Tuần 9: LT Bài 8.pdf

- Điểm danh trên gg classroom của lớp.
- Mở R, thiết lập thư mục làm việc cho R.
- Mở file “Bai 8.pdf”: đọc và thực hành các bài tập.

Thực hành xác suất thống kê

Bài 8:KIỂM ĐỊNH GIẢ THUYẾT THỐNG KÊ

1. Giới thiệu

Bài toán: 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

$\begin{cases} \boldsymbol{H_0: \theta = \theta_0} \\ \boldsymbol{H_1: \theta \neq \theta_0} \end{cases}$	$\begin{cases} \boldsymbol{H_0: \theta \geq \theta_0} \\ \boldsymbol{H_1: \theta < \theta_0} \end{cases}$	$\begin{cases} \boldsymbol{H_0: \theta \leq \theta_0} \\ \boldsymbol{H_1: \theta > \theta_0} \end{cases}$
Hai phía (‘two.sided’)	Một phía, bên trái (‘less’)	Một phía, bên phải (‘geater’)

2. Kiểm định giả thuyết cho kỳ vọng

Bài toán: Xét biến ngẫu nhiên $X_1, \dots, X_n$ 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 chưa biết.

$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$	$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$
Hai phía (‘two.sided’)	Một phía, bên trái (‘less’)	Một phía, bên phải (‘greater’)

2. Kiểm định giả thuyết cho kỳ vọng

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 $T_0 > t_{1-\alpha}^{n-1}$

Trong **R**

$t_{1-\frac{\alpha}{2}}^{n-1}$ = `qt(1 – alpha/2, n - 1)`

2. Kiểm định giả thuyết cho kỳ vọng

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**

$P(T_{n-1} \leq t_0) =$ `pt(t0, n - 1)`

$P(T_{n-1} \geq t_0) =$ `pt(t0, n – 1,lower.tail = FALSE)`

2. Kiểm định giả thuyết cho kỳ vọng

Sử dụng hàm t.test để kiểm định

```
t.test(x, alternative = “đối thuyết”, mu = mu0, conf.level = 0.95)
```

Trong đó

`x`: là véc-tơ dữ liệu.

`alternative`: là đối thuyết gồm hai phía (“`two.sided`”), bên trái (“`less`”), bên phải (“`greater`”), mặc định là “`two.sided`”.

`mu0`: giá trị cần kiểm định.

`conf.level`: Độ tin cậy (= 1 - α), mặc định là 0.95

2. Kiểm định giả thuyết cho kỳ vọng

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)
hist(heights)
t.test(heights,mu = 160, conf.level=0.95)
```

2. Kiểm định giả thuyết cho kỳ vọng

```
load('heights.rda')
summary(heights)
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
```

2. Kiểm định giả thuyết cho kỳ vọng

Giá trị thống kê kiểm định $T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} =$ `t = 2.6175`

Bậc tự do = `df = 124`

p - giá trị = `p-value = 0.009959`

Đối thuyết: $H_1: \mu \neq 160$ `alternative hypothesis: true mean is not equal to 160`

KTC 95% cho trung bình: `[160.6941 ; 164.9990]`

Trung bình mẫu: `mean of x  162.8465`

2. Kiểm định giả thuyết cho kỳ vọng

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 $T_0 > t_{1-\alpha}^{n-1}$

Trong **R**

$t_{1-\frac{\alpha}{2}}^{n-1}$ = `qt(1 – alpha/2, n - 1)`

2. Kiểm định giả thuyết cho kỳ vọng

Sử dụng giá trị thống kê kiểm định $T_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ = `t = 2.6175`

Ta có: $t^{n-1}_{1-\frac{\alpha}{2}} = t^{124}_{0.975} \approx 1.97928$ (= `qt(0.975,124)`)

Như vậy: t = 2.6175 > 1.97928 => Bác bỏ H_0.

Sử dụng p – giá trị: p - giá trị = `0.009959`

Ta có p - giá trị = 0.009959 < 0.05 = alpha => Bác bỏ H_0.

Với mức ý nghĩa 5%, thì chiều cao trung bình của thanh niên trong khu vực khác 160 cm.

Bài tập

Bài 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.

```
#Bai 1.
# Dua du lieu vao R:
data <- read.csv("profit.csv")
x    <- data$profit
# Cau a: Ve do thi cua du lieu
hist(x)
# Nhan xet: Du lieu da cho co dang dieu xap xi phan phoi chuan
```

Bài tập

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

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 thiết doanh số bán của những ngày bán đắt hàng là đại lượng ngẫu nhiên phân phối theo quy luật chuẩn).

Bài tập

```
# Cau b: Uoc luong khoang cho ky vong voi alpha = 1% cho mau co pp
chuan:
#Loc ra nhung ngay co doanh so ban hang lon hon 65tr
hi.pro <- x[x > 65]
#Tinh trung binh mau, do lech chuan mau
x.bar <- mean(hi.pro)
s <- sd(hi.pro)
n <- length(hi.pro)
alpha <- 0.01
epsilon = qnorm(1-alpha/2)*s/sqrt(n)
mu.lower = x.bar - epsilon
mu.upper = x.bar + epsilon
cat('Khoang tin cay',100*(1-alpha),'% cho ky vong mu la:\n')
cat('[',mu.lower,';',mu.upper,']\n')
```

Bài tập

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

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%.

Bài tập

```
# Cau c: Su dung ham t.test de kiem dinh gia thuyet cho ky vong voi
muc y nghia 1%.
# Dat gia thuyet:
# H0: "mu <= 60"
# H1: "mu > 60"

t.test(x, alternative = "greater", mu = 60, conf.level = 0.99)

# Nhan xet: Ket qua cho ta p-value = 0.9699 > 0.01 = 1% = alpha
# do do chua du co so de bac bo gia thuyet H0: "mu <= 60" voi muc y
nghia 1%.

# Ket luan: Voi muc y nghia 1%, phuong thuc ban hang moi
# khong mang lai hieu qua hon phuong thuc ban hang truoc do.
```

Bài tập

Bài 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.

Bài tập

```
#Bai 2:

# Cau a: Bien doi du lieu ve dang vecto
xi <- c(5,6,7,8,9,10)
ni <- c(5,10,15,20,12,8)
x <- rep(xi,ni); x

# Ve bieu do stem & leaf cho du lieu x:
stem(x)
```

Bài tập

Bài 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%.

b) Viết hàm test.leq.oneside(x, μ_0,alpha) để kiểm định giả thiết H_0: $\mu = \mu_0$ và đối thiết H_1: $\mu < \mu_0$. Xuất ra thông báo và p - giá trị. Áp dụng để kiểm định H_1: $\mu < 8$.

```
# Cau b:
# Viet ham test.leq.oneside:

test.leq.oneside <- function(x, mu_0,alpha){
        x.bar = mean(x)
        s = sd(x)
        n = length(x)
        t_0 = (x.bar - mu_0)*sqrt(n)/s
        p.value = pt(t_0,n-1)
        cat('Voi muc y nghia alpha =',alpha, ':\n')
        if(p.value <= alpha)
        cat('Bac bo H0 voi p-value =',p.value)
        else
        cat('Chua du co so de bac bo H0 voi p-value =',p.value)}

# Ap dung

test.leq.oneside(x,8,0.05)

# Kiem tra lai bang ham t.test:
t.test(x, alternative = "less", mu = 8, conf.level = 0.95)

# Nhan xet: Ket qua cho ta p-value = 0.0332 < 5% = alpha
# do do bac bo gia thuyet H0:"mu > 8" voi muc y nghia 5%.

# Ket luan: Voi muc y nghia 5%, diem trung binh cua hoc vien khong lon hon
8.
```

Bài tập

Bài 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%.

c) Viết hàm test.geq.oneside(x, μ_0, alpha) để kiểm định giả thiết H_0: $\mu = \mu_0$ và đối thiết H_1: $\mu > \mu_0$. Xuất ra thông báo và p - giá trị. Áp dụng để kiểm định H_1: $\mu > 8$.

```r
# Cau c:
# Viet ham test.geq.oneside:

test.geq.oneside <- function(x, mu_0,alpha){
        x.bar = mean(x)
        s = sd(x)
        n = length(x)
        # tinh gia tri thong ke kiem dinh t0:
        t_0 = (x.bar - mu_0)*sqrt(n)/s
        # tinh p - value
        p.value = 1- pt(t_0,n-1)
        cat('Voi muc y nghia alpha =',alpha, ':\n')
        if(p.value <= alpha)
        cat('Bac bo H0 voi p-value =',p.value)
        else
        cat('Chua du co so de bac bo H0 voi p-value =',p.value)}

# Ap dung:

test.geq.oneside(x,8,0.05)

# Kiem tra lai bang ham t.test:
t.test(x, alternative = "greater", mu = 8, conf.level = 0.95)

# Nhan xet: Ket qua cho ta p-value = 0.9668 > 5% = alpha
# do do chua du co so de bac bo gia thuyet H0:"<=8" voi muc y nghia 5%.
# Ket luan: Voi muc y nghia 5%, diem trung binh cua hoc vien khong lon hon 8
```

3. Kiểm định giả thuyết cho tỷ lệ

Bài toán: Cần kiểm định tỷ lệ thỏa tính chất A trong tổng thể. Khảo sát một mẫu cỡ n gồm n biến ngẫu nhiên $Y_1, \ldots, Y_n$ 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 thỏa tính chất A trong n phần tử khảo sát, suy ra tỷ lệ mẫu $\hat{p} = \frac{y}{n}$.

$\begin{cases} \boldsymbol{H_0}: \mathbf{p} = \mathbf{p_0} \\ \boldsymbol{H_1}: \mathbf{p} \neq \boldsymbol{p_0} \end{cases}$	$\begin{cases} \boldsymbol{H_0}: \mathbf{p} = \mathbf{p_0} \\ \boldsymbol{H_1}: \mathbf{p} < \boldsymbol{p_0} \end{cases}$	$\begin{cases} \boldsymbol{H_0}: \mathbf{p} = \mathbf{p_0} \\ \boldsymbol{H_1}: \mathbf{p} > \boldsymbol{p_0} \end{cases}$
Hai phía (‘two.sided’)	Một phía, bên trái (‘less’)	Một phía, bên phải (‘geater’)

3. Kiểm định giả thuyết cho tỷ lệ

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 $Z_0 > z_{1-\alpha}$

Trong **R**

$z_{1-\frac{\alpha}{2}}$ = `qnorm(1 - alpha/2)`

3. Kiểm định giả thuyết cho tỷ lệ

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 hóa N(0;1).

Trong **R**

$$\Phi(Z_0) = \texttt{pnorm(Z0)}$$

3. Kiểm định giả thuyết cho tỷ lệ

Sử dụng hàm prop.test để kiểm định

```
prop.test(y,n, p = p0, alternative = “đối thuyết”, conf.level = 0.95)
```

Trong đó

`y`: là số phần tử thỏa tính chất A trong n phần tử khảo sát.

`n`: là cỡ mẫu.

`alternative`: là đối thuyết gồm hai phía (“`two.sided`”), bên trái (“`less`”), bên phải (“`greater`”), mặc định là “`two.sided`”.

`p0`: giá trị cần kiểm định.

`conf.level`: Độ tin cậy (= 1 - α), mặc định là 0.95

3. Kiểm định giả thuyết cho tỷ lệ

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 cho ý kiến ủng hộ ông A. Hãy xét xem tuyên bố của ông A về tỷ lệ ủng hộ từ cử tri có đúng không với mức ý nghĩa 1%.

Ta có:

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

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

3. Kiểm định giả thuyết cho tỷ lệ

Sử dụng 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 = 11.281, 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
```

3. Kiểm định giả thuyết cho tỷ lệ

Giá trị cần kiểm định: $p_0 = 0.5$

Giá trị thống kê kiểm định $Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$ = `11.281`

Bậc tự do = `df = 1`

p - giá trị = `p-value = 0.0003915`

Đối thuyết: $H_1: p > 0.5$ `alternative hypothesis: true p is greater than 0.5`

KTC 99% cho trung bình: `[0.5182781 ; 1.0000000]`

Tỉ lệ mẫu: `sample estimates: p = 0.56`

3. Kiểm định giả thuyết cho tỷ lệ

p - giá trị = 0.0003915 < 1% = alpha

=> Bác bỏ $H_0: p = 0.5$

Kết luận: Với mức ý nghĩa 1%, tỷ lệ người dân ủng hộ ông A trong thành phố trên 50%.

3. Kiểm định giả thuyết cho tỷ lệ

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

```
n = 800; y = 448
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
```

Bài tập

Bài 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.

Bài tập

```
# Bai 6
# Nhap du lieu vao R
dat <- read.csv("times.csv", header = T)
attach(dat); names(dat)


# a) Su dung prop.test de kiem dinh
n = length(KHTN)
y = length(KHTN[KHTN > 5])
n;y
prop.test(y,n,0.5,alternative = "two.sided", conf.level =
0.95)
```