Author: Tan Duc Nguyen Email: tanduc307@gmai.com Updated 08/02/21

Population mean
Trung bình tổng thể

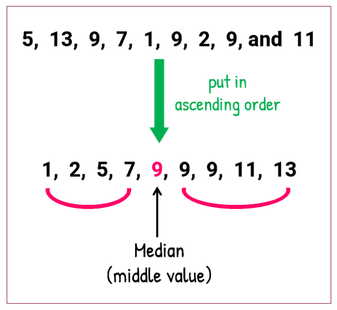
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

N = number of items in the population

Sample mean
Trung bình của mẫu

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

n = number of items in the sample



1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= **4.5**

LƯU Ý:
Muốn kiểm tra lại kết quả thì copy dữ liệu này vào R nhưng nhớ là copy đủ hết các chữ số sau dấu phẩy nhé!

object	a
object	table_a
No.	Value
1	11.198390805906300
2	11.053263022578600
3	10.382705473424200
4	9.345206992010660
5	8.555006222382860
6	9.484966278938550
7	8.909901662984320
8	9.568169663346200
9	9.780520600260730
10	10.162300023414300
11	9.114191705124020
12	11.148396732135000
13	10.876914178707100
14	8.734608046701930
15	7.666503610462450
16	10.440860682753600
17	10.219574160326100
18	11.936864710027000
19	9.054128980162910
20	9.911834153929750
21	10.131061187099700
22	8.111884514089440
23	10.546298568504000
24	11.004267393572100
25	10.702056739105900
26	10.830809121142900
27	7.619064391273120
28	8.713936955810440
29	8.613955400248210
30	10.577363603353500
31	11.456215881253500
32	9.544904647442960
33	10.800376406015400
34	8.911564832929680
35	8.905465400351500
36	10.454934717780400
37	10.330119741927400
38	10.392370649903500
39	9.375425213272920
40	8.607389635423660
41	11.666664648959500
42	10.485056111293600
43	10.615103712927800
44	10.614566092864200
45	9.860151050519850
46	9.030035815765350
47	11.625027191807500
48	9.082426981534370
49	9.014947651309000
50	10.156433318001200
51	10.401904110253400
52	10.128284111801900
53	9.688521027985390
54	11.647319235869300
55	10.559115196678200
56	10.453469999576100
57	11.053631369693900
58	9.910021038846760
59	8.868427934496110
60	9.757769838295860
61	10.242963058168200
62	10.833191018994100
63	8.624560307513080
64	10.448723886918600
65	9.719696121766920
66	10.011084658112500
67	8.498110057007280
68	11.527431339949400
69	10.160726103482000
70	9.785764464876790
71	11.352136308791000
72	9.781107600060460
73	10.482257645132300
74	11.827449528390100
75	10.894586443580600
76	10.771098026904400
77	10.544594119565900
78	10.484488858570300
79	10.453292650134900
80	9.999333663835750
81	9.429949815418860
82	10.564398345884500
83	7.844235146847470
84	7.840269251378430
85	8.387553012797840
86	9.848167032900210
87	9.496114935897990
88	11.406217075685700
89	9.196013618457260
90	12.378941573522200
91	9.077365278045710
92	10.167681080102600
93	9.800497822596300
94	11.043914238978700
95	8.907868481264600
96	9.093803415684730
97	10.299991520067600
98	9.780735497672850
99	11.183526791607700
100	9.136437475445020

Giải thích ý nghĩa population mean và sample mean

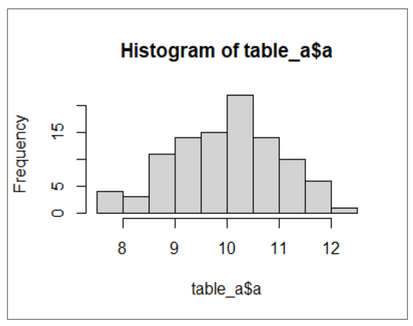
Trong R gõ code sau:

```
a = rnorm(n = 100, mean = 10, sd = 1)
#Có nghĩa là tạo ra một dãy gồm 100 số, có trung bình tổng thể là 10 và độ lệch chuẩn tổng thể là 1.
Thường chỗ này hay có ví dụ là giả sử đây là số đo bất kỳ của 100 cây con, có chiều cao trung bình (tổng thể) và độ lệch chuẩn (tổng thể) TỰ CHO TRƯỚC là mean=10 và sd=1).
Và dãy số này sẽ được tạo ra ngẫu nhiên làm sao để tuân theo quy luật phân phối chuẩn (vì dùng hàm rnorm).

table_a = data.frame(a)
#Tạo ra một object tên là table_a để nhét dãy dữ liệu vào thành dataset, ở dạng table có dòng và cột để thuận tiện phân tích.
Kiểm tra coi hình dáng table_a dùng hàm
fix(table_a) hoặc View(table_a) #chú ý chữ V trong view viết hoa

#Để copy data.frame từ R qua Excel thì dùng code này.
write.table(table_a, file="D:/abc.csv", row.names=F, sep=",")
Trong đó,
table_a là tên đối tượng chứa dữ liệu bảng
file = "D:/abc.csv" là đường dẫn lưu ở dạng file .csv
row.names=F là lệnh nói khi export qua .csv thì xóa đi cột số thứ tự (do mặc định khi tạo data.frame là R sẽ tạo ra cột này để thuận tiện kiểm tra, tương ứng từng dòng row gọi là observation).
sep="," là thông số để chọn cách export file .csv theo nhiều kiểu khác nhau.

dim(table_a) #Kiểm tra có bao nhiêu hàng và cột trong object table_a
names(table_a) #Kiểm tra liệt kê tên biến (variable) hay là cột trong object table_a

Để kiểm tra phân bố tần số histogram, ta cần dùng lệnh
attach(table_a) #Có nghĩa là đưa object table_a vào phân tích, lúc này chỉ cần gọi trực tiếp tên cột để xử lý. Nếu đã xử lý xong bảng ở object table_a này rồi thì nên detach(table_a) để trở về bình thường.
Lúc này sẽ gọi trực tiếp
hist(a)
hoặc
hist(table_a$a) #Nếu không dùng lệch attach thì để gọi biến a trong object table_a thì phải có dấu dollar sign.

Xác định đặc trưng thống kê về giá trị mean và sd của mẫu (từ 100 giá trị của tổng thể)
mean(a)
hoặc
mean(table_a$a) #Tìm giá trị trung bình của mẫu trên 100 giá trị trong object table_a

sd(a)
hoặc
sd(table_a$a) #Tìm giá trị độ lệch chuẩn của mẫu trên 100 giá trị trong object table_a
```

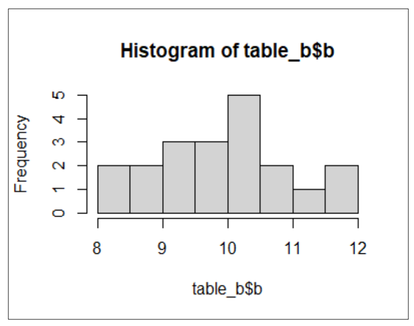
Tiếp tục, ta lại tạo ra một bộ dữ liệu khác

```
b = rnorm(n = 20, mean = 10, sd = 1)
#Có nghĩa là tạo ra một dãy gồm 20 số, có trung bình tổng thể là 10 và độ lệch chuẩn tổng thể là 1. Ý tưởng ở đây là để kiểm tra coi cùng với thông số mặc định TỰ CHO TRƯỚC mean=10 và sd=1 như đã nói ở trên, nhưng giờ nếu chỉ lấy có 20 giá trị bất kỳ từ tổng thể (n giá trị vẫn chưa biết lớn bao nhiêu!) thì mean và sd của mẫu giá trị sẽ thay đổi bao nhiêu, có gần với mean và sd của tổng thể hay không?

Tương tự, tạo table_b và histogram trên 20 giá trị này.

So sánh mean và sd của mẫu (với 20 giá trị từ tổng thể) bạn sẽ thấy là nếu lấy ít giá trị thì kết quả mean và sd của mẫu sẽ khó gần với mean và sd của tổng thể (TỰ CHO TRƯỚC), đây gọi là định lý số lớn Law of Large Numbers (LLN).

hist(table_b$b)
mean(table_b$b)
sd(table_b$b)
```

Histogram vẽ từ 100 giá trị table_a, xoay quanh giá trị trung bình của tổng thể mean = 10 và sd tổng thể = 1



Histogram vẽ từ 20 giá trị table_b, xoay quanh giá trị trung bình của tổng thể mean = 10 và sd tổng thể = 1



Từ 100 giá trị		
mean của mẫu		9.991189
sd của mẫu		1.021439

Tỷ lệ sai số so với mean tổng thể (%)	0.09
Tỷ lệ sai số so với sd tổng thể (%)	2.14

Từ 20 giá trị		
mean của mẫu		9.910932
sd của mẫu		1.060459

Tỷ lệ sai số so với mean tổng thể (%)	0.89
Tỷ lệ sai số so với sd tổng thể (%)	6.05

Từ n giá trị rất lớn không biết được			
mean của tổng thể (tự cho trước)			10
sd của tổng thể (tự cho trước)			1

Ta thấy rõ là tỷ lệ sai số so với tổng thể khi TĂNG số lượng lấy mẫu sẽ giảm gần về 0 hơn so với khi GIẢM số lượng lấy mẫu.

Đây là ý nghĩa của định lý số lớn.
According to the law, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed.

Tìm hiểu về định lý số nhỏ Law of Small Numbers và phân biệt với Law of Truly Large Numbers.

Note: Độ lệch chuẩn là căn bậc hai của phương sai.

object	b
object	table_b
No.	Value
1	9.139600569468480
2	9.659737841753540
3	9.985955950207770
4	10.422471092093500
5	10.539870227320600
6	8.154095832416740
7	9.272581948687760
8	11.505456631208600
9	8.605308752325180
10	9.088472277179400
11	11.728176048131800
12	8.008229223460620
13	10.194129585880100
14	10.732986757208000
15	8.896683633724700
16	10.234936414336000
17	9.849881322593620
18	10.443336739632000
19	11.451542288548600
20	10.305195996545600

Đây là bộ dữ liệu được tạo ra theo số liệu cho trước
Chọn ngẫu nhiên 100 giá trị từ trong n giá trị của tổng thể, mean tổng thể = 10, sd tổng thể = 1

mean của mẫu (trên 100 giá trị này)

Kết quả từ R	9.991189	hàm excel =AVERAGE

sd của mẫu (trên 100 giá trị này, ở ý nghĩa là 100 giá trị này là một phần của tổng thể)

Kết quả từ R	1.021439	hàm excel =STDEV.S
	1.021439	hàm excel =STDEV (cũ)

sd tổng thể (trên 100 giá trị này, NHƯNG ở ý nghĩa là 100 giá trị này là toàn bộ tổng thể)

	1.016319	hàm excel =STDEV.P
	1.016319	hàm excel =STDEVP (cũ)

Đây là bộ dữ liệu được tạo ra theo số liệu cho trước
Chọn ngẫu nhiên 20 giá trị từ trong n giá trị của tổng thể, mean tổng thể = 10, sd tổng thể = 1

mean của mẫu (trên 20 giá trị này)

Kết quả từ R	9.910932	hàm excel =AVERAGE

sd của mẫu (trên 20 giá trị này, ở ý nghĩa là 20 giá trị này là một phần của tổng thể)

Kết quả từ R	1.060459	hàm excel =STDEV.S