ĐÁP ÁN THAM KHẢO ĐỀ THI THỬ 1

Câu 1. Trong các đoạn code R sau, đoạn nào trả ra kết quả là TRUE .

A. (2!= 2) && (3 <= 3) **# FALSE vì 2 != 2 (2≠2) => FALSE; 3 <= 3 => TRUE; FALSE && TRUE = FALSE. (&& là phép giao)**

B. (3 != 4) || (4 >= 3) **#TRUE vì 3 != 4 => FALSE; 4 >= 3 => TRUE; FALSE || TRUE = TRUE (| là OR ; || là phép hợp).**

C. (2 <= 5) && (4 <= 3) **# FALSE vì 2 <= 5 => FALSE; 4 <= 3 => FALSE; FALSE && FALSE = FALSE.**

D. isTRUE(4 == 5) **#FALSE ; hàm isTRUE(x): hàm xét tính đúng sai của biểu thức x, vì 4 ≠ 5 nên kết quả là FALSE**

Cho đoạn code sau, trả lời các câu hỏi 2-3

```
perm <- function(n){
   if(n ==1){
       return( matrix (1))
   } else{
       sp <- perm(n -1)
       p <- nrow(sp)
       A <- matrix( nrow =n*p, ncol =n)
      for (i in 1:n){
            A[(i -1)*p+1:p ,] <- cbind(i,sp +(sp >=i))
       }
       return(A)
    }
}
Z= perm (3);
```

Câu 2. Cho biết số dòng của ma trận Z

```
Z

     [,1] [,2] [,3]

[1,]    1    2    3

[2,]    1    3    2

[3,]    2    1    3

[4,]    2    3    1

[5,]    3    1    2

[6,]    3    2    1
```

```
# ---> so dong la 6
```

D. 6

Câu 3. Tìm số tự nhiên *n* sao cho ma trận *perm(n)* có 24 dòng
B. 4

Câu 4. Câu lệnh R nào dùng để mô phỏng việc tung đồng xu cân đối đồng chất 5 lần.
 # lưu ý lệnh `sample(x, size, replace = FALSE, prob = NULL)`
x là vectơ dữ liệu gồm 1 hoặc nhiều thành phần.
size là số lượng mẫu phát sinh (cỡ mẫu).
replace: FALSE: các giá trị trong x khác nhau; TRUE: các giá trị trong x có thể lập lại. Mặc định là FALSE.
prob là tham số chứa các xác suất tương ứng với các giá trị trong x, nếu để trống => mặc định xác suất các giá trị bằng nhau.

Do đồng xu cân đối đồng chất nên xác suất xuất hiện mỗi mặt là 1/2 và tung đồng xu 5 lần nhưng chỉ có 2 giá trị nên cần tham số replace = TRUE.

A. `sample(c('H','T'),5).` **#cỡ mẫu nhiều hơn giá trị nhưng lại thiếu replace = TRUE.**
B. `sample(c(0,1), 5, prob=c(1/2,1/2)).` **#cỡ mẫu nhiều hơn giá trị nhưng lại thiếu replace = TRUE**
C. `sample(c(0,1), 5, replace= T).`
D. `sample(5, c('H','T'), replace= T).` **# Sai cấu trúc**

Cho đoạn code sau, trả lời các câu hỏi 5-6

```
x <- c( rep(1, 3), rep(2, 4), rep(3 ,5) , rep(4 ,4) ,rep(5 ,3))
a = mean(x) # trung bình mẫu của x.
b = length(x) # độ dài của x (cỡ mẫu).
c = median(x) # trung vị của x.
```

 # Lưu ý: rep(x,n): lập lại giá trị x với n lần.

```
> x
 [1] 1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5
> a
[1] 3
> b
[1] 19
> c
[1] 3
```

(đề gõ sai `b*a` => `a*c`)
Câu 5. Kết quả của `a*b` bằng **57** và `a*c` bằng **9**

Câu 6. Giá trị của `a*b - c` bằng

A. 54. **B**. 24. **C**. 10. **D**. 60

Câu 7. Câu lệnh R: `3*length(sample(1:6,5))+10` cho ra kết quả nào sau đây ?

A. 25. **B**. 28. **C**. 43. **D**. 13.

Cho đoạn code sau, trả lời các câu hỏi 8-9

```
df <- data.frame(
STT = c( seq(1 ,6 ,1)) ,
Name = c(" Pearson ", " Neymann "," Fisher "," Gosset "," Bayes "," Poisson "),
Birday = c(1857 ,1894 ,1890 ,1876 ,1702 ,1781) ,
Died =c(1936 , 1981 , 1962 , 1937 , 1761 , 1840)
)
```

Câu 8. Dựa vào dataframe df, hãy cho biết năm sinh và năm mất của cha đẻ kiểm định t (Gosset, William Sealy Gosset).

A. 1890-1962. **B**. 1936-1937. **C**. 1876-1937. **D**. 1890-1962.

```
> df
  STT      Name Birday Died
1   1  Pearson    1857 1936
2   2  Neymann    1894 1981
3   3   Fisher    1890 1962
4   4   Gosset    1876 1937
5   5    Bayes    1702 1761
6   6  Poisson    1781 1840
```

Câu 9. Lệnh `subset(Name,Birday%%25>13)` cho ra tên của hai nhà thống kê đã có quan điểm đối lập nhau trong vấn đề kiểm định

A. [1] "Bayes" "Poisson". **B**. [1] "Pearson" "Neymann" .
C. [1] "Gosset" "Poisson". **D**. [1] "Neymann" "Fisher" .

```
> subset(df$Name,df$Birday%%25>13)
[1]  Neymann   Fisher
Levels:  Bayes   Fisher   Gosset   Neymann   Pearson   Poisson
```

Câu 10. Giá trị của `ppois(x0,lambda)` bằng với

A. Giá trị của hàm phân phối (tích lũy) của biến ngẫu nhiên phân phối Poisson P(λ) tại x_0.

B. $\sum_{k\in\mathbb{Z},0\le k\le x_0} \frac{e^{-\lambda}\lambda^k}{k!}$.

C. $\mathbb{P}(X \le x_0)$ trong đó $X \sim$P(λ) .

D. Tất cả các giá trị liệt kê ở trên.

Câu 11. Trong các lệnh sau, lệnh nào có thể vẽ được hình bên dưới.

A. `curve(dbinom(x, 9, 0.5), from = 0, to = 10).`

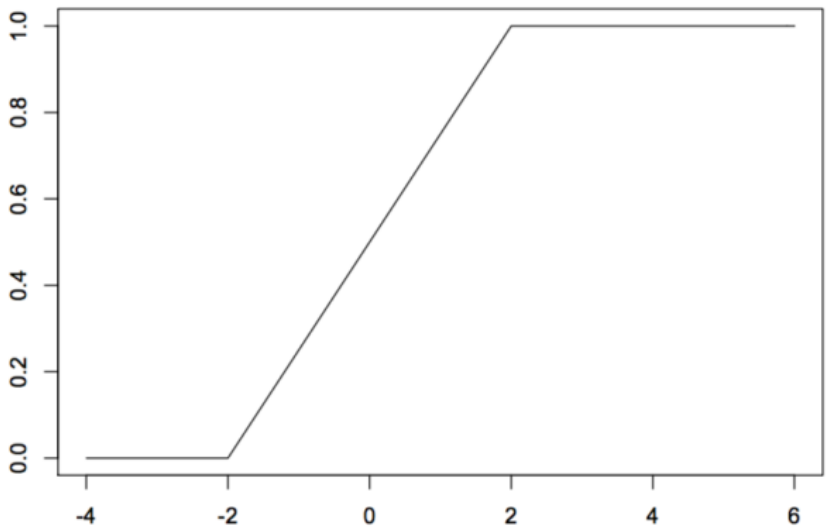
B. `curve(dnorm(x, 2, 1), from = -1, to = 5).`

C. `hist(c(0:9),dbinom(0:9,9, 0.5)).`

D. `plot(0:9,dbinom(0:9,9, 0.5), type='h', ylab = "P(X = k)").`

 # Quan sát biểu đồ, nhận diện đồ thị thống kê sau đó chọn câu lệnh phù hợp với loại đồ thị đó.

Câu 12. Cho biết đồ thị hàm phân phối (tích lũy), như Hình 2, là của phân phối nào ?



Hình 2: Hàm phân phối tích lũy

A. PP chuẩn N (1; 1).

B. PP đều U([−2; 2]).

C. PP mũ Exp(2) .

D. PP Student(10)

Câu 13. Để phát sinh hàm mật độ của biến ngẫu nhiên có phân phối chuẩn $\mathcal{N}(\mu = 2, \sigma^2 = 1$ và mẫu 1000 phần tử là biến ngẫu ngẫu nhiên có phân phối Poisson $\mathrm{P}(\lambda = 2)$ lần lượt bằng các lệnh `dnorm(x,2,1)` và `rpois(1000,2)`

Đề thi cuối kỳ môn THỰC HÀNH XÁC SUẤT THỐNG KÊ dạng trắc nghiệm có 50 câu hỏi, mỗi câu 5 đáp án. Sinh viên A không học bài, khi đi thi thì xác suất trả lời đúng của mỗi câu là như nhau. Sinh viên B học khá trong lớp, cảm thấy xác suất để mình chọn đúng mỗi câu là 0.6. Đặt

$$X = \sum_{i=1}^{50} X_i \text{ và } Y = \sum_{i=1}^{50} Y_i$$

trong đó, X_i, Y_i lần lượt là các biến ngẫu nhiên phản ánh kết quả chọn câu thứ i là đúng của sinh viên A và sinh viên B (với $i = 1, \ldots, 50$), trả lời các câu hỏi 13 và 14:

Câu 14. X và Y có phân phối gì ?

A. Nhị thức B(50, 0.2) và B(50, 0.4) .
B. Nhị thức B(50, 0.2) và B(50, 0.6).
C. Nhị thức B(50, 0.8) và B(50, 0.6) .
D. Nhị thức B(50, 0.8) và B(50, 0.4).

Câu 15. Sinh viên sẽ qua môn nếu trả lời đúng ít nhất 25 câu hỏi. Xác suất sinh viên A không qua môn; được xấp xỉ bằng câu lệnh nào sau đây ?

A. `pnorm(15/sqrt(8)).`
B. `pnorm(15/8).`
C. `1-pnorm(15/sqrt(8)).`
D. `dnorm(0.5).`

Cần tính $\mathbb{P}(X < 25)$

 # Lưu ý 1: khi n đủ lớn, với X ~ B(n,p) => x ~ $\mathcal{N}(np, npq)$

 # np = 50*0.2 = 10

 # npq = 50*0.2*(1-0.2) = 50*0.2*0.8 = 8

 # => X ~ $\mathcal{N}(10, 8)$

 # Lưu ý 2: pnorm(x, mean, sd)

 # Mặc định: mean = 0 , sd = 1

 # Tức là nếu chỉ có pnorm(x) [= pnorm(x, mean = 0 , sd = 1)]

 # được hiểu là tính giá trị của hàm mật độ pp Chuẩn tắc $\mathcal{N}(0, 1)$

 # Lưu ý 3: Với X ~ $\mathcal{N}(\mu, \sigma^2)$

=> Z = $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

=> P(X < 25) = P(Z ≤ (25 - 10)/sqrt(8)) = P(Z ≤ 15/sqrt(8))

= pnorm(15/sqrt(8))

---> Đáp án: A. pnorm(15/sqrt(8))

Câu 16. Các cuộc gọi điện đến tổng đài tuân theo phân phối Poisson với mức λ trên mỗi phút. Từ kinh nghiệm có được trong quá khứ, ta biết rằng xác suất nhận được chính xác một cuộc gọi trong một phút bằng ba lần xác suất không nhận được cuộc gọi nào trong cùng thời gian. Ta xét khoảng 100 khoảng thời gian một phút liên tiếp và gọi U là số khoảng thời gian một phút không nhận được cuộc gọi nào. Viết câu lệnh tính $\mathbb{P}(U \leq 1)$

```
#lambda = 3

# Xác suất không nhận được cuộc gọi nào trong thời gian 1 phút là

p=dpois(0, lambda=3)

# U ~ B(100; p): Vay P(U ≤ 1) = F (1)

pbinom(1,100,p)
```

Câu 17. Cho

$z_{1-\alpha/2} \triangleq$ qnorm(1 - alpha/2) và $t^{n-1}_{1-\alpha/2} \triangleq$ qt(1 - alpha/2, df = n-1)

hoàn thành các chỗ trống trong đoạn code sau

```
path = 'D:// Works '
setwd ( path )
dtf = read .csv('data01 . csv ', header = TRUE )
Age = dtf $ Age
KTC _ mean <- function(data , alpha , sig = 'None '){
                n = length( data )
                m = mean( data )
                sd = sd( data )
                zalp = qnorm(1 - ... alpha.... /2)
                talp = qt(1 - alpha /2, ... n-1...)
                if(sig != 'None ')
                      eps = sig * ..... zalp..... / sqrt (n)
                else  if(sig == 'None ')
                      if( n < 30)
                            eps = sd* talp / sqrt (n)
                      else if(n >=...30...)
```

```
                eps = sd* ..... zalp..... / sqrt (n)
                return (c(m - eps , m + eps ))
}
KTC_mean (Age , 0.05)
```

Câu 18. Hàm KTC_mean cho biết

A. Input các tham số dữ liệu mẫu (data), độ tin cậy (alpha) và giả thiết về sigma.

=> SAI !

B. Output là khoảng tin cậy của trung bình trong các trường hợp biết phương sai, không biết phương sai và cỡ mẫu.

C. A, B đều sai .

D. A, B đều đúng.

 # Lưu ý: Độ tin cậy = 1 – alpha

 # Mức ý nghĩa = alpha

Câu 19.

```
path = 'D:// Works '
setwd ( path )
dtf = read .csv('data01 . csv ', header = TRUE )
Age = dtf $ Age
U70 = Age [ Age > 70]
KTC _ prop <- function ( data .p, data , alpha ){
phat = length ( data .p)/ length ( data ) ## ty le mau
eps = qnorm (1 - alpha /2)* sqrt ( phat *(1- phat )/n)
print ('KTC cho ty le la ')
return (c( phat - eps , phat + eps ))
}
KTC _ prop (U70 , Age , 0.05)
```

Hàm KTC _ prop cho biết

A. Input các tham số dữ liệu mẫu (data), dữ liệu thỏa tính chất nào đó để truy xuất tỷ lệ mẫu (data.p) và mức ý nghĩa (alpha).

B. Output là khoảng tin cậy cho tỷ lệ p với độ tin cậy α. => SAI !

C. A, B đều sai .

D. A, B đều đúng.

Xem đoạn code và kết quả sau

```
path = 'D:// Works '
setwd ( path )
data = read . csv ('rocket . motor . csv ', header = TRUE )
SK = data $ streng ; mu_0 = 2000
test = t.test(SK , alternative = " two . sided ", mu = mu_0, conf .
level = 0.95) => 1 − α = 95% #20.A
                One Sample t- test
data : SK
t = 1.9799 #21.B , df = 19, p- value = 0.06238
alternative hypothesis:true mean is not equal to 2000 => μ ≠
2000#20.A
95 percent confidence interval :
    1992.438    2272.377
sample estimates :
mean of x
   2132.407
```

Câu 20. Hàm **`t.test(SK , alternative = " two . sided ", mu = mu_0, conf . level = 0.95)`** dùng để

A. Khoảng tin cậy cho trung bình của mẫu với đối thuyết $\mu \neq 2000$ và độ tin cậy $1 - \alpha = 95\%$.

B. Khoảng tin cậy cho trung bình của mẫu với đối thuyết $\mu < 2000$ và độ tin cậy $1 - \alpha = 95\%$.

C. Khoảng tin cậy cho trung bình của mẫu với đối thuyết $\mu \neq 2000$ và mức ý nghĩa $1 - \alpha = 0.05$.

D. Khoảng tin cậy cho trung bình của mẫu với đối thuyết $\mu > 2000$ và độ tin cậy $1 - \alpha = 0.95$.

Câu 21. Kết quả của lệnh bằng

A. 0.9218 . **B.** 1.979949. **C.** 19 . **D.** 0.06238 .

Câu 22. Để kiểm định trung bình của hai mẫu X,Y độc lập với đối thuyết $\mu_X > \mu_Y$ và độ tin cậy $1 - \alpha = 0.95$, hãy viết một đoạn code thực hiện điều đó

```
t.test(X,Y,alternative="greater")
```

hoặc viết đầy đủ

```
t.test(X,Y,alternative="greater",conf.level=0.95,paired=F).
```

 # Lưu ý: hàm t.test:

 # alternative là đối thuyết kiểm định (mặc định là "two.sided").

 # conf.level = độ tin cậy = 1 – alpha (mặc định = 0.95).

Xét dữ liệu trong file house.price.csv với các tên biến như đoạn lệnh bên dưới, hãy trả lời các câu hỏi 23 và 24.

```
dtf = read .csv('house . price . csv ', header = TRUE )
Tax = dtf $ taxes
Sales = dtf$ sale . price
Tax2 = Tax [Tax > 8]
Sale2 = Sales[ Sales > 35]
```

Câu 23. Cho kết quả của kiểm định sau

```
test1
         Welch Two Sample t- test
data : Tax and sales
t = -22.2571 , df = 26.179 , p- value = 1 #4
alternative hypothesis : true difference in means #1 is greater than
0 #2
95 percent #3 confidence interval :
    -30.36865      Inf
sample estimates:
mean of x mean of y
  6.404917   34.612500
```

Hãy cho biết kết quả trên nói về kiểm định của **trung bình hai mẫu** (xem #1 phần màu xanh lá), trong đó đối thuyết của kiểm định là $H_1: \mu_1 - \mu_2 > 0$ (xem #2 phần màu tím) cùng với mức ý nghĩa 0.05 (xem #3 phần màu xanh dương độ tin cậy 95% => alpha = 5% = 0.05) , ta có thể kết luận rằng **chưa đủ cơ sở để bác bỏ H_0, nghĩa là trung bình của biến Tax không lớn hơn trung bình của biến Sales với mức ý nghĩa 5%** (xem #4 phần màu xám, bác bỏ H0 khi p-value ≤ alpha).

Câu 24. Cho kết quả của kiểm định sau

```
test2
           2- sample test for equality of proportions #1 with
continuity correcti
data : y out of n
X- squared = 2.2222 , df = 1, p- value = 0.06802 #3
alternative hypothesis : less #2
95 percent confidence interval :
   -1.00000000     0.01374729
sample estimates :
prop 1 prop 2
   0.25    0.50
```

Hãy cho biết kết quả trên nói về kiểm định của **tỷ lệ hai mẫu** (xem #1 xanh lá), trong đó đối thuyết của kiểm định là $H_1: p_1 - p_2 < 0$ (xem #2 xanh dương) , cùng với p – giá trị bằng

0.06802 (xem #3 tím) , ta có thể kết luận rằng **chưa đủ cơ sở để bác bỏ H_0, nghĩa là tỷ lệ mẫu 1 không nhỏ hơn tỷ lệ mẫu 2 với mức ý nghĩa 5%** (vì p – giá trị = 0.06802 > 0.05 = alpha).

Câu 25. Cho mẫu X, dùng những hàm có sẵn hãy viết các đoạn lệnh thực hiện tính *trung bình mẫu:* **mean(X)** *, phương sai mẫu:* **var(x)** *, trung vị mẫu:* **median(X)** *và độ lệch chuẩn (mẫu):***sd(X)**

Câu 26. Để tính p – giá trị của Y = 18 với Y ~ B(50, 0.5) với đối thuyết $H_1: p \neq p_0$, hãy hoàn thành đoạn code sau

z_0 = ...
p. value = 2* min (pnorm (.....) , 1 - .....(z_ 0))

ĐỀ LỖI => BỎ

Xem đoạn lệnh sau và kết quả của nó để trả lời các câu hỏi 27 và 28.

```
df = read . csv ('chloride . csv ', header = TRUE )
y = df$y ## Nong do cloride
x = df$x ## ty le phan tram
lm(y ~ x)

Call :
lm( formula = y ~ x)

Coefficients
( Intercept )                    x
0.4705                     20.5673
```

Câu 27. Đoạn lệnh trên cho biết giá trị các hệ số ước lượng $\widehat{\beta_0}$ và $\widehat{\beta_1}$ lần lượt bằng **0.4705 và 20.5673**

Câu 28. Đoạn lệnh trên cho biết

A. Kết quả mô hình hồi quy $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, cho y- nồng độ clorua(đv: mg/l) theo x- diện tích ở đầu nguồn x(đv:%) với các hệ số hồi quy $\widehat{\beta_0} = 20.5673$ và $\widehat{\beta_1} = 0.4705$.

B. Kết quả mô hình hồi quy $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, cho y- nồng độ clorua(đv: mg/l) theo x- diện tích ở đầu nguồn x(đv:%) với các hệ số hồi quy $\widehat{\beta_0} = 0.4705$ và $\widehat{\beta_1} = 20.5673$.

C. Kết quả mô hình hồi quy $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, cho x- diện tích ở đầu nguồn x(đv:%) theo y- nồng độ clorua(đv: mg/l) với các hệ số hồi quy $\widehat{\beta_0} = 20.5673$ và $\widehat{\beta_1} = 0.4705$.

D. Kết quả mô hình hồi quy $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, cho x- diện tích ở đầu nguồn x(đv:%) theo y- nồng độ clorua(đv: mg/l) với các hệ số hồi quy $\widehat{\beta_0} = 0.4705$ và $\widehat{\beta_1} = 20.5673$.

 # lm(y~x): hồi quy y theo x ---> loại C,D

 # hệ hố B0, B1 lần lượt là 0.4705 và 20.5673

---> Đáp án: B

Câu 29. Để tìm tính khoảng tin cậy 99% cho β_0, hãy hoàn chỉnh vào đoạn lệnh sau , biết rằng $\beta_0 \in \left[\hat{\beta}_0 - t^{n-2}_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}}{Sxx}\right)}; \hat{\beta}_0 + t^{n-2}_{1-\alpha/2}\sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}}{Sxx}\right)}\right]$ trong đó $\hat{\beta}_0$ là hệ số góc trong mô hình hồi quy và MSE, SSE, Sxx thoả các công thức sau

$$MSE = \frac{SSE}{n-2}, \quad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad Sxx = \sum_{i=1}^{n}(x_i - \bar{x})^2].$$

```
beta0_CI <- function(x, y, ..alpha..){

                n = length(x); x.mean = mean(x)
                result = lm(y~x) # kết quả của mô hình hồi quy tuyến tính y theo x
được gán vào biến result
                res = resid( result ) ## Tính các giá trị thặng dư (y_i − ŷ_i)
                beta0.hat = ( coef ( result ))[[1]]
                MSE = ....... sum( res^2 )/(n-2)..... ##SSE= sum(
res^2 )
                Sxx = sum((x-x.mean)^..2.)
                eps = qt (1 - alpha /2, df=n -2) * sqrt (..MSE...
*(1/n       + x. mean **2/ ...Sxx..))
                print ('KTC cho beta 0')
                return (c( beta0 . hat - eps , beta0 . hat + eps ))
}
beta0 _CI(x, y, 0.01)
```

Câu 30. Một nhóm sinh viên đo nhiệt độ ở những độ cao khác nhau và thu được bảng số liệu sau:

Elevation(ft)	600	1000	1250	1600	1800	2100	2500	2900
Temperature(F)	56	54	56	50	47	49	47	45

Sử dụng các câu lệnh trong R để vẽ đồ thị phân tán và đường hồi quy nhiệt độ theo độ cao cùng hệ trục tọa độ.

 # Lưu ý cấu trúc lm(y~x): xấp xỉ y = f(x). Chú ý thứ tự giữa x và y !

```
Elevation<-c(600,1000,1250,1600,1800,2100,2500,2900)

Temperature<-c(56,54,56,50,47,49,47,45)

plot(Elevation,Temperature)

abline(lm(Temperature Elevation))
```