BÀI THỰC HÀNH SỐ 5

1. Phân phối mẫu

a) Tham sô đặc trưng thống kê mẫu

Trung bình mẫu: Cho n giá trị quan trắc $X_1, \dots, X_n$, ta có:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Phương sai mẫu: Cho n giá trị quan trắc $X_1, \dots, X_n$, ta có:

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Độ lệch chuẩn mẫu: là căn bậc hai của phương sai

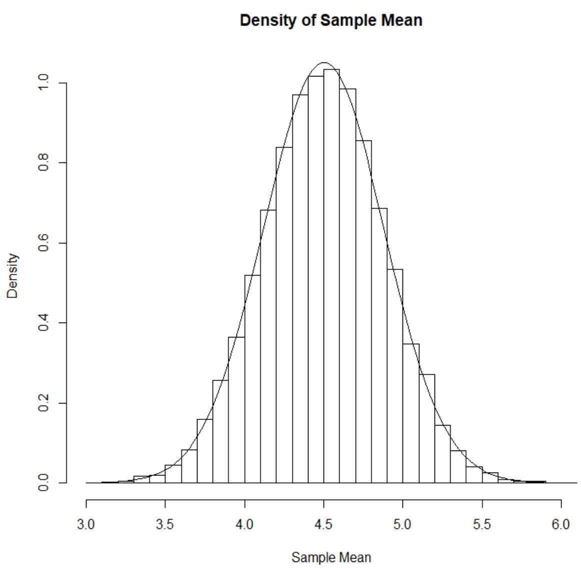
b) Định lý

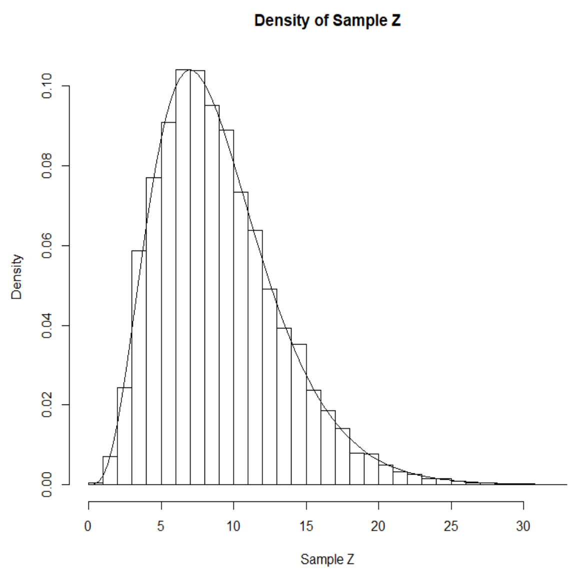
Cho $X_1, X_2 \dots, X_n$ là mẫu ngẫu nhiên lấy từ phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$. Ta có:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad và \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

```
mu = 4.5
sigma = 1.2
n=10
m=10000
#################################################
MeanX = function(n){
  X = rnorm(n,mu,sigma)
  mean(X)
}

SampleMeanX = function(n,m){
  replicate(m,MeanX(n))
}

hist(SampleMeanX(n,m),freq = 0,breaks = 40, main = "Density
of Sample Mean", xlab = "Sample Mean")
curve(dnorm(x,mu,sigma/sqrt(n)),add=TRUE)

##################################################
Z = function(n){
  X = rnorm(n,mu,sigma)
  ((n-1)*var(X))/(sigma^2)
```

```
}
SampleZ = function(n,m){
  replicate(m,Z(n))
}

hist(SampleZ(n,m),freq = 0,breaks = 40, main = "Density of
Sample Z", xlab = "Sample Z")
curve(dchisq(x,df=n-1),add = TRUE)
```





2. Định lý giới hạn - Định lý giới hạn trung tâm

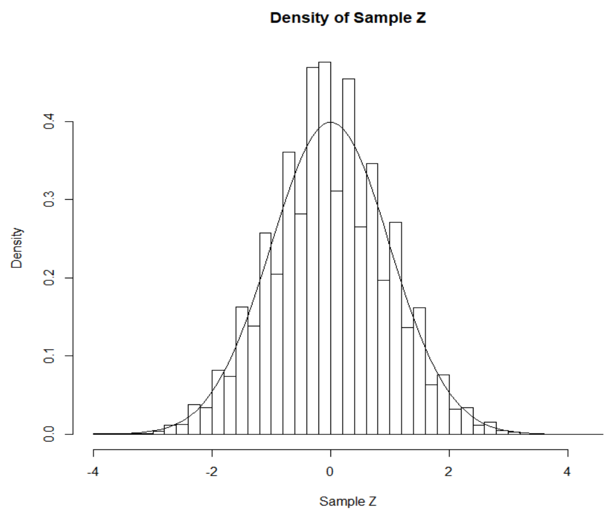
a) Định lý giới hạn trung tâm

Cho $X_1, X_2 \dots, X_n$ là mẫu ngẫu nhiên của tổng thể (có thể hữu hạn hoặc vô hạn) với một phân phối có trung bình μ và phương sai σ^2. Nếu $\hat{\mu} = \bar{X}$ là trung bình mẫu thì ta có:

$$Z_n = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0,1) \quad khi \quad n \to \infty$$

```
size = 10
prob = 0.2
n = 100
m = 10000
####################################
Z = function(n){
  X = rbinom(n,size,prob)
  mu = size*prob
  var = size*prob*(1-prob)
  (mean(X)-mu)/(sqrt(var)/sqrt(n))
}
```

```
SampleZ = function(n,m){
  replicate(m,Z(n))
}

hist(SampleZ(n,m),freq = 0,breaks = 40,main = "Density of
Sample Z", xlab = "Sample Z")
curve(dnorm(x,0,1),add = TRUE)
```

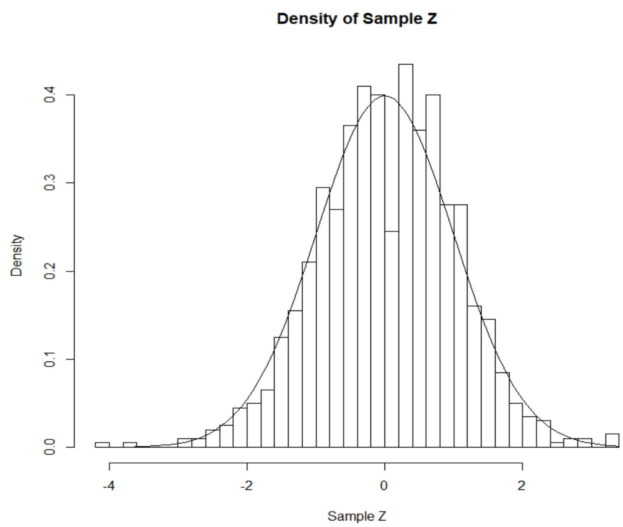


b) Định lý giới hạn (Sử dụng phân phối chuẩn tắc để xấp xỉ phân phối nhị thức)

Cho $X \sim B(n,p)$ thì biến ngẫu nhiên $Z = \frac{X-np}{\sqrt{np(1-p)}}$ được xấp xỉ bởi phân phối chuẩn tắc khi $np > 5$ và $n(1-p) > 5$.

```
size = 10
prob = 0.3
n = 100
m = 1000
##########################################################
Z = function(n){
  X = rbinom(n,size,prob)
  (mean(X)-(size*prob))/(sqrt(size*prob*(1-prob))/sqrt(n))
}

SampleZ = function(n,m){
  replicate(m,Z(n))
}
```

```
hist(SampleZ(n,m),freq = 0,breaks = 40,main = "Density of
Sample Z", xlab = "Sample Z")
curve(dnorm(x,0,1),add = TRUE)
```



3. <u>Các lệnh cơ bản trong R về lý thuyết mẫu</u>

- `max(x)`, `min(x)`: giá trị lớn nhất, bé nhất của x.
- `sum(x)`: tổng các giá trị trong x
- `mean(x)`: trung bình của x
- `median(x)`: trung vị của x
- `range(x)`: bằng max(x) – min(x)
- `var(x)`: phương sai của x
- `sd(x)`: độ lệch chuẩn của x
- `quantile(x)`: tính các phân vị của x
- `sort(x)`: sắp xếp x, mặc định theo thứ tự tăng dần
- `order(x)`: trả về các vị trí của x khi đã sắp theo thứ tự tăng dần
- `cumsum(x)`: tổng tích lũy
- `cumprod(x)`: tích tích lũy

4. <u>Ước lượng khoảng cho tham số thống kê</u>

a) Khoảng tin cậy cho trung bình

Các bước thực hiện

B1:Tìm trung bình mẫu $\bar{X}$ `mean(X)` và phương sai mẫu s^2 `var(X)` hoặc độ lệch chuẩn mẫu s `sd(X)`

B2:Xác định các trường hợp

TH1: $n \geq 30$ (hoặc $n < 30$, X có phân phối chuẩn) và σ^2 đã biết
TH2: $n \geq 30$ và σ^2 chưa biết
TH3: $n < 30$, X có phân phối chuẩn và σ^2 chưa biết

B3:Tìm phân vị

- Nếu là TH1 và TH2 thì tìm $z_{\alpha/2}$ `qnorm(1-alpha/2)`
- Nếu là TH3 thì tìm $t_{\alpha/2}^{n-1}$ `qt(1-alpha/2, df = n-1)`

Chú ý: $z_{\alpha/2}$ thỏa $\mathbb{P}\left(Z \leq z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$ với $Z \sim \mathcal{N}(0,1)$

$t_{\alpha/2}^{n-1}$ thỏa $\mathbb{P}\left(|T| > t_{\alpha/2}^{n-1}\right) = \alpha$ với $T \sim t(n-1)$

B4:Tìm dung sai

$$\varepsilon = \begin{cases} z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}} & nếu\ TH1 \\ z_{\alpha/2}\dfrac{s}{\sqrt{n}} & nếu\ TH2 \\ t_{\alpha/2}^{n-1}\dfrac{\sigma}{\sqrt{n}} & nếu\ TH3 \end{cases}$$

B5:Kết luận khoảng tin cậy $100(1-\alpha)\%$ cho trung bình của tổng thể là

$$[\bar{X} - \varepsilon\ , \bar{X} - \varepsilon]$$

b) Khoảng tin cậy cho tỉ lệ của phân phối nhị thức

Các bước thực hiện

B1:Tìm tỉ lệ mẫu $\hat{p} = \frac{X}{n}$ `rbinom(m,n,p)/n`
B2:Kiểm tra điều kiện $n\hat{p} \geq 5$ và $n(1-\hat{p}) \geq 5$
B3:Tìm phân vị $z_{\alpha/2}$ `qnorm(1-alpha/2)`
B4: Tìm dung sai $\varepsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
B5:Kết luận khoảng tin cậy $100(1-\alpha)\%$ cho tỉ lệ tổng thể là

$$\hat{p} - \varepsilon\,, \hat{p} + \varepsilon$$

5. **Bài tập**

Bài 1: Cho X_1, X_2 là mẫu ngẫu nhiên kích thước 2 lấy từ phân phối chuẩn tắc $\mathcal{N}(0,1)$. Dùng hàm `rnorm()` phát sinh X_1, X_2 và $Y = {X_1}^2 + {X_2}^2$. Xây dựng hàm SampleY phát sinh mẫu ngẫu nhiên kích thước n cho Y. Lần lượt phát sinh mẫu ngẫu nhiên kích thước 100, 1000,10000 cho Y, vẽ biểu đồ tần suất `hist()` và đồ thị hàm mật độ xác suất của phân phối Chi – bình phương với 2 bậc tự do cho từng trường hợp.

Bài 2:

a) Tạo ngẫu nhiên 100 giá trị có phân phối nhị thức, với n = 60 và xác suất thành công mỗi lần 0.4. Vẽ biểu đồ tần số.
b) Tạo ngẫu nhiên 100 giá trị có phân phối Poisson với $\lambda = 4$, vẽ biểu đồ tần số.

c) Tạo ngẫu nhiên 100 giá trị có phân phối chuẩn có trung bình là 50 và độ lệch tiêu chuẩn 4. Vẽ hàm phân phối, hàm mật độ.
d) Tạo ngẫu nhiên 100 giá trị có phân phối mũ với $\lambda = \frac{1}{25}$. Vẽ hàm phân phối, hàm mật độ.

<u>Bài 3</u>: File *diesel_engine.dat* và *diesel_time.xls* chứa số liệu về hoạt động của các động cơ chạy bằng dầu diesel. Thực hiện:

a) Đọc số liệu từ hai file này, gán và hai dataframe, đặt tên hai dataframe cùng tên với file.
b) Liệt kê tên các biến có trong hai dataframe vừa nhập.
c) Xác định có bao nhiêu dữ liệu bị khuyết (missing data) trong *diesel_engine*. Thay thế các giá trị khuyết trong biến *speed* bằng 1500, biến *load* bằng 20.
d) Tính: trung bình, phương sai, độ lệch tiêu chuẩn, giá trị lớn nhất, nhỏ nhất của biến *alcohol* trong dataframe *diesel_engine*.
e) Ghép hai dataframe *diesel_engine* và *diesel_time* lại thành một frame có tên là *diesel*.
f) Trích giá trị của biến *run* (số thứ tự các động cơ) mà có thời gian trễ (biến *delay*) dưới 1.000.
g) Đếm xem có bao nhiêu động cơ có *timing* bằng 30.
h) Vẽ biểu đồ boxplot cho các biến *speed, timing* và *delay*. (dùng hàm `boxplot`)
i) Vẽ biểu đồ phân tán cho các cặp biến (*timing, speed*), (*temp, press*). (dùng hàm `plot`)
j) Chuyển biến *load* sang biến nhân tố.
k) Chia phạm vi giá trị của biến *delay* thành 4 đoạn đều nhau và đếm số giá trị nằm trong các đoạn đó. Tạo bảng thống kê và vẽ biểu đồ cột.
l) Chia phạm vi giá trị của biến *delay* thành 4 đoạn như sau: (0.283, 0.7], (0.7, 0.95], (0.95, 1.2], (1.2, 1.56]. Tạo bảng thống kê và vẽ biểu đồ cột.

<u>Bài 4</u>: Cho số liệu sau:

year	*snow.cover*
1970	6.5
1971	12.0
1972	14.9
1973	10.0
1974	10.7
1975	7.9
1976	21.9
1977	12.5
1978	14.5
1979	9.2

a) Nhập số liệu trên vào R.

b) Vẽ *snow.cover* theo *year*.
c) Vẽ biểu đồ histogram cho *snow.cover*.
d) Lặp lại câu b. và c. sau khi lấy logarit của biến *snow.cover*.

<u>Bài 5</u>: Cho số liệu sau:

Temperature (F)	*Erosion incidents*	*Blowby incidents*	*Total incidents*
53	3	2	5
57	1	0	1
63	1	0	1
70	1	0	1
70	1	0	1
75	0	2	1

Nhập số liệu trên vào một dataframe, vẽ đồ thị biểu diễn tổng số *incidents* theo *temperature*.

<u>Bài 6</u>: Thống kê số liệu tỉ lệ lạm phát tại 4 nước trong giai đọan 1960-1980 được thu thập trong 2 bảng số liệu sau (Đvt: %)

Nam	**US**	**Anh**	**Nhat**	**Duc**
1960	**1.5**	**1**	**3.6**	**1.5**
1961	**1.1**	**3.4**	**5.4**	**2.3**
1962	**1.1**	**4.5**	**6.7**	**4.5**
1963	**1.2**	**2.5**	**7.7**	**3**
1964	**1.4**	**3.9**	**3.9**	**2.3**
1965	**1.6**	**4.6**	**6.5**	**3.4**
1966	**2.8**	**3.7**	**6**	**3.5**
1967	**2.8**	**2.4**	**4**	**1.5**
1968	**4.2**	**4.8**	**5.5**	**18**
1969	**5**	**5.2**	**5.1**	**2.6**
1970	**5.9**	**6.5**	**7.6**	**3.7**
1971	**4.3**	**9.5**	**6.3**	**5.3**
1972	**3.6**	**6.8**	**4.9**	**5.4**
1973	**6.2**	**8.4**	**12**	**7**
1974	**10.9**	**16**	**24.6**	**7**
1975	**9.2**	**24.2**	**11.7**	**5.9**
1976	**5.8**	**16.5**	**9.3**	**4.5**
1977	**6.4**	**15.9**	**8.1**	**3.7**
1978	**7.6**	**8.3**	**3.8**	**2.7**
1979	**11.4**	**13.4**	**3.6**	**4.1**
1980	**13.6**	**18**	**8**	**5.5**

a) Nhập dữ liệu trên vào 2 data.frame *lamphat1* và *lamphat2* trong R bằng 3 cách.
b) Trộn 2 data.frame trên vào 1 data.frame duy nhất là *lamphat* theo Nam.
c) Đếm số năm các nước US, Anh, Nhật, Đức có tỉ lệ lạm phát trên 5%.

d) Vẽ đồ thị phân tán về tỉ lệ lạm phát cho mỗi quốc gia theo thời gian. Cho nhận xét tổng quát về lạm phát của 4 nước?
e) Tính trung bình, trung vị, Max, Min, độ lệch chuẩn, sai số chuẩn của từng nước?
f) Để xác định lạm phát nước nào biến thiên nhiều hơn, ta cần dựa vào tham số thống kê nào? Kết luận?
g) Tạo một data.frame mới *lamphat1* với số biến như trong data.frame *lamphat* nhưng không chứa dữ liệu của năm 1980.
h) Ta biết rằng hệ số của phương trình hồi quy tuyến tính $\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2}\widehat{X_i}$ được xác định như sau:

$$\widehat{\beta_2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} {X_i}^2 - n(\bar{X})^2}$$
$$\widehat{\beta_1} = \bar{Y} - \widehat{\beta_2}\bar{X}$$

Xác định các hệ số này trong mô hình hồi quy: lạm phát theo thời gian cho US bằng cách sử dụng data.frame *lamphat1*. Vẽ đồ thị phương trình hồi quy này.

i) Sử dụng phương trình hồi quy trong câu h) hãy xác định tỉ lệ lạm phát trong năm 1980 của US. So sánh số liệu thực tế

<u>Bài 7</u>: Tạo ngẫu nhiên 35 giá trị của biến ngẫu nhiên có phân phối chuẩn với trung bình bằng 10 và độ lệch chuẩn 5. Tìm khoảng tin cậy 95% cho kỳ vọng của biến ngẫu nhiên chuẩn dựa vào số liệu vừa tạo.

<u>Bài 8</u>: Số liệu thống kê về doanh số bán hàng của một siêu thị cho ở file data31.xls:

a) Đọc dữ liệu từ file data31.xls vào R.
b) Viết hàm ***ci.mean(x, alpha)*** xuất ra khoảng tin cậy cho kỳ vọng, với x là vec-tơ dữ liệu, (1-alpha) là độ tin cậy. Áp dụng để tìm khoảng tin cậy 95% và 99% cho doanh số bán hàng trung bình ở siêu thị.

<u>Bài 9</u>: File data32.xls chứa số liệu về thời gian tự học của 120 sinh viên trường ĐH Khoa học Tự nhiên.

a) Hãy ước lượng thời gian học nhóm trung bình của sinh viên trường ĐH KHTN, độ tin cậy là 95%. (Dùng hàm ***ci.mean(x, alpha)***)
b) Viết hàm hàm ***ci.prop(f, n, alpha)*** xuất ra khoảng tin cậy cho tỷ lệ, với n là cỡ mẫu; f: số các phần tử thỏa yêu cầu (với tỷ lệ p cần tìm); (1-alpha) là độ tin cậy. Áp dụng để tìm khoảng tin cậy 90%; 95% và 99% cho tỷ lệ sinh viên có thời gian tự học trên 5 giờ mỗi ngày.

<u>Bài 10</u>: Bảng sau thống kê chiều cao (Đv: m) của 125 thanh niên 18 tuổi trong một khu vực:

Chiều cao	[1.2,1.4)	[1.4,1.6)	[1.6,1.8)	[1.8,2.0)	[2.0,2.2)
Số thanh niên	6	34	31	42	12

a) Chuyển bảng tần số dạng khoảng ở trên thành dữ liệu dạng véc-tơ cột. Áp dụng hàm ***ci.mean***() đã ở bài **8** để tìm khoảng tin cậy 95% cho chiều cao trung bình của thanh niên trong khu vực.

b) Những người có chiều cao từ 1.7 m trở lên được xếp vào sức khỏa loại A. Sử dụng hàm ***ci.prop()*** ở bài **9** để tìm khoảng tin cậy 95% cho tỷ lệ thanh niên đạt sức khoẻ loại A.

<u>Bài 11</u>: Viết hàm ***ktc.tb()*** để tìm khoảng tin cậy cho trung bình biết:

- Input: là trung bình mẫu $\bar{x}$ độ lệch chuẩn của tổng thể σ (có thể biết trước hoặc không), trường hợp không biết α thì phải nhập độ lệch chuẩn của mẫu s, kích thước mẫu n, và mức ý nghĩa α.
- Output: khoảng tin cậy cho trung bình.

<u>Bài 12</u>: Từ hàm được viết trong bài 11 hãy viết hàm ***ktc.tb.mau()*** để tìm khoảng tin cậy cho trung bình biết:

- Input: vecto dữ liệu mẫu x, độ lệch chuẩn của tổng thể σ (có thể biết trước hoặc không), và mức ý nghĩa α
- Output: khoảng tin cậy cho trung bình.

<u>Bài 13</u>: Đo đường kính của một chi tiết máy do một máy tiện tự động sản xuất, ta ghi nhận được số liệu như sau:

X	12.00	12.05	12.10	12.15	12.20	12.25	12.30	12.35	12.40
n	2	3	7	9	10	8	6	5	3

Bằng cách sử dụng hàm ***ktc.tb.mau()*** trong câu 6), hãy ước lượng khoảng tin cậy 95% cho đường kính trung bình.