**THỰC HÀNH BÀI TẬP XSTK TRÊN R**

**Bài 1**. Cho mẫu thực nghiệm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | 20 | 22 | 24 | 26 | 28 |
|  | 5 | 6 | 7 | 6 | 4 |

Tính và s

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(20,22,24,26,28), times=c(5,6,7,6,4))

1. Tính các đặc trưng của mẫu theo cú pháp  
   > mean(x)

[1] 23.85714

> var(x)

[1] 7.089947

> sd(x)

[1] 2.662695

**Bài 2.** Cho mẫu thực nghiệm

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 41 | 43 | 45 | 47 | 49 | 51 | 53 |
|  | 10 | 25 | 33 | 55 | 36 | 21 | 12 |

Tính và s

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(41,43,45,47,49,51,53), times=c(10,25,33,55,36,21,12))

1. Tính các đặc trưng của mẫu theo cú pháp  
   > mean(x)

[1] 47.01042

> var(x)

[1] 9.444917

> sd(x)

[1] 3.073258

**Bài 3.** Để kiểm tra mức xăng hao phí của một loại xe ô tô. Người ta chọn ngẫu nhiên 28 chiếc xe và cho chạy trên cùng một đoạn đường 300 km. Kết quả thu được như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *X* | 4*,* 6 *–* 4*,* 8 | 4*,* 8 *–* 5*,* 0 | 5*,* 0 *–* 5*,* 2 | 5*,* 2 *–* 5*,* 4 | 5*,* 4 *–* 5*,* 6 |
|  | 6 | 5 | 9 | 4 | 4 |

Với độ tin cậy *γ* = 0*,* 95, hãy tìm khoảng ước lượng của lượng xăng hao phí trung bình. Biết mức xăng hao phí có phân phối chuẩn.

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(4.7, 4.9, 5.1, 5.3, 5.5), times=c(6,5,9,4,4))

1. Tính các đặc trưng của mẫu theo cú pháp  
   > mean(x)

[1] 5.064286

> var(x)

[1] 0.07126984

> sd(x)

[1] 0.2669641

3. Công thức khoảng ước lượng cho kỳ vọng là

4. Định nghĩa biến ghi phân vị theo hàm qt

> n<-length(x)

> phanvi <- qt(0.975,n-1)

Đọc giá trị phân vị

> phanvi

[1] 2.051831

5. Định nghĩa biến ghi bán kính ước lượng và đọc giá trị của bán kính

> bkinh <- phanvi\*sd(x)/sqrt(n)

> bkinh

[1] 0.1035179

6. Tính khoảng ước lượng với độ tin cậy 95%

> mean(x)+c(-1,1)\*bkinh

[1] **4.960768** **5.167804**

**Bài 4**. Giả sử rằng thu nhập *X* hàng tháng của một kỹ sư sau khi ra trường 5 năm là một biến ngẫu nhiên có phân phối chuẩn . Lấy một mẫu quan sát của *X* có kích thước *n* = 64 ta có kết quả:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *X* (triệu) | 4 *-* 6 | 6 *-* 8 | 8 *-* 10 | 10 *-* 12 | 12 *-* 14 |
| (người) | 12 | 24 | 18 | 7 | 3 |

Với độ tin cậy *γ* = 0*,* 99 hãy xây dựng khoảng tin cậy cho thu nhập trung bình hàng tháng của kỹ sư.

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(5, 7, 9, 11, 13), times=c(12,24,18,7,3))

1. Tính các đặc trưng của mẫu theo cú pháp  
   > mean(x)

[1] 7.90625

> var(x)

[1] 4.5625

> sd(x)

[1] 2.136001

3. Công thức khoảng ước lượng cho kỳ vọng là

4. Định nghĩa kích thước mẫu và biến ghi phân vị theo hàm qnorm

> n<-length(x)

> phanvi <- qnorm (0.995)

Đọc giá trị phân vị

> phanvi

[1] 2.575829

5. Định nghĩa biến ghi bán kính ước lượng và đọc giá trị của bán kính

> bkinh <- phanvi\*sd(x)/sqrt(n)

> bkinh

[1] 0.6877467

6. Tính khoảng ước lượng với độ tin cậy 95%

> mean(x)+c(-1,1)\*bkinh

[1] **7.218503 8.593997**

**Chú ý**. Trong R ta vẫn dùng được lệnh phanvi <-qt(0.995,n-1) (do R không giới hạn số bậc tự do n<=30) và nhận được [1] 2.656145.

**Bài 5**. Dung lượng (đơn vị Ampe-giờ) của 10 chiếc pin được ghi lại dưới đây  
 140*;* 136*;* 150*;* 144*;* 148*;* 152*;* 138*;* 141*;* 143*;* 151  
a) Ước lượng giá trị phương sai tập chính .  
b) Tìm khoảng tin cậy 98% cho .

**Thực hành trên R**

1. Nhập dữ liệu

> dluong <-c(140*,* 136*,* 150*,* 144*,* 148*,* 152*,* 138*,* 141*,* 143*,* 151)

1. Ước lượng giá trị phương sai tập chính

> var(dluong)

[1] 32.23333

1. Định nghĩa các biến ghi kích thước, 2 phân vị bên trái, bên phải và đọc giá trị từ R

> n<-length(dluong)

> n

[1] 10

> pvtrai<-qchisq(0.01,9)

> pvtrai

[1] 2.087901

> pvphai<-qchisq(0.99,9)

> pvphai

[1] 21.66599

1. Công thức khoảng ước lượng là
2. Tính toán khoảng ước lượng trong R

> c((n-1)\*var(dluong)/pvphai, (n-1)\*var(dluong)/pvtrai)

[1] **13.38965 138.94339**

**Bài 6**. Kiểm tra ngẫu nhiên 400 người đi xe máy ở khu vực có 500*.*000 người đi xe máy thấy có 360 người có bằng lái. Với độ tin cậy 0*,*95 hãy ước lượng số người đi xe máy có bằng lái trong khu vực.

**Thực hành trên R**

1. Nhập dữ liệu cho R  
   > N <-500000

> n <-400

> m <-360

2. Định nghĩa tần số, phân vị và đọc các giá trị này từ R  
 > f<-m/n

> f

[1] 0.9

> phanvi<-qnorm(0.975)

> phanvi

[1] 1.959964

3. Công thức khoảng ước lượng cho tỷ lệ p (tỷ lệ có bằng lái) là

4. Tính toán khoảng ước lượng cho p trong R

> c(f-phanvi\*sqrt(f\*(1-f)/n), f+phanvi\*sqrt(f\*(1-f)/n))

[1] **0.8706005 0.9293995**

5. Tính toán khoảng ước lượng cho M (số lượng người có bằng lái) trong R

> c(N\*f-N\*phanvi\*sqrt(f\*(1-f)/n), N\*f+N\*phanvi\*sqrt(f\*(1-f)/n))

[1] **435300.3 464699.7**

**Chú ý:** Có thể tính khoảng ước lượng bằng cách sử dụng biến bkinh (tính bán kính của khoảng).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| (gram) | 29*,*5 *–* 29*,*7 | 29*,*7 *–* 29*,*9 | 29*,*9 *–* 30*,*1 | 30*,*1 *–* 30*,*3 | 30*,*3 *–* 30*,*5 |
| số SP | 7 | 9 | 12 | 5 | 3 |

**Bài 7**. Hao phí nguyên liệu cho 1 đơn vị sản phẩm là đại lượng ngẫu nhiên có phân phối chuẩn với độ lệch chuẩn *σ* = 0*,*03 và hao phí trung bình là 29*,* 9 gram. Nghi ngờ máy móc  
trục trặc làm cho hao phí nguyên liệu cho 1 đơn vị sản phẩm tăng lên. Người ta kiểm tra 36 sản phẩm thì có bảng số liệu sau:

Hãy kết luận ý kiến nêu trên với mức ý nghĩa *α* = 0*,*05.

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(29.6, 29.8, 30, 30.2, 30.4), times=c(7,9,12,5,3))

1. Tính trung bình của mẫu theo cú pháp  
   > mean(x)

[1] 29.93333

3. Đưa ra kích thước của mẫu và đọc dữ liệu từ R

> n<- length(x)

> n

[1] **36**

4. Xác định cặp giả thuyết, đối thuyết của bài toán

5. Tiêu chuẩn kiểm định

6. Nhập giá trị

> sigma <-0.03

7. Công thức miền bác bỏ và xác định giá trị phân vị

> phanvi <-qnorm(0.95)

> phanvi

[1] 1.644854

8. Nhập công thức tính và đọc giá trị từ R  
 > Zqs <- (mean(x)-29.9)\*sqrt(n)/sigma

> Zqs

[1] 6.666667

9. Đưa ra kết luận

nên : Ta bác bỏ và thay thế bằng .

**Chú ý**: Để thực hiện ***kiểm định p-value*** ta sử dụng các cú pháp

> library(BSDA)

> z.test(x, alternative="greater", mu=29.9, sigma.x=0.03, conf.level=0.95)

Kết quả được R đưa ra như sau:

One-sample z-Test

data: x

z = 6.6667, p-value = 1.308e-11

alternative hypothesis: true mean is greater than 29.9

95 percent confidence interval:

29.92511 NA

sample estimates:

mean of x

29.93333

***Giải thích kết quả:*z = 6.6667** là giá trị của

**p-value = 1.308e-11** nghĩa là ứng với mẫu thực nghiệm này của X, giả thuyết gốc bị bác bỏ với tất cả các mức ý nghĩa lớn hơn hoặc bằng giá trị .

**Bài 8.** Trọng lượng những bao phân đạm do nhà máy sản xuất ra là 1 biến ngẫu nhiên *X* tuân theo luật phân bố chuẩn, với *EX* = 50 kg. Khách hàng cho rằng trọng lượng đó đã thay đổi và ít hơn 50 kg. Cân thử 100 bao, ta có bảng số liệu dưới đây:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *xi* | 47 *-* 48 | 48 *-* 49 | 49 *-* 50 | 50 *-* 51 | 51 *-* 52 |
| *ni* | 30 | 40 | 20 | 5 | 5 |

Với mức ý nghĩa *α* = 0*,*05, hãy cho kết luận về ý kiến của khách hàng.

**Thực hành trên R**

1. Nhập dữ liệu có tần số theo cú pháp

x <- rep(c(47.5, 48.5, 49.5, 50.5, 51.5), times=c(30,40,20,5,5))

1. Tính trung bình, phương sai và độ lệch chuẩn của mẫu theo cú pháp  
   > mean(x)

[1] 48.65

> var(x)

[1] 1.138889

> sd(x)

[1] 1.067187

3. Đưa ra kích thước của mẫu và đọc dữ liệu từ R

> n<- length(x)

> n

[1] **100**

4. Xác định cặp giả thuyết, đối thuyết của bài toán

5. Tiêu chuẩn kiểm định

6. Công thức miền bác bỏ và xác định giá trị phân vị

> phanvi <-qt(0.95,99)

> phanvi

[1] 1.660391

**Chú ý:** *R không giới hạn số bậc tự do nên ta có thể tính các phân vị của biến student với số bậc tự do . Nếu sử dụng bảng giấy (kém chính xác hơn) thì giá trị phân vị được lấy gần đúng là (hoặc là sử dụng máy Casio để thu được giá trị 1,644853667) và ta chấp nhận sai số khi chuyển đổi theo kiểu này ở mức độ .*

7. Nhập công thức tính và đọc giá trị từ R  
 > Tqs <- (mean(x)-50)\*sqrt(n)/sd(x)

> Tqs

[1] -12.65007

8. Đưa ra kết luận

nên : Ta bác bỏ và thay thế bằng .

**Chú ý**: Để thực hiện ***kiểm định p-value*** ta sử dụng cú pháp

> t.test(x, alternative="less", mu=50, conf.level=0.95)

Kết quả được R đưa ra như sau:

One Sample t-test

data: x

t = -12.65, df = 99, p-value < 2.2e-16

alternative hypothesis: true mean is less than 50

95 percent confidence interval:

-Inf 48.82719

sample estimates:

mean of x

48.65

Với kết quả này **p-value < 2.2e-16** cónghĩa là ứng với mẫu thực nghiệm này của X, giả thuyết gốc bị bác bỏ với tất cả các mức ý nghĩa lớn hơn hoặc bằng giá trị p mà ở đây

**Bài 9**. Một tòa báo thanh niên thông báo có 35% học sinh phổ thông trung học là độc giả thường xuyên. Phỏng vấn 400 em thì có 136 em đọc báo đó thường xuyên. Hãy kiểm định tính chính xác của thông báo trên với mức ý nghĩa 0*,*05.

**Thực hành trên R**

1. Nhập dữ liệu của bài toán

> n <- 400

> m <-136

> p0 <-0.35

1. Kiểm tra điều kiện về kích thước mẫu

> n\*p0

[1] 140

> n\*(1-p0)

[1] 260

3. Xác định cặp giả thuyết, đối thuyết của bài toán

4. Tiêu chuẩn kiểm định

5. Công thức miền bác bỏ và xác định giá trị phân vị

> phanvi <-qnorm(0.975)

> phanvi

[1] 1.959964

6. Nhập công thức tính và đọc giá trị từ R  
 > Tqs <- (m/n-0.35)\*sqrt(n)/sqrt(0.35\*(1-0.35))

> Tqs

[1] -0.4193139

7. Đưa ra kết luận

nên : Ta chưa có cơ sở để bác bỏ .

**Chú ý**: Để thực hiện ***kiểm định p-value*** ta sử dụng cú pháp

> prop.test(136, 400, p=0.35, alternative="two.sided", conf.level=0.95, correct=FALSE)

Kết quả được R đưa ra như sau:

1-sample proportions test without continuity correction

data: 136 out of 400, null probability 0.35

X-squared = 0.17582, df = 1, p-value = 0.675

alternative hypothesis: true p is not equal to 0.35

95 percent confidence interval:

0.2952956 0.3877483

sample estimates:

p

0.34

Với kết quả này **p-value = 0.675** cónghĩa là ứng với mẫu thực nghiệm này, giả thuyết gốc bị bác bỏ với tất cả các mức ý nghĩa lớn hơn hoặc bằng giá trị p = 0,675. R không đưa ra giá trị -0.4193139 mà đưa ra giá trị X-squared = 0.17582, ở đây X-squared = .

**Bài 10**. Một xí nghiệp sử dụng hai dây chuyền tự động khác nhau để sản xuất. Sau khi theo dõi người ta thu được kết quả sau

|  |  |  |  |
| --- | --- | --- | --- |
| Dây chuyền | Số ca  theo dõi | Số sản phẩm  trung bình | Độ lệch tiêu chuẩn thực nghiệm |
| *I  II* | 100  150 | 306  375 | 10 25 |

Với mức ý nghĩa *α* = 0*,*05, có thể kết luận dây chuyền thứ hai có sản lượng tốt hơn dây chuyền thứ nhất hay không? Giả thiết sản lượng tuân theo phân phối chuẩn.

**Thực hành trên R**

1. Nhập dữ liệu về các biến X,Y của bài toán  
 > n1<-100

> xtb<-306

> Sx<-10

> n2<-150

> ytb<-375

> Sy<-25

2. Xác định cặp giả thuyết, đối thuyết của bài toán

3. Tiêu chuẩn kiểm định

trong đó df (số bậc tự do) được cho theo công thức Welch

4. Nhập số bậc tự do và đọc giá trị từ R

> df=(Sx^2/n1+Sy^2/n2)^2/( ( Sx^4/ (n1^2\*(n1-1)) )+( Sy^4/(n2^2\*(n2-1)) ) )

> df

[1] 210.8257

5. Công thức miền bác bỏ và xác định giá trị phân vị

> phanvi <-qt(0.95,df)

> phanvi

[1] 1.652113

**Chú ý 1.** *Việc sử dụng bảng giấy* tương ứng với việc lấy xấp xỉ phân vị là và ta chấp nhận sai lệch .

6. Nhập công thức tính và đọc giá trị từ R  
 > Tqs <- (xtb-ytb) /sqrt( Sx^2/n1 +Sy^2/n2 )

> Tqs

[1] -30.35595

7. Đưa ra kết luận

nên : Ta bác bỏ và thay thế bằng .

**Chú ý 2.** Ta có thể thực hiện bằng ***kiểm định p-value*** cú pháp

> library(BSDA)

> tsum.test(mean.x=306, s.x = 10, n.x = 100, mean.y=375, s.y=25, n.y=150, alternative="less",mu=0,var.equal=FALSE)

Kết quả được R đưa ra như sau:

Welch Modified Two-Sample t-Test

data: Summarized x and y

t = -30.356, df = 210.83, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

NA -65.2447

sample estimates:

mean of x mean of y

306 375

Với kết quả này **p-value < 2.2e-16** cónghĩa là ứng với mẫu thực nghiệm này của X, giả thuyết gốc bị bác bỏ với tất cả các mức ý nghĩa lớn hơn hoặc bằng giá trị p mà ở đây

**Chú ý 3.** Ta có thể thực hiện ***kiểm định p-value*** bằng hàm rnorm. Cụ thể hơn với các cú pháp

> x <- rnorm(100, mean=306, sd=10)

> y <- rnorm(150, mean=375, sd=25)

R sẽ tạo cho chúng ta 2 mẫu dữ liệu ngẫu nhiên. Mẫu thứ nhất ứng với biến x, kích thước 100, trung bình lý thuyết là 306, độ lệch lý thuyết là 10. Mẫu thứ hai ứng với biến y, kích thước 150, trung bình lý thuyết là 375, độ lệch lý thuyết là 25. *Các giá trị mean và sd trên x, y sẽ có sai khác (ngẫu nhiên) với giá trị được nhập vào*.

Việc kiểm định sẽ được thực hiện qua cú pháp  
> t.test(x, y, alternative="less", mu=0, var.equal=FALSE, paired=FALSE)

Kết quả do R trả ra màn hình có thể khác nhau (trên 2 máy tính khác nhau hoặc giữa 2 lần chạy khác nhau). Nguyên nhân là do x,y được tạo ngẫu nhiên và khác nhau giữa các lần. Chẳng hạn dưới đây là kết quả của 2 lần khác nhau trên cùng một máy tính:  
  
**Lần 1**

> x <- rnorm(100, mean=306, sd=10)

> y <- rnorm(150, mean=375, sd=25)

> t.test(x, y, alternative="less", mu=0, var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test

data: x and y

t = -29.12, df = 207.73, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -63.23793

sample estimates:

mean of x mean of y

306.5781 373.6198

**Lần 2**

> x <- rnorm(100, mean=306, sd=10)

> y <- rnorm(150, mean=375, sd=25)

> t.test(x, y, alternative="less", mu=0, var.equal=FALSE, paired=FALSE)

Welch Two Sample t-test

data: x and y

t = -31.718, df = 200.71, p-value < 2.2e-16

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -63.85919

sample estimates:

mean of x mean of y

307.5025 374.8715

**Bài 11.** Một kỹ sư đưa ra một quy trình sản xuất mới để giảm tỷ lệ phế phẩm của nhà máy. Kiểm tra về hai quy trình sản xuất ta có bảng số liệu sau

|  |  |  |
| --- | --- | --- |
| Quy trình | Số sản phẩm kiểm tra | Số phế phẩm |
| Cũ |  |  |
| Mới |  |  |

Với mức ý nghĩa 0,05 hãy đánh giá ý kiến cửa kỹ sư trên.

**Thực hành trên R**

1. Nhập dữ liệu của bài toán  
 > n1<-250

> n2<-350

> m1<-18

> m2<-35

2. Xác định cặp giả thuyết, đối thuyết của bài toán

Ở đây là tỷ lệ phế phẩm của quy trình sản xuất cũ, là tỷ lệ phế phẩm của quy trình sản xuất mới.

3. Định nghĩa các biến mô tả tần số và đọc giá trị từ R

> f1<-m1/n1

> f2<-m2/n2

> fchung<-(m1+m2)/(n1+n2)

> f1

[1] 0.072

> f2

[1] 0.1

> fchung

[1] 0.08833333

4. Kiểm tra điều kiện về kích thước mẫu (so sánh với 10)

> n1\*f1

[1] 18

> n1\*(1-f1)

[1] 232

> n2\*f2

[1] 35

> n2\*(1-f2)

[1] 315

5. Tiêu chuẩn kiểm định

5. Công thức miền bác bỏ và xác định giá trị phân vị

> phanvi <-qnorm(0.95)

> phanvi

[1] 1.644854

6. Nhập công thức tính và đọc giá trị từ R  
 > Tqs <- (f1-f2) /sqrt(fchung\*(1-fchung)\*(1/n1+1/n2))

> Tqs

[1] -1.191533

7. Đưa ra kết luận

nên : Ta chưa có cơ sở để bác bỏ .

**Chú ý**: Để thực hiện ***kiểm định p-value*** ta sử dụng cú pháp

> prop.test(c(18,35), c(250,350), alternative="greater", conf.level=0.95, correct=FALSE)

Kết quả được R đưa ra màn hình là

2-sample test for equality of proportions without continuity correction

data: c(18, 35) out of c(250, 350)

X-squared = 1.4198, df = 1, p-value = 0.8833

alternative hypothesis: greater

95 percent confidence interval:

-0.06566704 1.00000000

sample estimates:

prop 1 prop 2

0.072 0.100

Kết quả trên không đưa ra giá trị mà đưa ra bình phương của nó X-squared = 1.4198 = . Giá trị **p-value = 0.8833** là mức ý nghĩa nhỏ nhất ứng với việc bác bỏ .