

TRƯỜNG ĐẠI HỌC NHA TRANG
KHOA CÔNG NGHỆ THÔNG TIN

---*****---



BÀI THI GIỮA KÌ
MÔN: THỐNG KÊ MÁY TÍNH
LỚP : 61. CNTT-1

Nhóm sinh viên thực hiện

- | | |
|---------------------------|------------------------|
| 1. Hoàng Minh Tâm | 6. Nguyễn Đình Sơn |
| 2. Trương Thị Diễm Quỳnh | 7. Nguyễn Văn Tâm |
| 3. Hoàng Minh Quân | 8. Trần Công Quyền |
| 4. Trương Minh Phi | 9. Nguyễn Phú Tâm |
| 5. Nguyễn Hoàng Minh Phúc | 10. Phan Trần Hữu Phúc |

Năm học: 2021 - 2022

Mục lục

Lời cảm ơn!.....	3
Tài liệu tham khảo	3
Bài 3.37: (Quỳnh)	4
Bài làm.....	5
Bài 3.38: (M.Tâm, Quỳnh)	8
Bài làm.....	9
Bài 3.39: (Quân)	10
Bài làm.....	11
Bài 3.40: (Phi)	13
Bài làm.....	13
Bài 3.41: (M.Phúc).....	16
Bài làm.....	16
Bài 3.42: (Sơn)	18
Bài làm.....	18
Bài 3.43: (V.Tâm)	21
Bài làm.....	21
Bài 3.44: (Quỳnh).....	24
Bài làm.....	25
Bài 3.45: (P.Tâm).....	29
Bài làm.....	30
Bài 3.46: (H.Phúc)	34
Bài làm.....	34
Bài 3.47: (M.Tâm)	36
Bài làm.....	36

Lời cảm ơn!

Lời đầu tiên, cho phép nhóm chúng em gửi lời cảm ơn sâu sắc và chân thành nhất đến quý thầy/cô và các bạn học đã tạo điều kiện giúp chúng em trong suốt quá trình học tập và thực hiện đề tài. Sự quan tâm và giúp đỡ của quý thầy/cô và các bạn học là nguồn động viên rất lớn giúp chúng em hoàn thành tốt dự án này.

Với lòng biết ơn sâu sắc nhất, chúng em xin gửi đến quý thầy/cô ở Khoa Công Nghệ Thông Tin đã truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Nhờ có những lời hướng dẫn, dạy bảo của các thầy cô nên dự án nghiên cứu của chúng em mới có thể hoàn thiện tốt nhất.

Một lần nữa, em xin chân thành cảm ơn những thầy/cô – người đã trực tiếp giúp đỡ, quan tâm, hướng dẫn chúng em hoàn thành tốt bài báo cáo này trong thời gian qua.

Bước đầu đi vào thực tế của chúng em còn nhiều hạn chế và bất ngờ nên không tránh khỏi những thiếu sót, nhóm chúng em rất mong nhận được những ý kiến đóng góp quý báu của quý thầy/cô để kiến thức của chúng em trong lĩnh vực này được hoàn thiện hơn đồng thời có điều kiện bổ sung, nâng cao ý thức của mình.

Nhóm chúng em xin chân thành cảm ơn!

Tài liệu tham khảo

[1] **Giáo trình thống kê máy tính (Computational Statistics)_TS.Nguyễn Đức Thuần**

[2] **An introduction to Statistical Methods & Data Analysis**

[3] **Video thống kê của thầy Nguyễn Văn Tuấn**

<https://www.youtube.com/channel/UC21dOPe-YHO3Gw6BRbyeotQ>

Bài 3.37: (Quỳnh)

Tính giá trị trung bình, giá trị trung vị và độ lệch chuẩn cho các tỷ lệ sở hữu nhà được đưa ra trong Bài tập 3.10.

- So sánh giá trị trung bình và giá trị trung vị của 3 năm dữ liệu. Giá trị nào, nghĩa là hoặc trung bình, có thích hợp hơn cho các tập dữ liệu này không? Giải thích câu trả lời của bạn.
- So sánh mức độ thay đổi trong tỷ lệ sở hữu nhà trong 3 năm.

Dữ liệu đầu vào:

State	1985	1996	2002	State	1985	1996	2002
Alabama	70.4	71.0	73.5	Montana	66.5	68.6	69.3
Alaska	61.2	62.9	67.3	Nebraska	68.5	66.8	68.4
Arizona	64.7	62.0	65.9	Nevada	57.0	61.1	65.5
Arkansas	66.6	66.6	70.2	New Hampshire	65.5	65.0	69.5
California	54.2	55.0	58.0	New Jersey	62.3	64.6	67.2
Colorado	63.6	64.5	69.1	New Mexico	68.2	67.1	70.3
Connecticut	69.0	69.0	71.6	New York	50.3	52.7	55.0
Delaware	70.3	71.5	75.6	North Carolina	68.0	70.4	70.0
Dist. of Columbia	37.4	40.4	44.1	North Dakota	69.9	68.2	69.5
Florida	67.2	67.1	68.7	Ohio	67.9	69.2	72.0
Georgia	62.7	69.3	71.7	Oklahoma	70.5	68.4	69.4
Hawaii	51.0	50.6	57.4	Oregon	61.5	63.1	66.2
Idaho	71.0	71.4	73.0	Pennsylvania	71.6	71.7	74.0
Illinois	60.6	68.2	70.2	Rhode Island	61.4	56.6	59.6
Indiana	67.6	74.2	75.0	South Carolina	72.0	72.9	77.3
Iowa	69.9	72.8	73.9	South Dakota	67.6	67.8	71.5
Kansas	68.3	67.5	70.2	Tennessee	67.6	68.8	70.1
Kentucky	68.5	73.2	73.5	Texas	60.5	61.8	63.8
Louisiana	70.2	64.9	67.1	Utah	71.5	72.7	72.7
Maine	73.7	76.5	73.9	Vermont	69.5	70.3	70.2
Maryland	65.6	66.9	72.0	Virginia	68.5	68.5	74.3
Massachusetts	60.5	61.7	62.7	Washington	66.8	63.1	67.0
Michigan	70.7	73.3	76.0	West Virginia	75.9	74.3	77.0
Minnesota	70.0	75.4	77.3	Wisconsin	63.8	68.2	72.0
Mississippi	69.6	73.0	74.8	Wyoming	73.2	68.0	72.8
Missouri	69.2	70.2	74.6				

Source: U.S. Bureau of the Census, <http://www.census.gov/ftp/pub/hhes/www/hvs.html>.

Bài làm

Đọc dữ liệu từ file

```
setwd("D:/HK1-3/THONGKEMT/BTN")
getwd()
data<-read.csv("dulieu.csv",header=TRUE)
data
names(data)
#Trung bình mỗi năm
mean(data$nam1985)
mean(data$nam1996)
mean(data$nam2002)
```

Kết quả trình trung bình mỗi năm

```
> names(data)
[1] "State"      "nam1985"    "nam1996"    "nam2002"
> #Trung bình mỗi năm
> mean(data$nam1985)
[1] 65.87647
> mean(data$nam1996)
[1] 66.84314
> mean(data$nam2002)
[1] 69.44902
```

Kết quả tính trung vị mỗi năm

```
> median(data$A)
[1] 67.9
> median(data$B)
[1] 68.2
> median(data$B.1)
[1] 70.2
```

Chúng ta có thể dùng hàm `str()` để xem mối quan hệ giữa các biến với giá trị từng biến

```
> str(data)
```

```
'data.frame': 51 obs. of 4 variables:
 $ State : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...
 $ nam1985: num 70.4 61.2 64.7 66.6 54.2 63.6 69 70.3 37.4 67.2 ...
 $ nam1996: num 71 62.9 62 66.6 55 64.5 69 71.5 40.4 67.1 ...
 $ nam2002: num 73.5 67.3 65.9 70.2 58 69.1 71.6 75.6 44.1 68.7 ...
```

Ngoài hàm `mean()` và hàm `median()` có thể dùng `summary()` để thống kê đơn giản như tìm max, min, giá trị trung bình, tính trung vị, ...

```
> summary(data)
```

State	nam1985	nam1996	nam2002
Length:51	Min. :37.40	Min. :40.40	Min. :44.10
Class :character	1st Qu.:63.15	1st Qu.:64.55	1st Qu.:67.25
Mode :character	Median :67.90	Median :68.20	Median :70.20
	Mean :65.88	Mean :66.84	Mean :69.45
	3rd Qu.:69.95	3rd Qu.:71.20	3rd Qu.:73.50
	Max. :75.90	Max. :76.50	Max. :77.30

- Với tập dữ liệu trên ta nên sử dụng giá trị trung vị(median) bởi vì giá trị trung bình là đại diện cho giá trung tâm của tập hợp, nhưng vì tập dữ liệu trên có chứa các phân tử ngoại lệ làm kéo giá trị trung bình theo hướng của giá trị ngoại lai(phần tử ngoại lệ) để tìm điểm cân bằng, do đó làm sai lệch giá trị trung bình làm thước đo giá trị trung tâm.
- So sánh mức độ biến động của tỷ lệ sở hữu nhà trong 3 năm:
 - Để rõ hơn ta hãy đi tìm độ phân tán của tỷ lệ sở hữu nhà qua các năm. Vì trong R không có hàm để tìm độ phân tán (CV) nên ta phải tự thực hiện thủ tục hàm đó. Từ dữ liệu đã có ta có thể tính được giá trị trung bình và độ lệch chuẩn của từng năm.

```

> CV<-function(x){
+   sd<-sd(x)
+   xn<-mean(x)
+   Nam<-sd/xn*100
+   c(CV=Nam)
+ }
> CV(data$A)
      CV
10.22278
> CV(data$B)
      CV
10.00555
> CV(data$B.1)
      CV
8.873993

```

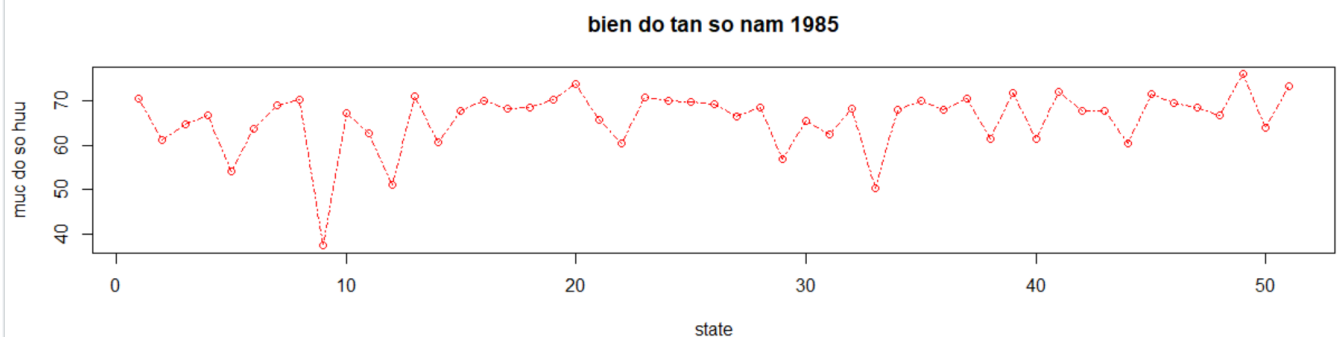
- Ta thấy qua từng năm độ phân tán ngày càng giảm xuống chứng tỏ mức độ biến động của dữ liệu theo chiều hướng giảm dần và điều này chứng tỏ mức độ biến động của tỷ lệ sở hữu nhà là khá lớn trong 3 năm.

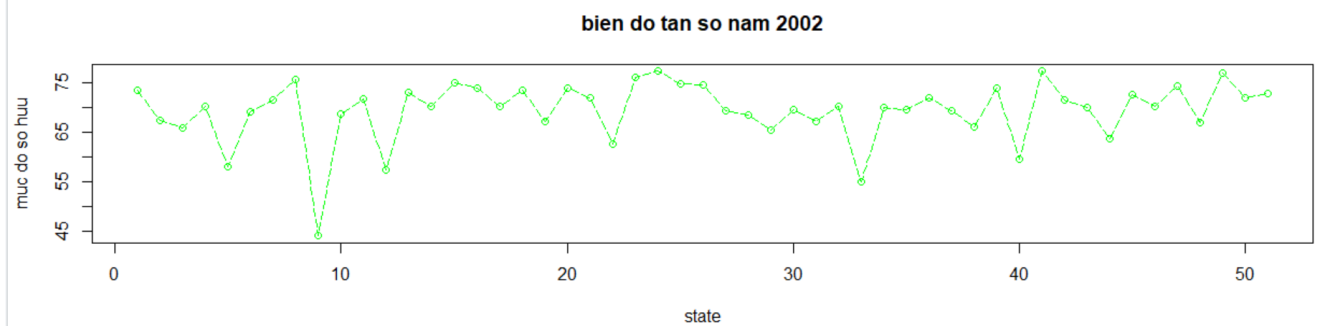
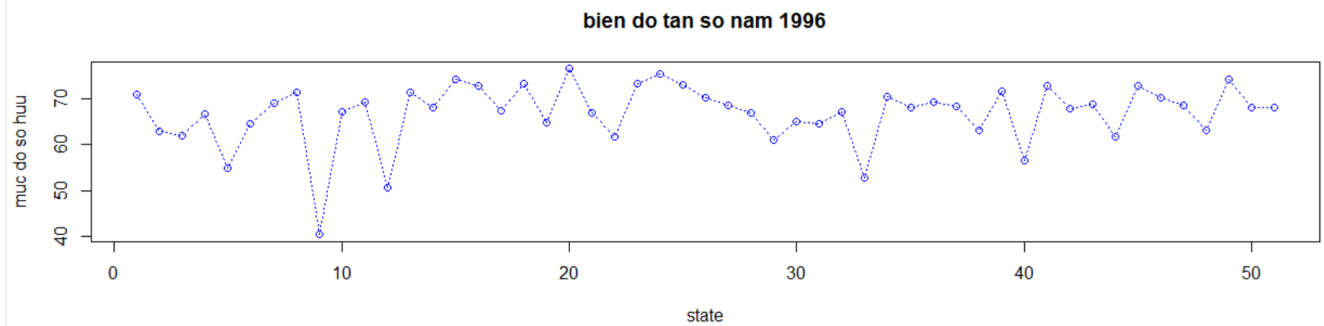
Một số biểu đồ thể hiện sự biến động theo từng năm

```

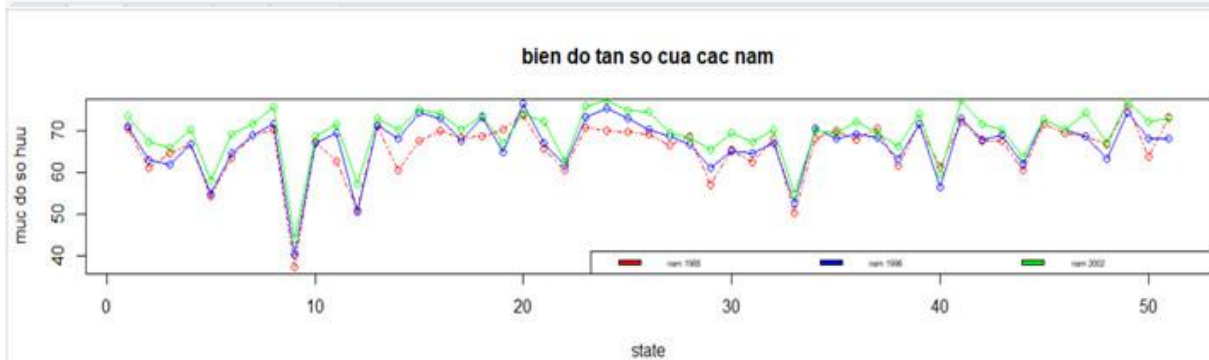
#biểu đồ đường
nam1985=(data$A)
nam1985
plot(nam1985,type="o",col="red",main="bien do tan so nam 1985",xlab="state",ylab="muc do so huu",lty=4)
nam1996=(data$B)
nam1996
plot(nam1996,type="o",col="blue",main="bien do tan so nam 1996",xlab="state",ylab="muc do so huu",lty=3)
nam2002=(data$C)
nam2002
plot(nam2002,type="o",col="green",main="bien do tan so nam 2002",xlab="state",ylab="muc do so huu",lty=5)

```





```
setwd("D:/HK1-3/THONGKEMT/BTN")
getwd()
data<-read.csv("dulieu.csv",header=TRUE)
data
names(data)
nam1985=(data$A)
plot(nam1985,type="o",col="red",main="bien do tan so cua cac nam",xlab="state",ylab="muc do so huu",lty=4)
nam1996=(data$B)
lines(nam1996,type="o",col="blue")
nam2002=(data$C)
lines(nam2002,type="o",col="green")
legend("bottomright", c("nam 1985", "nam 1996", "nam 2002"),fill=c("red","blue","green"),horiz=TRUE,cex=0.5)
```



Bài 3.38: (M.Tâm, Quỳnh)

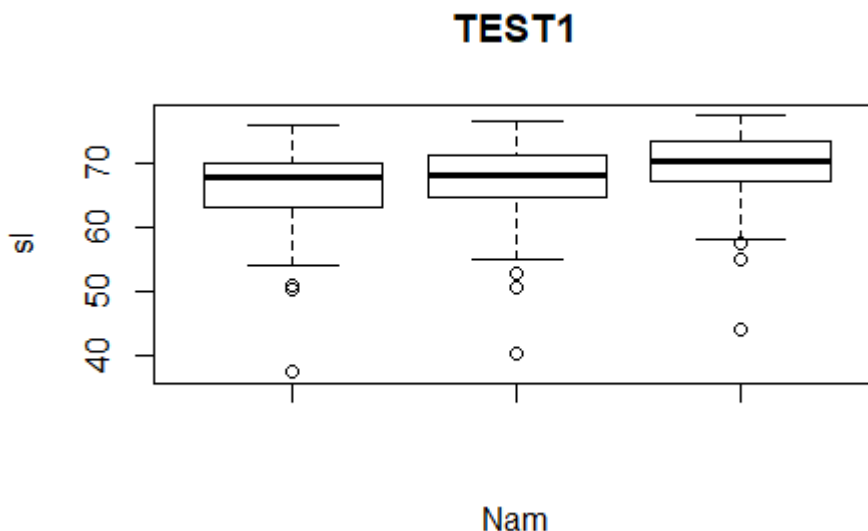
Đối với các ô vuông được xây dựng cho các tỷ lệ sở hữu nhà được đưa ra trong Bài tập 3.36, hãy đặt ba ô hộp trên cùng một bộ trục

a. Sử dụng boxplot cạnh nhau này để thảo luận về những thay đổi về tỷ lệ sở hữu nhà trung bình trong 3 năm.

- b. Sử dụng boxplot cạnh nhau này để thảo luận về những thay đổi trong sự thay đổi trong các tỷ lệ này trong 3 năm.
- c. Có tiểu bang nào có tỷ lệ sở hữu nhà cực kỳ thấp không?
- d. Có tiểu bang nào có tỷ lệ sở hữu nhà cực kỳ cao không?

Bài làm

```
> setwd("D:/HK1(2021)_Nam3/ThongKeMayTinh")
> getwd()
[1] "D:/HK1(2021)_Nam3/ThongKeMayTinh"
> data<-read.csv("dulieu.csv")
> boxplot(data$A, data$B, data$B.1, xlab = "Nam",
+         ylab = "sl", type="l", main="TEST1")
```



- a) Qua sơ đồ ta có thể thấy giá trị trung vị của dữ liệu tăng dần qua từng năm từ 67,9 lên 68,2 và đến 70.2.
- b) Tỷ lệ số người sở hữu nhà ở mức phân vị 75%(Q3) cũng tăng dần qua các năm từ 69.95 lên 71.2 và cuối cùng là 73.5
- c) Dựa vào biểu đồ ta thấy giá trị tỷ lệ thấp nhất là của năm đầu tiên 37.40 của Dist.of Columbia
- d) Giá trị tỷ lệ cao nhất là năm thứ ba 77.3 của SouthCarolina

Note: Có thể dùng summary () để thống kê đơn giản như tìm max, min, giá trị trung bình, tính trung vị, ...

```

> summary(data)
  State      nam1985      nam1996      nam2002
Length:51    Min.    :37.40    Min.    :40.40    Min.    :44.10
Class :character 1st Qu.:63.15    1st Qu.:64.55    1st Qu.:67.25
Mode  :character Median :67.90    Median :68.20    Median :70.20
              Mean  :65.88    Mean  :66.84    Mean  :69.45
              3rd Qu.:69.95    3rd Qu.:71.20    3rd Qu.:73.50
              Max.   :75.90    Max.   :76.50    Max.   :77.30

> # Nam 1985
> kq<-summary(data$A)
> Q3<-kq[5]
> Q3
3rd Qu.
  69.95

> # Nam 1996
> kq<-summary(data$B)
> Q3<-kq[5]
> Q3
3rd Qu.
  71.2

> # Nam 2002
> kq<-summary(data$B.1)
> Q3<-kq[5]
> Q3
3rd Qu.
  73.5

```

Bài 3.39: (Quân)

Trong tờ " *Demographic Implications of Socioeconomic Transition Among the Tribal Populations of Manipur, India* " [Human Biology (1998) 70(3):597–619]", tác giả đã mô tả lại sự thay đổi lớn diễn ra ở quy mô toàn bộ Manipur, Ấn Độ, bắt đầu từ thế kỉ 20 trở đi.

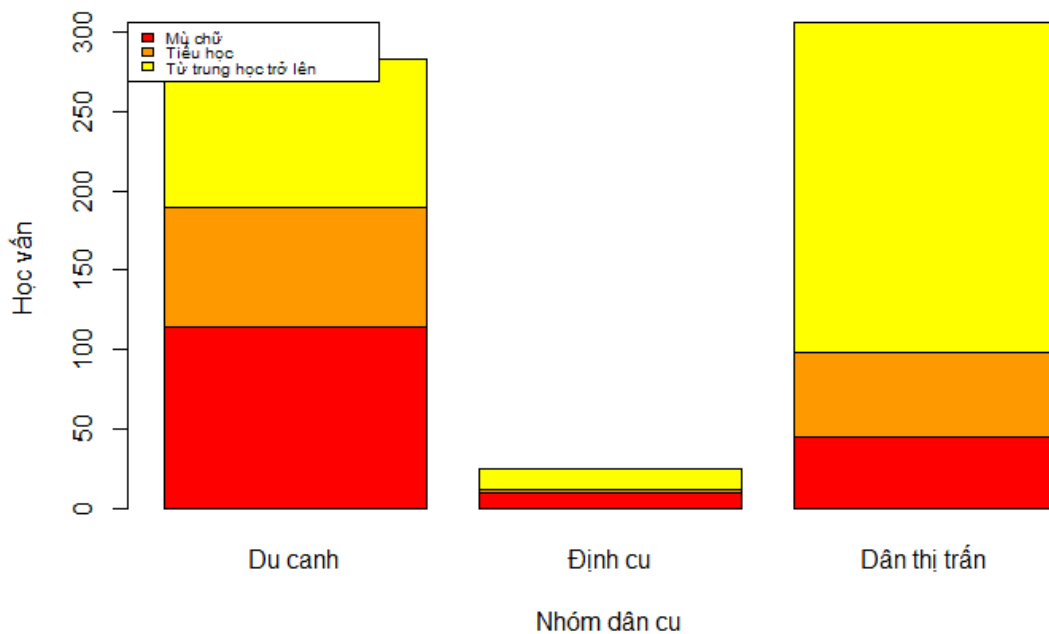
Người dân đang có xu hướng chuyển đổi từ nền kinh tế truyền thống tự cung tự cấp sang nền kinh tế định hướng thị trường. Bảng dưới đây thể hiện mối quan hệ giữa trình độ học vấn với nhóm người sinh sống với mẫu thử là: 614 người đã kết hôn.

Trình độ học vấn

Nhóm sinh hoạt	Tiểu học	Ít nhất là Trung học	Mù chữ
Du canh du cư(người dân thiểu số)	10	45	114
NN định cư	2	53	76
Cư dân thị trấn	13	208	93

Bài làm

a. Biểu diễn dữ liệu dạng biểu đồ cột chồng



Code:

```
a<-matrix(c(114,10,45,76,2,53,93,13,208),nrow=3,ncol=3,byrow=T)
```

```
a
      [,1] [,2] [,3]
1,]  114   10   45
2,]   76    2   53
3,]   93   13  208
```

```
data.frame(a)
  x1 x2 x3
114 10 45
 76  2 53
 93 13 208
```

```
barplot(a,xlab = "Nhóm dân cư",ylab = "Học vấn", names.arg=c("Du canh","Định cư","Dân thị trấn"),col = heat.colors(3))
legend("topleft",c("Mù chữ","Tiểu học","Tù trung học trở lên"),fill=colors(3),cex=0.6)
legend("topleft",c("Mù chữ","Tiểu học","Tù trung học trở lên"),fill=heat.colors(3),cex=0.6)
```

b. So sánh tỉ lệ phần trăm dựa trên hàng và cột, ta thấy

+) Tính phần trăm từng thành phần = <thành phần>/(tổng*100)

->Theo đó ta có tỉ lệ % về trình độ học vấn khu vực Manipur nói chung:

$$\text{Tỷ lệ mù chữ} = (114+96+73) / (614*100) \approx 46,1\%$$

$$\text{Tỷ lệ qua tiểu học} = (10 + 12 + 13) / (614*100) \approx 4,1\%$$

$$\text{Trung học trở lên} = (45 + 53 + 208) / (614*100) \approx 49,8\%$$

->Tỉ lệ thành phần nhóm người khu vực Manipur nói chung:

$$\text{Du canh/cư: } 27,5\%$$

$$\text{NN định cư: } 21,4\%$$

$$\text{Cư dân thị trấn: } 51,1\%$$

=> Qua mẫu, và biểu đồ, ta thấy tỉ lệ % người mù chữ so với người có trình độ khá là xấp xỉ ngang bằng nhau, trong đó nhóm người du canh có trình độ học vấn thấp nhất so với hai thành phần nhóm người còn lại.

Code:

```

> T<-sum(data$TieuHoc)+sum(data$LHTH)+sum(data$MC)
> T
[1] 614
> TLMC<-(sum(data$MC)/T)*100
> TLMC
[1] 46.09121
> QTH<-(sum(data$TieuHoc)/T)*100
> QTH
[1] 4.071661
> TH<-(sum(data$LHTH)/T)*100
> TH
[1] 49.83713

```

Bài 3.40: (Phi)

Trong sản xuất kính áp tròng mềm, công suất (độ mạnh) của thấu kính cần phải rất gần với giá trị mục tiêu. Trong bài báo “Kiểm tra loại ANOM đối với sự khác biệt so với dân số bình thường” [Technometrics (1997) 39: 274–283], một số nhà cung cấp được thực hiện so sánh về tính nhất quán của công suất của ống kính. Bảng sau đây chứa độ lệch so với giá trị công suất mục tiêu của thấu kính được sản xuất bằng vật liệu từ ba nhà cung cấp khác nhau:

NSX	Sai lệch với giá trị công suất mục tiêu								
1	189.9	191.9	190.9	183.8	185.5	190.9	192.8	188.4	189.0
2	156.6	158.4	157.7	154.1	152.3	161.5	158.1	150.9	156.9
3	218.6	208.4	187.1	199.5	202.0	211.1	197.6	204.4	206.8

- Tính giá trị trung bình mẫu và độ lệch chuẩn mẫu?
- Vẽ đồ thị dữ liệu độ lệch mẫu?
- Mô tả sự sai lệch so với quyền lực được chỉ định cho ba nhà cung cấp?
- Nhà cung cấp nào dường như cung cấp vật liệu sản xuất thấu kính có công suất gần nhất với giá trị mục tiêu?

Bài làm

- Tính giá trị trung bình mẫu và độ lệch chuẩn mẫu?
- **Đối với NSX1:**

$$\text{Trung bình mẫu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{(189,9 + 191,9 + \dots + 189,0)}{9} = 189,133$$

$$\text{Phương sai} = s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)} = \frac{((189,9-189,133)^2 + (191,9-189,133)^2 + \dots + 189,0-189,133)^2}{(9-1)} = 8,245$$

$$\text{Độ lệch chuẩn mẫu } s = \sqrt{s^2} = 2,8713$$

- **Đối với NSX2**, tính tương tự như công thức bên trên ta có:

Trung bình mẫu = 156,277

Phương sai = 10,886

Độ lệch chuẩn mẫu = 3,299

- **Đối với NSX3:**

Trung bình mẫu = 203,944

Phương sai = 80,215

Độ lệch chuẩn mẫu = 8,956

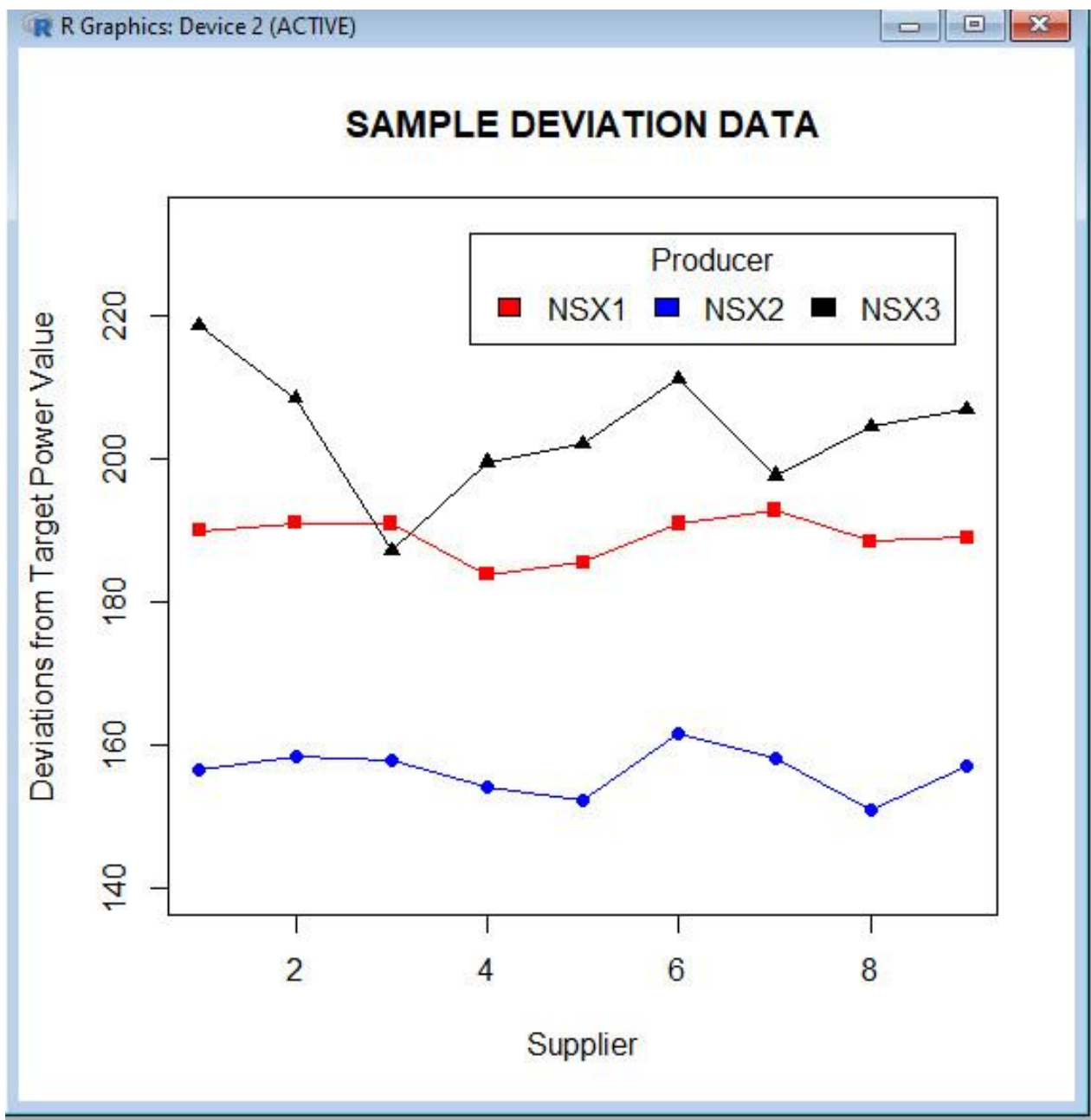
Thể hiện bằng ngôn ngữ R:

```
R Console
> nsx1<-c(189.9,191.0,190.9,183.8,185.5,190.9,192.8,188.4,189.0)
> mean(nsx1)
[1] 189.1333
> #Độ lệch chuẩn mẫu có thể tính trực tiếp trong R mà không cần thông qua phương sai
> sd(nsx1)
[1] 2.871411
> nsx2<-c(156.6,158.4,157.7,154.1,152.3,161.5,158.1,150.9,156.9)
> mean(nsx2)
[1] 156.2778
> sd(nsx2)
[1] 3.299537
> nsx3<-c(218.6,208.4,187.1,199.5,202.0,211.1,197.6,204.4,206.8)
> mean(nsx3)
[1] 203.9444
> sd(nsx3)
[1] 8.956298
> |
```

- b. Vẽ đồ thị dữ liệu độ lệch mẫu?

Thể hiện bằng ngôn ngữ R

```
> nsx1<-c(189.9,191.0,190.9,183.8,185.5,190.9,192.8,188.4,189.0)
> nsx2<-c(156.6,158.4,157.7,154.1,152.3,161.5,158.1,150.9,156.9)
> nsx3<-c(218.6,208.4,187.1,199.5,202.0,211.1,197.6,204.4,206.8)
> plot(nsx1,type="o",pch=15,col="red",xlab="Deviations from Target Power Value",ylab="Supplier",main="sample deviation data",ylim=c(140,max(nsx1)+40))
> lines(nsx2,type="o",pch=16,col="blue")
> lines(nsx3,type="o",pch=17,col="violet")
> |
```



c. Mô tả sự sai lệch so với quyền lực được chỉ định cho ba nhà cung cấp?

Nhìn chung thì sự sai lệch so với công suất mục tiêu của các nhà cung cấp là khá lớn. Cụ thể là từ 150 đến 219 đơn vị công suất.

d. Nhà cung cấp nào dường như cung cấp vật liệu sản xuất thấu kính có công suất gần nhất với giá trị mục tiêu?

Dựa vào biểu đồ thể hiện sự sai lệch với công suất mục tiêu giữa 3 nhà cung cấp, ta thấy nhà cung cấp thứ 1 có ít sự sai lệch nhất, vậy có thể nói nhà cung cấp thứ 1 có công suất gần với giá trị mục tiêu nhất.

Bài 3.41: (M.Phúc)

Bài làm

Chính phủ liên bang theo dõi chặt chẽ sự tăng trưởng tiền so với các mục tiêu đã được đặt ra cho sự tăng trưởng đó. Chúng tôi liệt kê hai thước đo cung tiền ở Hoa Kỳ, M2 (tiền gửi séc tư nhân, tiền mặt và một số khoản tiết kiệm) và M3 (M2 cộng với một số khoản đầu tư), được đưa ra ở đây trong 20 tháng liên tiếp.

Month	Money Supply (in trillions of dollars)		Month	Money Supply (in trillions of dollars)	
	M2	M3		M2	M3
1	2.25	2.81	11	2.43	3.05
2	2.27	2.84	12	2.42	3.05
3	2.28	2.86	13	2.44	3.08
4	2.29	2.88	14	2.47	3.10
5	2.31	2.90	15	2.49	3.10
6	2.32	2.92	16	2.51	3.13
7	2.35	2.96	17	2.53	3.17
8	2.37	2.99	18	2.53	3.18
9	2.40	3.02	19	2.54	3.19
10	2.42	3.04	20	2.55	3.20

- a. Biểu đồ tán xạ có mô tả mối quan hệ giữa M2 và M3 không?

Biểu đồ tán xạ thể hiện mối quan hệ giữa M2 và M3, chính phủ liên bang muốn xác định những thay đổi của M2 và M3 trong khoảng thời gian 20 tháng

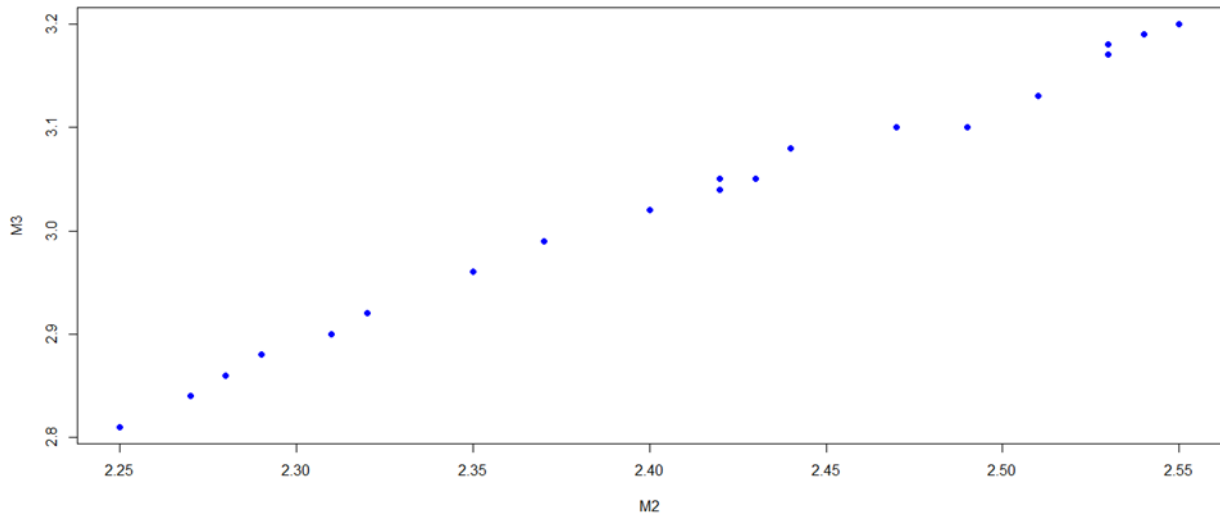
- b. Xây dựng biểu đồ phân tán. Có mối quan hệ rõ ràng?

Nhập dữ liệu:

	A	B	C
1	Month	M2	M3
2	1	2.25	2.81
3	2	2.27	2.84
4	3	2.28	2.86
5	4	2.29	2.88
6	5	2.31	2.9
7	6	2.32	2.92
8	7	2.35	2.96
9	8	2.37	2.99
10	9	2.4	3.02
11	10	2.42	3.04
12	11	2.43	3.05
13	12	2.42	3.05
14	13	2.44	3.08
15	14	2.47	3.1
16	15	2.49	3.1
17	16	2.51	3.13
18	17	2.53	3.17
19	18	2.53	3.18
20	19	2.54	3.19
21	20	2.55	3.2

Code:

```
R Console
> dl<-read.csv("Bai3_41.csv")
> dl
  Month  M2  M3
1     1 2.25 2.81
2     2 2.27 2.84
3     3 2.28 2.86
4     4 2.29 2.88
5     5 2.31 2.90
6     6 2.32 2.92
7     7 2.35 2.96
8     8 2.37 2.99
9     9 2.40 3.02
10    10 2.42 3.04
11    11 2.43 3.05
12    12 2.42 3.05
13    13 2.44 3.08
14    14 2.47 3.10
15    15 2.49 3.10
16    16 2.51 3.13
17    17 2.53 3.17
18    18 2.53 3.18
19    19 2.54 3.19
20    20 2.55 3.20
> plot(dl$M2,dl$M3,xlab="M2",ylab="M3",col="blue",pch=16)
> |
```



Dựa vào biểu đồ tán xạ trên ta thấy hai thước đo tuân theo mối quan hệ tuyến tính tăng dần.

Bài 3.42: (Son)

Tham khảo bài tập 3.41. Biểu đồ dữ liệu nào khác có thể được sử dụng để mô tả và tóm tắt những dữ liệu này? Lập cốt truyện và giải thích kết quả của bạn.

Bài làm

Biểu đồ thanh (barplot):

```
> a<-c (2.25,2.27,2.28,2.29,2.31,2.32,2.35,2.37,2.4,2.42)
```

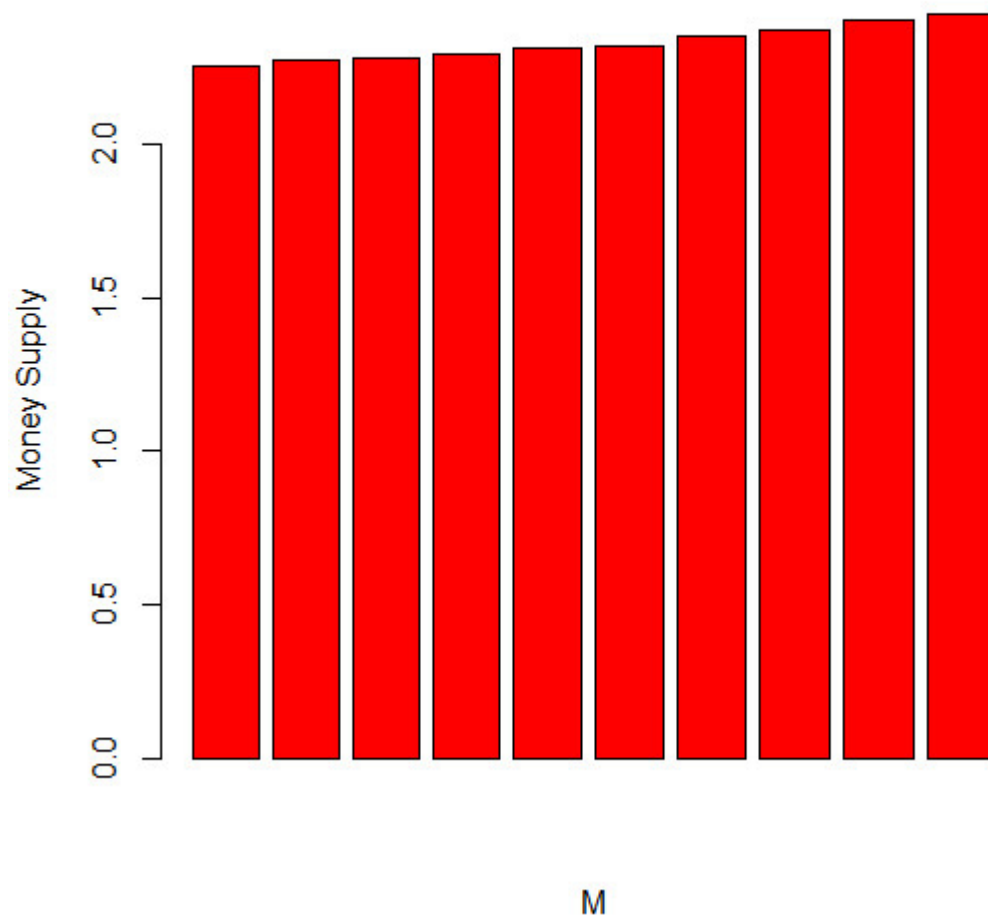
```
> b<-c (2.81,2.84,2.86,2.88,2.9,2.92,2.96,2.99,3.02,3.04)
```

```
barplot(a,b,xlab="M",ylab="Money Supply")
```

R Console

```
> a<-c(2.25,2.27,2.28,2.29,2.31,2.32,2.35,2.37,2.4,2.42)
> b<-c(2.81,2.84,2.86,2.88,2.9,2.92,2.96,2.99,3.02,3.04)
> barplot(a~b,xlab="M",ylab="Money Supply",col="green")
> barplot(a,b,xlab="M",ylab="Money Supply",col="red")
> |
```

R Graphics: Device 2 (ACTIVE)



→ Không hiệu quả

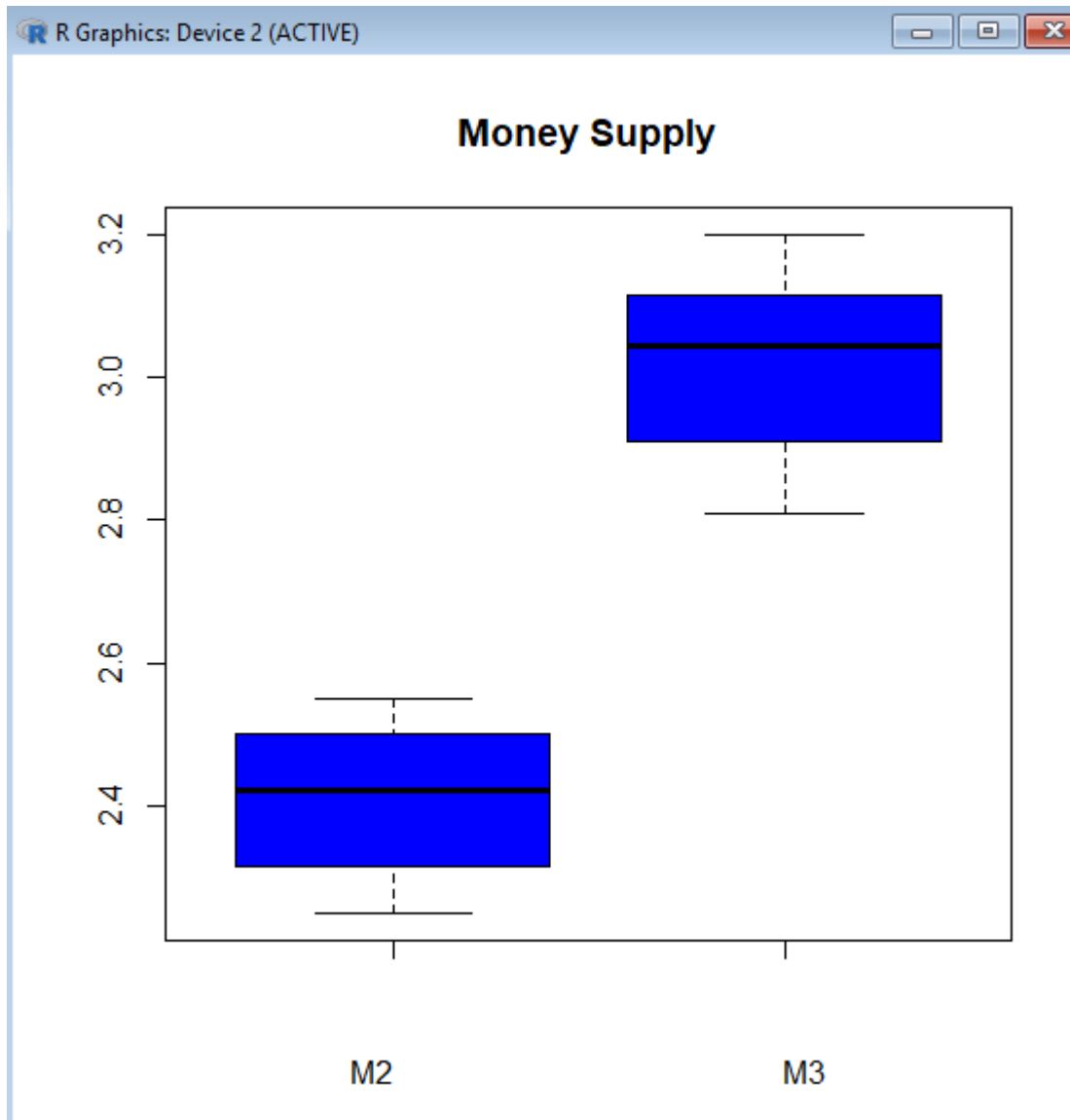
Biểu đồ dạng hộp

```
a<- c(2.25,2.27,2.28,2.29,2.31,2.32,2.35,2.37,2.4,
+2.42,2.43,2.42,2.44,2.47,2.49,2.51,2.53,2.53,2.54,2.55)
> b<-c(2.81,2.84,2.86,2.88,2.9,2.92,2.96,2.99,
```

+3.02,3.04,3.05,3.05,3.08,3.1,3.1,3.13,3.17,3.18,3.19,3.20)

Boxplot (a,b,xlab="M2 M3",main="Money Supply",col="blue")

```
R Console
> a<-c(2.25,2.27,2.28,2.29,2.31,2.32,2.35,2.37,2.4,2.42,2.43,2.42,2.44,2.47,2.49,2.51,2.53,2.53,2.54,2.55)
> b<-c(2.81,2.84,2.86,2.88,2.9,2.92,2.96,2.99,3.02,3.04,3.05,3.05,3.08,3.1,3.1,3.13,3.17,3.18,3.19,3.20)
> boxplot(a,b,xlab="M2 M3",main="Money Supply",col="blue")
> |
```



→ Hiệu quả hơn Biểu Đồ Thanh

+ Biểu đồ tròn (pie, pie3D): không hợp lệ vì trong chia thành 2 bảng và không thể hiện được mối quan hệ của M2 và M3.

+ Biểu đồ đường: không thể vẽ được vì biên số quá lớn

Bài 3.43: (V.Tâm)

Để kiểm soát rủi ro hư hỏng lõi nghiêm trọng trong một nhà máy điện hạt nhân thương mại tai nạn mất điện, độ tin cậy của máy phát điện diesel khẩn cấp khi khởi động theo yêu cầu phải được duy trì ở mức cao. Bài báo “Ước tính Bayes theo kinh nghiệm về độ tin cậy của máy phát điện diesel khẩn cấp năng lượng hạt nhân” chứa dữ liệu về lịch sử thất bại của bảy nhà máy điện hạt nhân. Dữ liệu sau đây là số lượng thành công nhu cầu giữa những lần hỏng hóc đối với máy phát điện diesel tại một trong những nhà máy này từ năm 1982 đến năm 1988.

(Lưu ý: sự cố của máy phát điện diesel không nhất thiết dẫn đến hư hỏng lõi hạt nhân vì tất cả các nhà máy điện hạt nhân đều có một số máy phát điện diesel khẩn cấp.)

Bài làm

- a. Tính giá trị trung bình và trung vị của các nhu cầu thành công giữa các lần thất bại.

Trung bình = $(28+50+193+55+\dots+26+6)/34=57.5294117$

me

Trung vị: sx dữ liệu tăng dần

0 0 0 1 4 4 4 6 7 7 9 10 10 15 26 26 28 40 41
46 50 54 55 55 62 76 84 105 128 147 164 193 226 273

Trung vị = $(28+40)/2=34$

R:

```
> dl
[1] 28 50 193 55 4 7 147 76 10 0 10 84 0 9 1 0 62 26 15
[20] 226 54 46 128 4 105 40 4 273 164 7 55 41 26 6
> mean(dl)
[1] 57.52941
> median(dl)
[1] 34
```

- b. Thước đo nào thể hiện tốt nhất trung tâm của dữ liệu?

Trung vị làm thước đo cho dữ liệu, vì dữ liệu có chênh lệch giá trị lớn (min=0, max=273), dùng trung bình để đo lường không chính xác về xu hướng tập trung dữ liệu.

- c. Tính phạm vi và độ lệch chuẩn, s.

Độ lệch chuẩn $s=70.1955$

R:

```
> dl
[1] 28 50 193 55 4 7 147 76 10 0 10 84 0 9 1 0 62 26 15
[20] 226 54 46 128 4 105 40 4 273 164 7 55 41 26 6
> sd(dl)
[1] 70.1955
```

d. Sử dụng xấp xỉ phạm vi để ước tính s. Khoảng cách gần đúng với giá trị đích thực?

Sử dụng xấp xỉ phạm vi để ước tính s: $s_{ut} \approx (\max - \min) / 4 = (273 - 0) / 4 = 68.25$

Độ lệch chuẩn ban đầu $s_{dt} = 70.1955$

=> Ta thấy được độ lệch chuẩn khi đo bằng sắp xỉ phạm vi có giá trị gần đúng với độ lệch chuẩn đích thực ($s_{ut} \approx s_{dt}$).

$$s = (\max - \min) / \sqrt{n}, \text{ nếu } n < 12$$

$$s = (\max - \min) / 4, \text{ nếu } 20 < n < 40$$

$$s = (\max - \min) / 5, \text{ nếu } n \approx 100$$

$$s = (\max - \min) / 6, \text{ nếu } n > 400$$

e. Xây dựng các khoảng

$$\bar{y} \pm s \quad \bar{y} \pm 2s \quad \bar{y} \pm 3s$$

Đếm số lần yêu cầu giữa các lần không đạt trong mỗi khoảng thời gian trong ba khoảng thời gian. Chuyển đổi những con số này thành tỷ lệ phần trăm và so sánh kết quả của bạn với Quy tắc thực nghiệm.

Đếm số lần yêu cầu giữa các lần không đạt trong mỗi khoảng thời gian trong ba khoảng thời gian. Chuyển đổi những con số này thành tỷ lệ phần trăm và so sánh kết quả của bạn với Quy tắc thực nghiệm.

$$\bar{y} \pm s = 28 (82\%) \neq 68\%$$

```

test<-function(x){
  print("Khoang thứ 1 68%: ")
  dem<-0
  gt<-mean(x) + 1*sd(x)
  gd<-mean(x) - 1*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}

[1] "Khoang thứ 1 68%: "
[1] 28
[1] -12.66609
[1] 127.7249
> (28/length(d1))*100
[1] 82.35294

```

$\bar{y} \pm 2s = 32$ (94%) ~ 95%

```

test<-function(x){
  print("Khoang thứ 2 95%: ")
  dem<-0
  gt<-mean(x) + 2*sd(x)
  gd<-mean(x) - 2*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}

```

```

[1] "Khoang thứ 2 95%: "
[1] 32
[1] -82.86159
[1] 197.9204
> (32/length(d1))*100
[1] 94.11765
 $\bar{y} \pm 3s = 33 (97\%) \sim 99,7\%$ 
test<-function(x){
  print("Khoang thứ 3 99.7%: ")
  dem<-0
  gt<-mean(x) + 3*sd(x)
  gd<-mean(x) - 3*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}
[1] "Khoang thứ 3 99.7%: "
[1] 33
[1] -153.0571
[1] 268.1159
> (33/length(d1))*100
[1] 97.05882

```

=> kết quả khác với quy tắc chuẩn

f. Tại sao bạn cho rằng Quy tắc thực nghiệm và tỷ lệ phần trăm của bạn không khớp nhau?

$Y \pm s \Rightarrow 68\% \rightarrow 82\%$

Ta thấy ở mức 68% thì kết quả không trùng khớp với Quy tắc thực nghiệm.

Bài 3.44: (Quyền)

Trường Cao đẳng Nha khoa tại Đại học Florida đã cam kết phát triển toàn bộ chương trình giảng dạy của mình xung quanh việc sử dụng các tài liệu giảng dạy tự nhíp như

băng video, băng trượt và giáo trình. Hy vọng rằng mỗi sinh viên sẽ tiến hành với một tốc độ tương xứng với khả năng của mình và nhân viên giảng dạy sẽ có nhiều thời gian rảnh hơn để tư vấn cá nhân trong tương tác sinh viên - giảng viên. Một mô-đun giảng dạy như vậy đã được phát triển và thử nghiệm trên 50 sinh viên đầu tiên tiến hành thông qua chương trình giảng dạy. Các phép đo sau đây đại diện cho số giờ mà những sinh viên này phải mất để hoàn thành vật liệu mô-đun cần thiết.

16	8	33	21	34	17	12	14	27	6
33	25	16	7	15	18	25	29	19	27
s5	12	29	22	14	25	21	17	9	4
12	15	13	11	6	9	26	5	16	5
9	11	5	4	5	23	21	10	17	15

a. Tính toán chế độ, trung vị và trung bình cho các thời gian hoàn thành được ghi lại này.

b. Đoán giá trị của s.

c. Tính toán bằng cách sử dụng công thức phím tắt và so sánh câu trả lời của bạn với câu trả lời của phần (b).

d. Bạn có mong đợi Quy tắc thực nghiệm mô tả đầy đủ sự thay đổi của các dữ liệu này không? Giải thích.

Bài làm

- a. Gọi X là biến ngẫu nhiên chỉ số giờ mà sinh viên hoàn thành vật liệu mô-đun cần thiết, ta có bảng phân phối xác suất cho biến ngẫu nhiên X như sau:

$X=x_i$	Tần số n_i	Xác suất P_i
4	2	0.04
5	5	0.1
6	2	0.04
7	1	0.02
8	1	0.02

9	3	0.06
10	1	0.02
11	2	0.04
12	3	0.06
13	1	0.02
14	2	0.04
15	3	0.06
16	3	0.06
17	3	0.06
18	1	0.02
19	1	0.02
21	3	0.06
22	1	0.02
23	1	0.02
25	3	0.06
26	1	0.02
27	2	0.04
29	2	0.04
33	2	0.04
34	1	0.02
	50	1.00

Kì vọng: $E(X) = \sum_{i=1}^{25} x_i P_i = 4 \times 0.04 + 5 \times 0.1 + 6 \times 0.04 + \dots + 33 \times 0.04 + 34 \times 0.02 = 15.96$

Số trung vị: 15

b. Đoán giá trị của s

$$E(X^2) = \sum_{i=1}^{25} x_i^2 P_i = 4^2 \times 0.04 + 5^2 \times 0.1 + 6^2 \times 0.04 + \dots + 33^2 \times 0.04 + 34^2 \times 0.02 = 325.52$$

Phương sai: $VAR(X) = E(X^2) - [E(X)]^2 = 325.52 - 15.96^2 = 70.7984$

Độ lệch chuẩn: $s = \sqrt{VAR(X)} = \sqrt{70.7984} = 8.414179$

c.

```
> setwd("/Users/Quyen/Desktop/TranCongQuyen_TKMT")
> Cau_44<-read.csv("Cau_44.csv")
> library(distrEx)
> X<-DiscreteDistribution(Cau_44$X, Cau_44$XS)
> E(X);var(X);sd(X)
[1] 15.96
[1] 70.7984
[1] 8.414179
> |
```

d. - Chọn khoảng:

+ y-s; y+s: 60% ~ 68%

```
test<-function(x){
  print("Khoang thứ 1 68%: ")
  dem<-0
  gt<-mean(x) + 1*sd(x)
  gd<-mean(x) - 1*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}
```

```

[1] 15
[1] 8.21605
[1] 25.70395
> (15/length(data$ĩ..X))*100
[1] 60

```

+ y-2s; y+2s: 100% khác 95%

```

test<-function(x){
  print("Khoang thứ 3 99.7%: ")
  dem<-0
  gt<-mean(x) + 2*sd(x)
  gd<-mean(x) - 2*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}

```

```

[1] 25
[1] -0.5279006
[1] 34.4479
> (25/length(data$ĩ..X))*100
[1] 100

```

+ y-3s; y+3s: 100% ~ 99,7%

```

test<-function(x){
  print("Khoang thứ 3 99.7%: ")
  dem<-0
  gt<-mean(x) + 3*sd(x)
  gd<-mean(x) - 3*sd(x)
  for(i in 1:length(x)){
    if(x[i] <= gt && x[i] >= gd){
      dem<-dem+1
    }
  }
  print(dem)
  print(gd)
  print(gt)
}

[1] 25
[1] -9.271851
[1] 43.19185
> (25/length(data$X))*100
[1] 100

```

Bài 3.45: (P.Tâm)

Số tháng 2 năm 1998 của Báo cáo Người tiêu dùng cung cấp dữ liệu về giá của 24 nhãn hiệu khăn giấy. Giá được đưa ra theo cả chi phí trên mỗi cuộn và giá mỗi tờ vì các thương hiệu đã số lượng tờ mỗi cuộn khác nhau.

Brand	Price per Roll	Number of Sheets per Roll	Cost per Sheet
1	1.59	50	.0318
2	0.89	55	.0162
3	0.97	64	.0152
4	1.49	96	.0155
5	1.56	90	.0173
6	0.84	60	.0140
7	0.79	52	.0152
8	0.75	72	.0104
9	0.72	80	.0090
10	0.53	52	.0102
11	0.59	85	.0069
12	0.89	80	.0111
13	0.67	85	.0079
14	0.66	80	.0083
15	0.59	80	.0074
16	0.76	80	.0095
17	0.85	85	.0100
18	0.59	85	.0069
19	0.57	78	.0073
20	1.78	180	.0099
21	1.98	180	.0100
22	0.67	100	.0067
23	0.79	100	.0079
24	0.55	90	.0061

- Tính toán độ lệch chuẩn cho cả giá mỗi cuộn và giá mỗi tờ giấy.
- Cái nào thay đổi nhiều hơn, giá mỗi cuộn hay giá mỗi tờ?

Bài làm

a. Tính toán độ lệch chuẩn cho cả giá mỗi cuộn và giá mỗi tờ giấy.

Công thức :

Tính trung bình cộng:

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

Trung bình cộng mỗi cuộn

$$\bar{x} = \frac{(1.59+0.89+\dots+0.55)}{24} = 0.9195833$$

Trung bình cộng mỗi tờ

$$\bar{x} = \frac{(0,0318+0,0162+\dots+0,0061)}{24} = 0.01465417$$

Tính phương sai mỗi cuộn:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n x_i^2 - \frac{n(\bar{x})^2}{n-1}$$

$$s^2 = \frac{(1.59^2 + 0.89^2 + \dots + 0.55^3) - 24(0.9195833)^2}{23} = 0.1791868$$

Tính phương sai mỗi tờ :

$$s^2 = \frac{(0,0318^2 + 0,0162^2 + \dots + 0,0162^2) - 24(0.01465417)^2}{23} = 0.0002878339$$

Bảng dữ liệu khi nhập vào exel có dạng:

Brand	Price per Roll	Number of Sheet per Roll	Cost per Sheet
1	1,59	50	0,0318
2	0,89	55	0,0162
3	0,97	64	0,0152
4	1,49	96	0,0155
5	1,56	90	0,0173
6	0,84	60	0,014
7	0,79	52	0,0152
8	0,75	72	0,0104
9	0,72	80	0,09
10	0,53	52	0,0102
11	0,59	85	0,0069
12	0,89	80	0,0111
13	0,67	85	0,0079
14	0,66	80	0,0083
15	0,59	80	0,0074
16	0,76	80	0,0095
17	0,85	85	0,01
18	0,59	85	0,0069
19	0,57	78	0,0073
20	1,78	180	0,0099
21	1,98	180	0,01
22	0,67	100	0,0067
23	0,79	100	0,0079
24	0,55	90	0,0061

Câu lệnh R

```
> install.packages("readxl")
```

```
> library(readxl)
```

```
> r<-read_excel("D:\\R\\b.xlsx",sheet=1)
```

```
> r
```


	Brand	Price per Roll	Number of Sheet per Roll	Cost per Sheet
1	1	1.59	50	0.0318
2	2	0.89	55	0.0162
3	3	0.97	64	0.0152
4	4	1.49	96	0.0155
5	5	1.56	90	0.0173
6	6	0.84	60	0.014
7	7	0.79	52	0.0152
8	8	0.75	72	0.0104
9	9	0.72	80	0.09
10	10	0.53	52	0.0102

... with 14 more rows

```
> s<-var(r$'Price per Roll')
> s
[1] 0.1791868
> alphas<-sqrt(s)
> alphas
[1] 0.4233046
```

```
> s2<-var(r$'Cost per Sheet')
> s2
[1] 0.0002878339
> alpha2<-sqrt(s2)
> alpha2
[1] 0.01696567
```

b.Cái nào thay đổi nhiều hơn, giá mỗi cuộn hay giá mỗi tờ?

” Giá mỗi cuộn” thay đổi nhiều hơn.

c.Trong so sánh của bạn ở phần (b), bạn nên sử dụng s hay CV? Biện minh cho câu trả lời của bạn

```
> X1n<-mean(r$'Price per Roll')
> X2n<-mean(r$'Cost per Sheet')
> CV1<-s/X1n*100
> CV2<-s2/X2n*100
> CV1
[1] 19.48565
> CV2
[1] 1.964178
> |
```

Theo so sánh thay đổi nên sử dụng CV vì CV cung cấp tỉ suất sai số của phương sai và giá trị trung bình

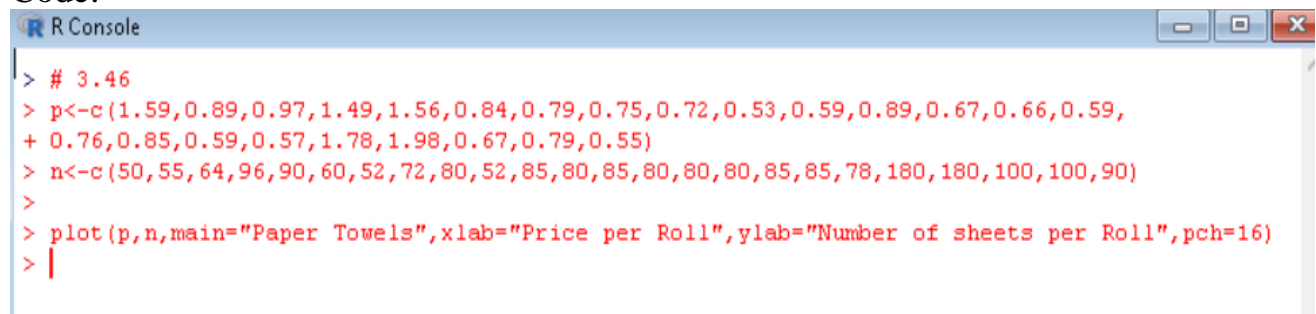
Bài 3.46: (H.Phúc)

Dựa vào bài 3.45, sử dụng biểu đồ tán xạ (scatter plot) để vẽ biểu đồ giá mỗi cuộn và số tờ mỗi cuộn.

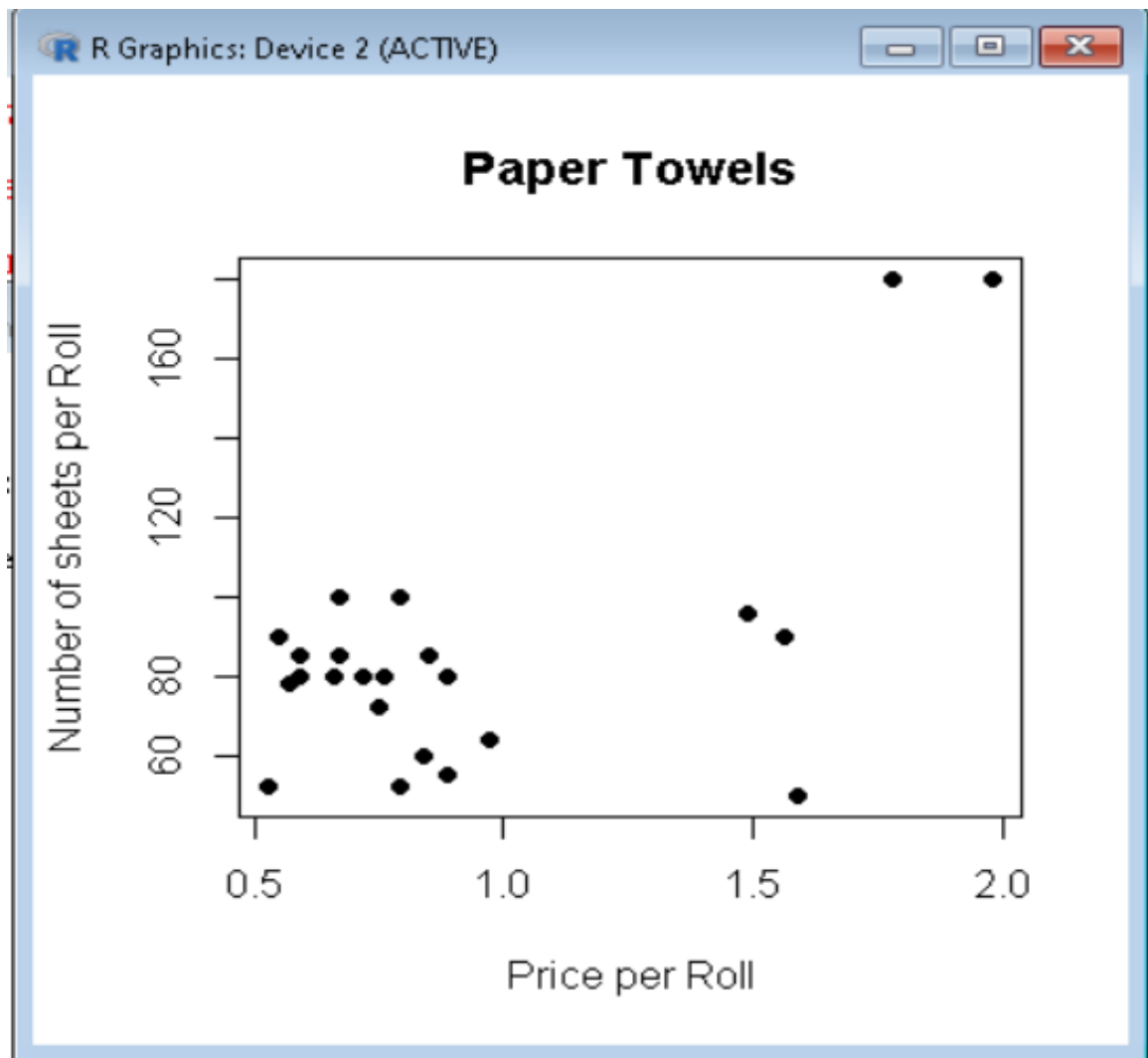
- 24 điểm có dường như nằm trên một đường thẳng không?
- Nếu không, có mối quan hệ nào khác giữa hai mức giá không?
- Những yếu tố nào có thể giải thích tại sao giá mỗi cuộn trên số tờ mỗi cuộn không phải là một hằng số?

Bài làm

Code:



```
> # 3.46
> p<-c(1.59,0.89,0.97,1.49,1.56,0.84,0.79,0.75,0.72,0.53,0.59,0.89,0.67,0.66,0.59,
+ 0.76,0.85,0.59,0.57,1.78,1.98,0.67,0.79,0.55)
> n<-c(50,55,64,96,90,60,52,72,80,52,85,80,85,80,80,80,85,85,78,180,180,100,100,90)
>
> plot(p,n,main="Paper Towels",xlab="Price per Roll",ylab="Number of sheets per Roll",pch=16)
> |
```



a) 24 điểm có dường như nằm trên một đường thẳng không?

Sau khi sử dụng ngôn ngữ R để vẽ biểu đồ tán xạ (scatter plot) của giá mỗi cuộn và số tờ mỗi cuộn của 24 hãng khăn giấy, thì ta có thể nhìn thấy 24 điểm không nằm trên một đường thẳng.

b) Nếu không, có mối quan hệ nào khác giữa hai mức giá không?

Nhìn vào biểu đồ, ta có thể thấy đa số các mức giá nằm trong khoảng từ 0.5 đến 1.0, tuy nhiên lại có 5 mức giá cá biệt nằm trong khoảng từ 1.5 đến 2.0.

c) Những yếu tố nào có thể giải thích tại sao giá mỗi cuộn trên số tờ mỗi cuộn không phải là một hằng số?

Những yếu tố như sự uy tín, mức độ nổi tiếng, và chất liệu làm giấy, ... của mỗi hãng khăn giấy có thể giải thích được tại sao giá mỗi cuộn trên số tờ không phải là một hằng số.

Bài 3.47: (M.Tâm)

Refer to Exercise 3.45. Construct boxplots for both price per roll and number of sheets per roll. Are there any “unusual” brands in the data? Are there any “unusual” brands in the data? (Tham khảo bài tập 3.45. Xây dựng biểu đồ boxplot cho **price per roll** và **number of sheets per roll**. Có bất kỳ phần tử "bất thường"(phần tử ngoại lệ) nào trong dữ liệu không?)

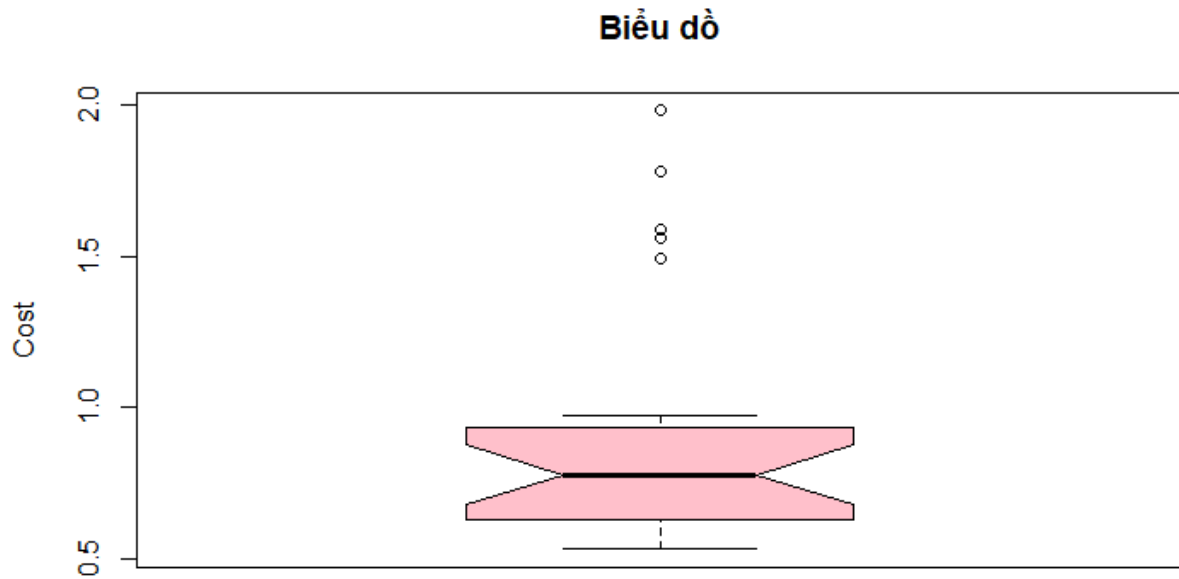
Bài làm

> data

	Brand	Price.per.Roll	Number.of.Sheets.per.Roll	Cost.per.Sheet
1	1	1.59	50	0.0318
2	2	0.89	55	0.0162
3	3	0.97	64	0.0152
4	4	1.49	96	0.0155
5	5	1.56	90	0.0173
6	6	0.84	60	0.0140
7	7	0.79	52	0.0152
8	8	0.75	72	0.0104
9	9	0.72	80	0.0090
10	10	0.53	52	0.0102
11	11	0.59	85	0.0069
12	12	0.89	80	0.0111
13	13	0.67	85	0.0079
14	14	0.66	80	0.0083
15	15	0.59	80	0.0074
16	16	0.76	80	0.0095
17	17	0.85	85	0.0100
18	18	0.59	85	0.0069
19	19	0.57	78	0.0073
20	20	1.78	180	0.0099
21	21	1.98	180	0.0100
22	22	0.67	100	0.0067
23	23	0.79	100	0.0079

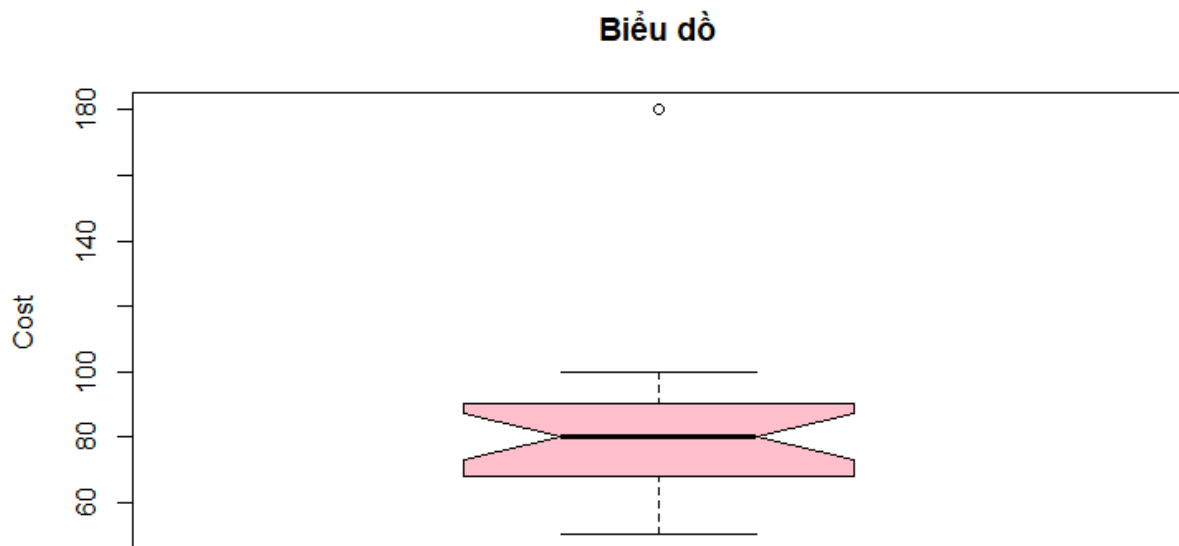
- Dữ liệu đầu vào:
- Biểu đồ của **price per roll**

Boxplot (data\$Price.per.Roll, main="Biểu đồ ", ylab="Cost", horizontal = FALSE, col = "pink")



- Biểu đồ của **number of sheets per roll**:

Boxplot (`data$Number.of.Sheets.per.Roll`, `main="Biểu đồ"`, `horizontal = FALSE`, `col = "pink"`, `notch = TRUE`, `ylab="Cost"`)



- Với bộ dữ liệu trên có sự tồn tại của các phần tử ngoại lệ, điều đó cũng đã được thể hiện rõ qua biểu đồ boxplot.