĐ Khoa học quyết định khảo sát ý kiến các giảg viên trong trường trước khi ra quyết định. Nếu như tỷ lệ giảng viên đồng ý với sự thay đổi trên 60% thì việc thay đổi thang điểm sẽ được thực hiện. Khảo sát ngẫu nhiên 80 giảng viên trong trường, gọi p là tỷ lệ giảng viên đồng ý với sự thay đổi.

a) Hãy chọn 1 cặp giả thuyết/ đối thuyết thích hợp với yêu cầu kiểm định của bài toán:

(1) $H_0 : p = 0.6$ và $H_1 : p < 0.6$

(2) $H_0 : p = 0.6$ và $H_1 : p > 0.6$

Giải thích sự lựa chọn.

b) Biến *survey* có trong tập tin data04.rda chứa kết quả khảo sát ý kiến của 80 giảng viên (0: không đồng ý, 1: đồng ý). Theo kết quả khảo sát, HĐ Khoa học có nên thay đổi thang điểm hay không? Mức ý nghĩa 5%.

5. Tại một đợt khám sức khỏe của trẻ em ở nhà trẻ, người ta khám ngẫu nhiên 100 cháu thấy có 20 cháu có hiện tượng còi xương do suy dinh dưỡng. Gọi *p* là xác suất để bắt gặp 1 trẻ mắc bệnh còi xương. Hãy kiểm định giả thuyết H_0: $p = 0,15$ và đối thuyết H_1: $p \neq 0,15$ ở mức $\alpha = 5\%$.

6. File ***times.csv*** chứa thời gian tự học mỗi ngày của sinh viên hai trường Khoa học Tự nhiên và Kinh tế.

a) Có ý kiến cho rằng tỷ lệ sinh viên có thời gian tự học trên 5 giờ mỗi ngày của sinh viên trường Khoa học Tự nhiên là 50%. Với mức ý nghĩa 5%, hãy kiểm tra ý kiến này.

b) Viết hàm proptest.geq(f, n, p_0, alpha) để kiểm định giả thuyết $H_0 : p = p_0$ và đối thuyết $H_1 : p > p_0$ trong đó f là số phần tử thoả tính chất quan tâm trong n phần tử khảo sát. Xuất ra kết quả và giá trị P. Áp dụng cho câu a. với đối thuyết H_1: p > 0.5. (Lưu ý: không gọi hàm prop.test hay binom.test)

c) Viết hàm proptest.leq(f, n, p_0, alpha) để kiểm định giả thuyết $H_0 : p = p_0$ và đối thuyết $H_1 : p < p_0$. Xuất ra kết quả và giá trị P. Áp dụng cho câu a. với đối thuyết H_1: p < 0.5. (Lưu ý: không gọi hàm prop.test hay binom.test)

7. Hãy viết hàm `test.mean( …)` với tính năng tương tự hàm `t.test` trong **R**: thực hiện kiểm định hai phía, một phía (bên trái, bên phải) cho kỳ vọng; xuất ra thông báo, giá trị trả về gồm trung bình mẫu của dữ liệu vào, p – giá trị. (Tham khảo hàm zpfunction(…) sau để làm mẫu)

```
zp <- function(y , n , p0, alpha = 0.05 ,HA = c('neq', 'greater' , 'smaller')){

     #Compute sample proportion
     p.hat <- y / n ;
     #Compute Standard Eror (SE)
     se <- sqrt(p0 * (1 - p0) / n)
     #Compute test statistic
     Z0 <- (p.hat - p0) / se
     #Compute critical value
     critical.z <- qnorm(1 - alpha/2)
     HO <- ifelse(abs(Z0) > critical.z , 'Reject' ,'Do not reject')
     HA <-match.arg(HA)
     if (HA == 'smaller') {
          critical.z <- qnorm(alpha)
          HO <- ifelse(Z0 < critical.z , 'Reject' ,'Do not reject')
     }
     if (HA == 'greater') {
          critical.z <- qnorm(1 - alpha)
          HO <- ifelse(Z0 > critical.z , 'Reject' ,'Do not reject')
     }
     results <- list(Z0 , critical.z , alpha , HO)
     names(results) <- c('Z' , 'critical.z' , 'alpha' , 'HO')
```

```
      class(results) <- 'table'
      print(results)
}
```

8. Ký hiệu $X \sim N(\mu, \sigma^2)$ chỉ X là biến ngẫu nhiên có phân phối chuẩn với kỳ vọng μ và phương sai σ^2; $Y \sim B(n, p)$ chỉ Y là biến ngẫu nhiên có phân phối nhị thức với tham số n và p (xác suất thành công). Hãy tính p – giá trị trong các trường hợp sau và cho biết chúng có ý nghĩa hay không?

1) $X = 1.96, X \sim N(0,1)$, kiểm định hai phía.
2) $X = 1.96, X \sim N(0,1)$, kiểm định một phía bên trái.
3) $X = 1.96, X \sim N(0,1)$, kiểm định một phía bên phải.
4) $X = 1.7, X \sim N(0,1)$, kiểm định hai phía.
5) $X = 1.7, X \sim N(0,1)$, kiểm định một phía bên trái.
6) $X = 1.7, X \sim N(0,1)$, kiểm định một phía bên phải.
7) $Y = 18, Y \sim B(50, 0.5)$, kiểm định hai phía.
8) $Y = 18, Y \sim B(50, 0.5)$, kiểm định một phía bên trái.
9) $Y = 18, Y \sim B(50, 0.5)$, kiểm định một phía bên phải.