

Bài giảng 7b: Biểu đồ với R, phần II

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia
Đại học Tôn Đức Thắng, Việt Nam

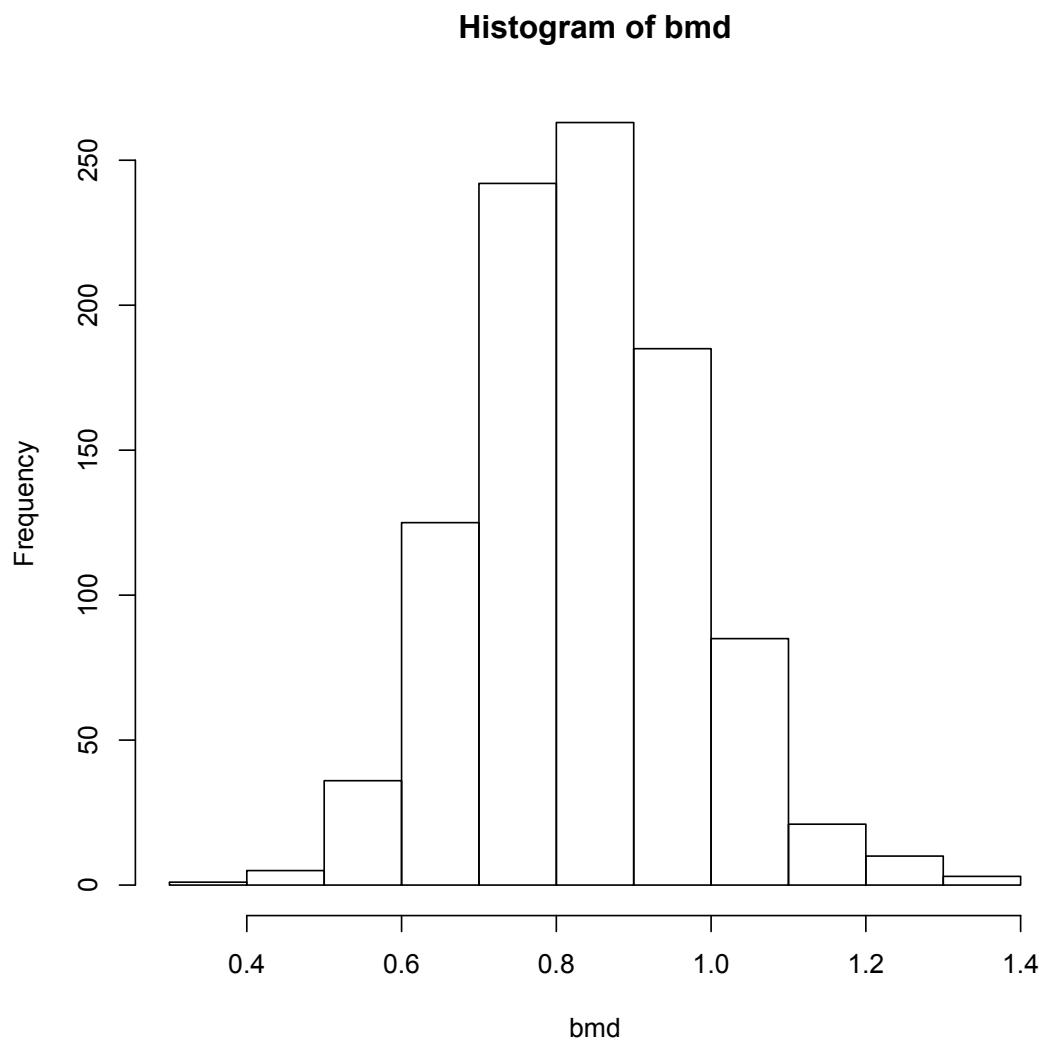
Một số biểu đồ chính trong khoa học

- Biểu đồ phân bố (histogram)
- Biểu đồ hộp (box plot)
- Biểu đồ thanh (bar plot)
- Biểu đồ tương quan (scatter plot)

Biểu đồ phân bố

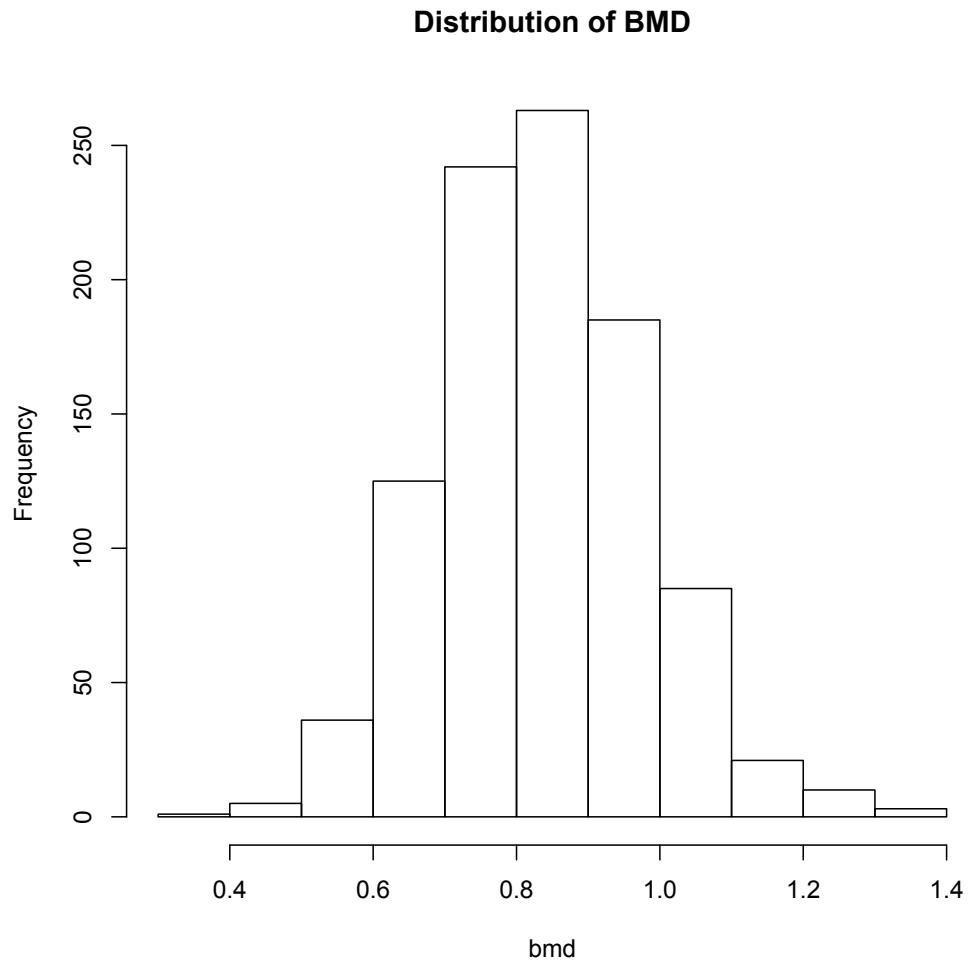
Phân bố dữ liệu: histogram

`hist(fnbmd)`



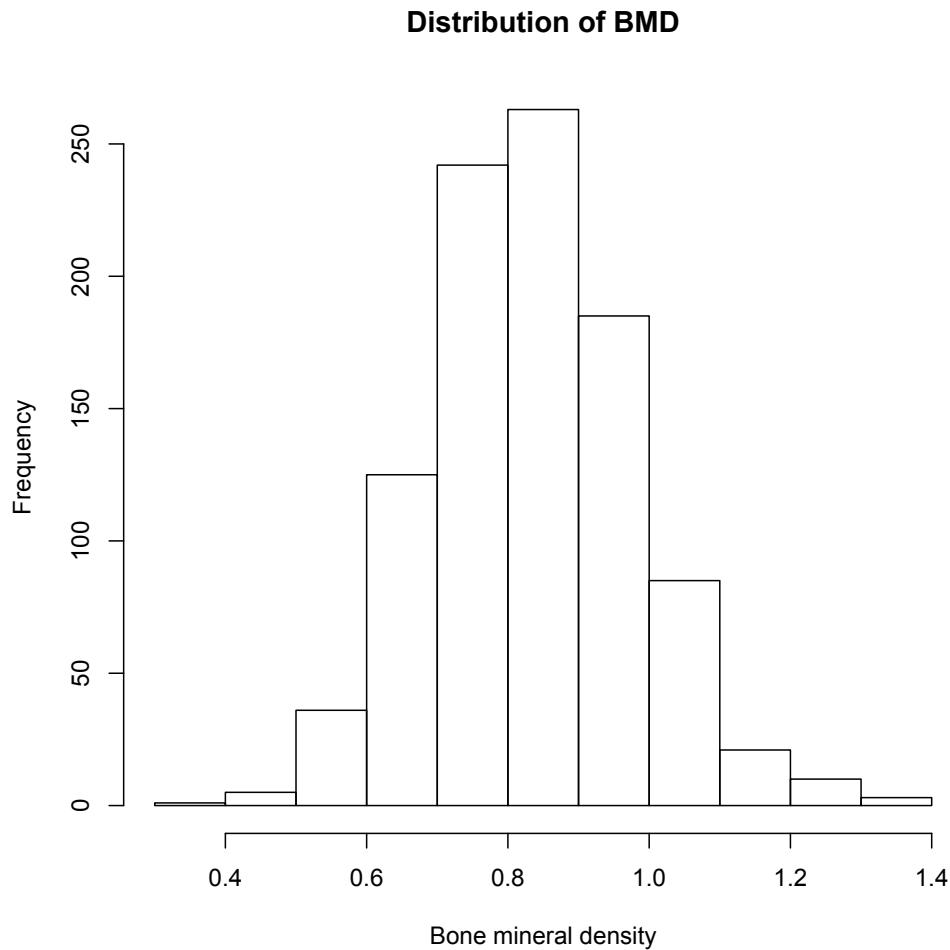
Thêm tựa đề ...

```
hist(fnbmd, main="Distribution of BMD")
```



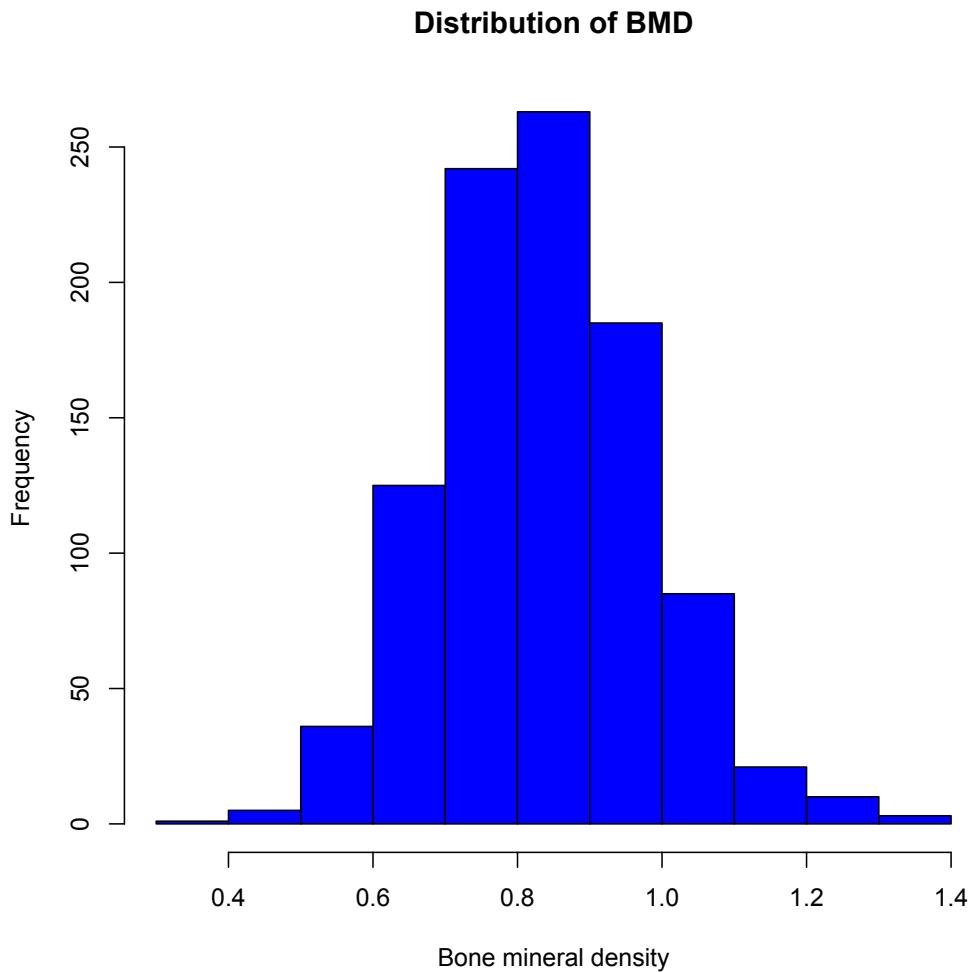
tên trục tung và trục hoành ...

```
hist(fnbmd, main="Distribution of BMD",  
xlab="Bone mineral density",ylab="Frequency")
```



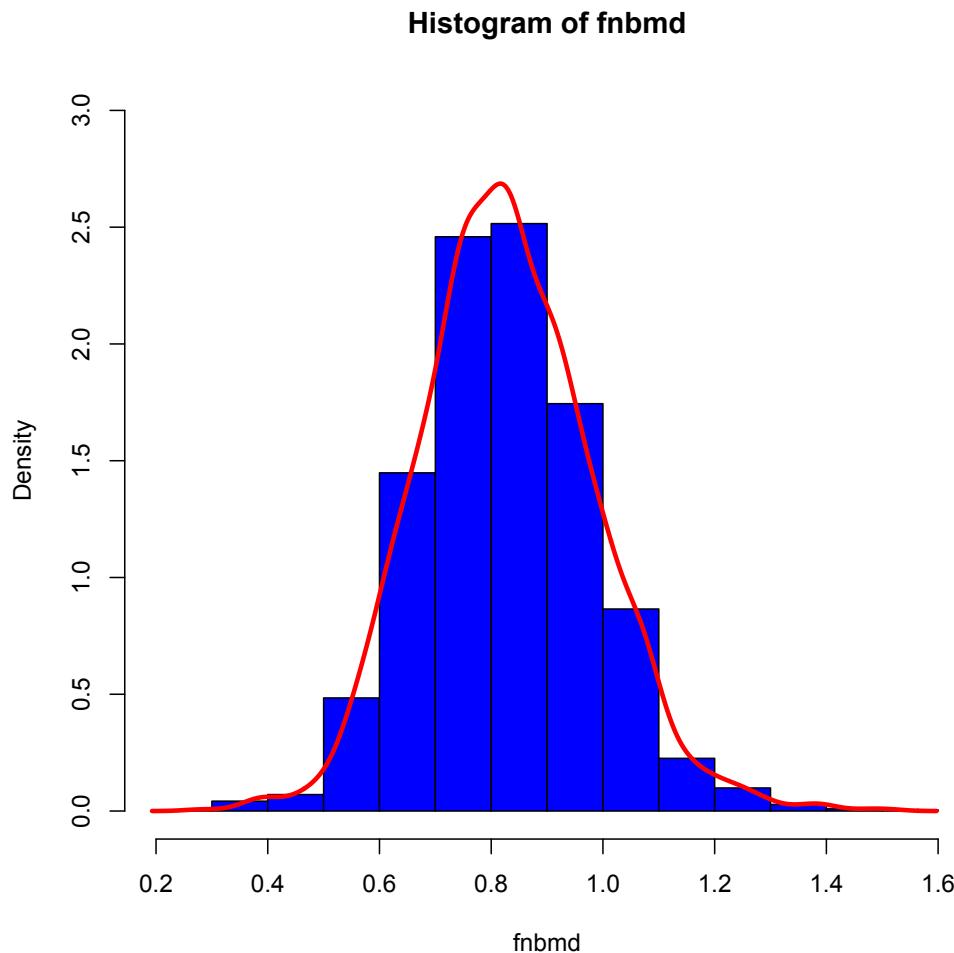
thêm màu ...

```
hist(fnbmd, main="Distribution of BMD", xlab="Bone  
mineral density", ylab="Frequency", col="blue")
```



thêm đường density curve

```
hist(fnbmd, prob=T, ylim=c(0,3), col="blue")
lines(density(na.omit(fnbmd)), col="red",
lwd=3)
```



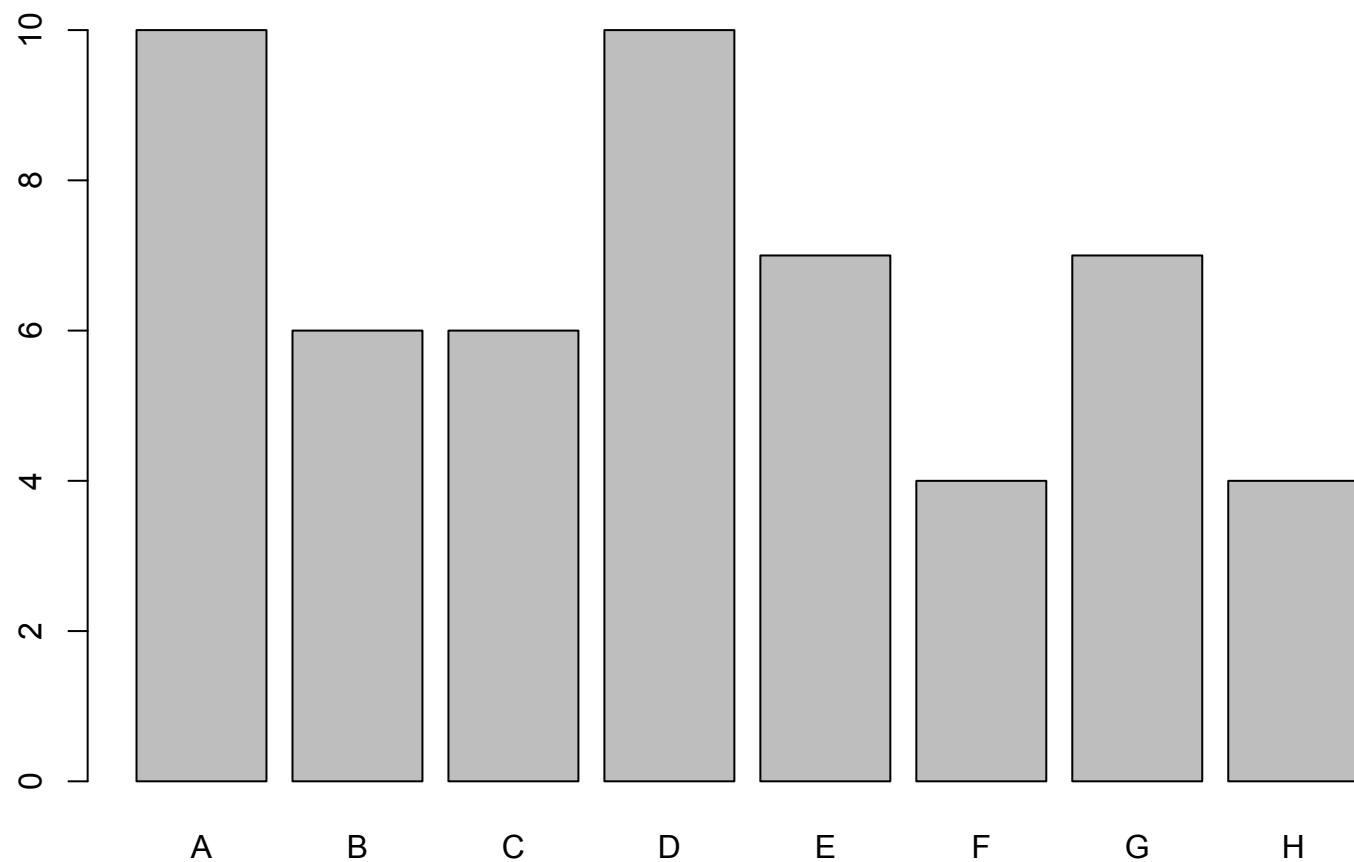
Biểu đồ thanh (bar plot)

Dữ liệu gốc (tần số)

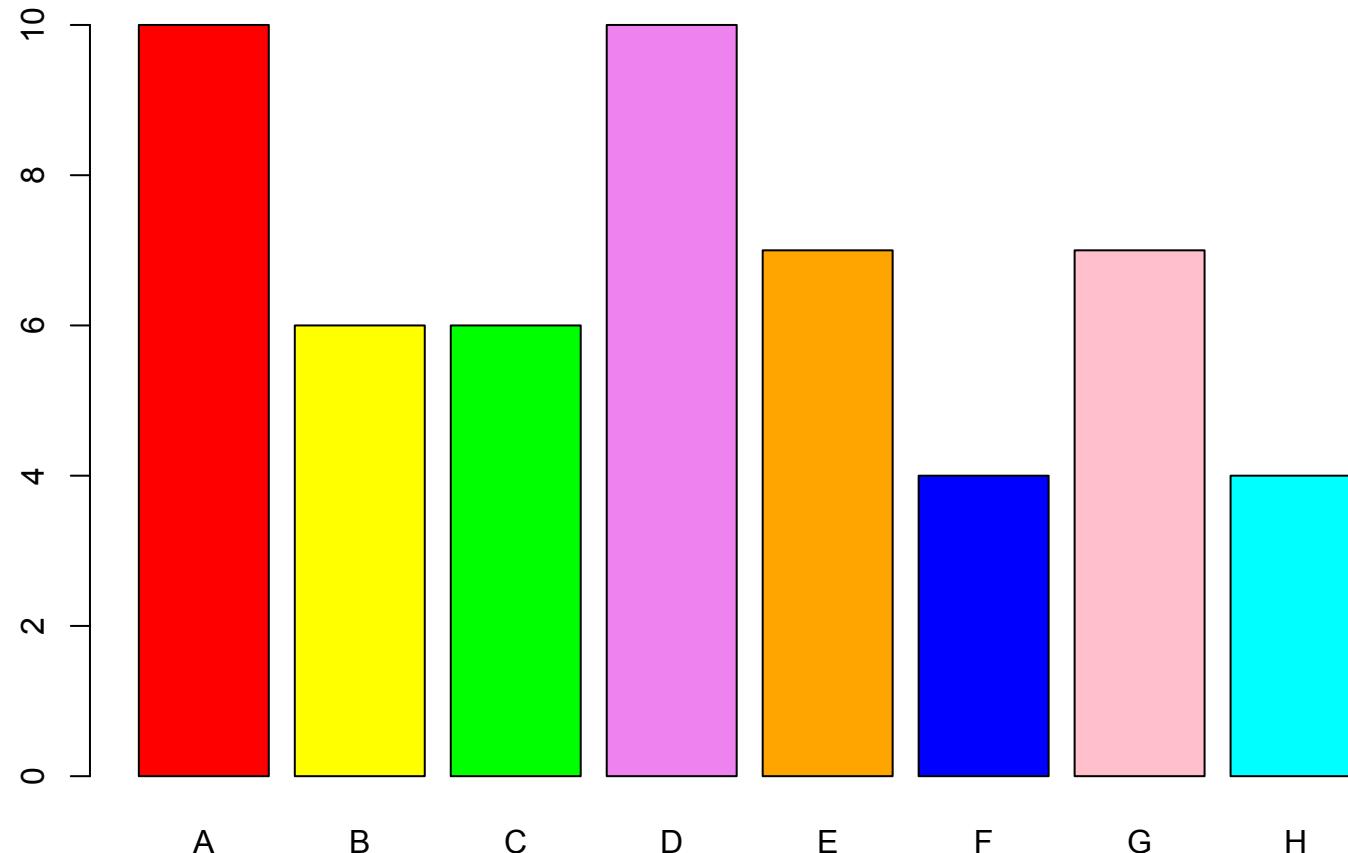
```
School = c("A", "A", "A", "A", "A", "A", "A", "A", "A", "A",  
"A", "B", "B", "B", "B", "B", "B", "C", "C", "C", "C",  
"C", "C", "D", "D", "D", "D", "D", "D", "D", "D", "D",  
"D", "D", "E", "E", "E", "E", "E", "E", "E", "F", "F", "F",  
"F", "G", "G", "G", "G", "G", "G", "G", "H", "H", "H", "H",  
"H")
```

```
Freq = table(School)
```

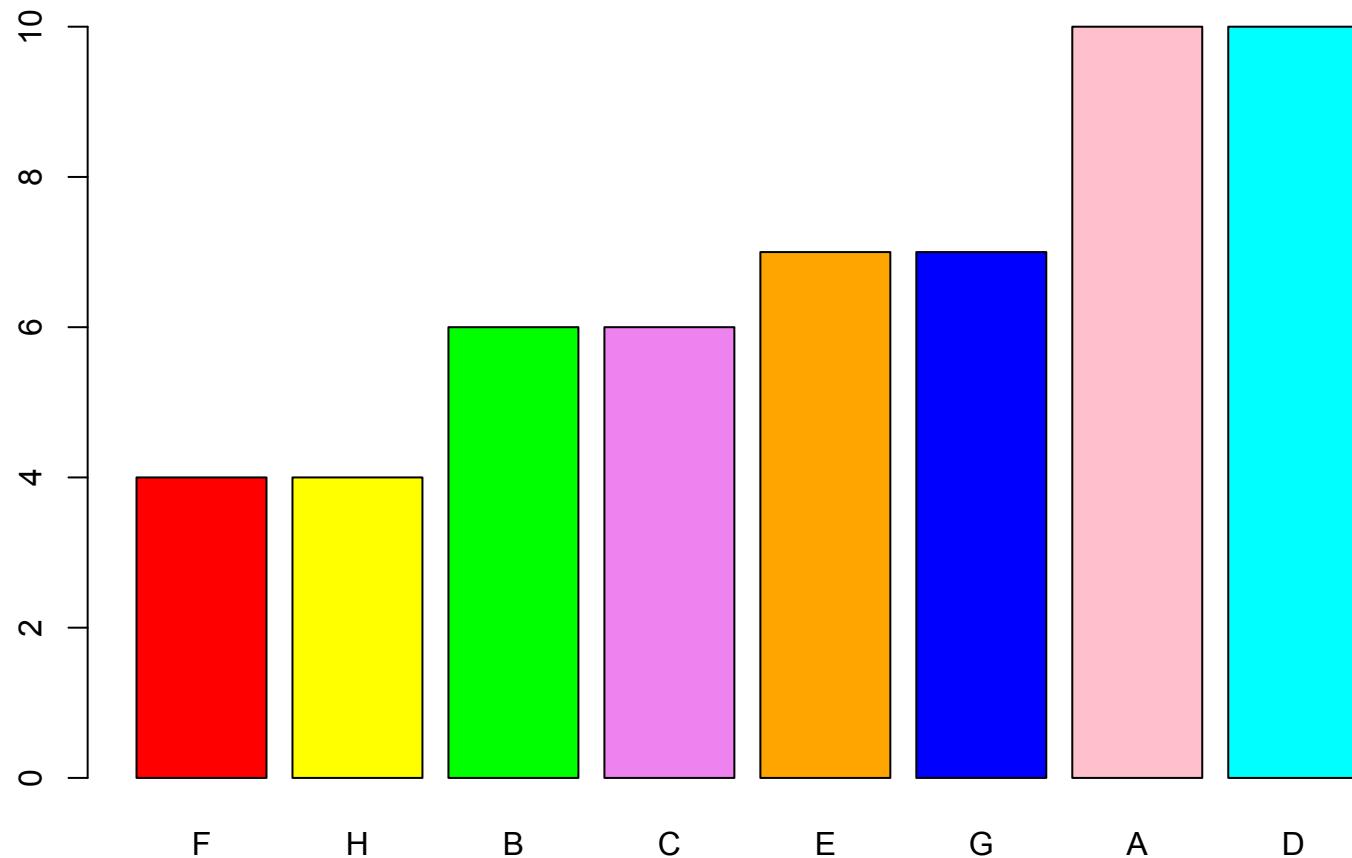
```
barplot(Freq)
```



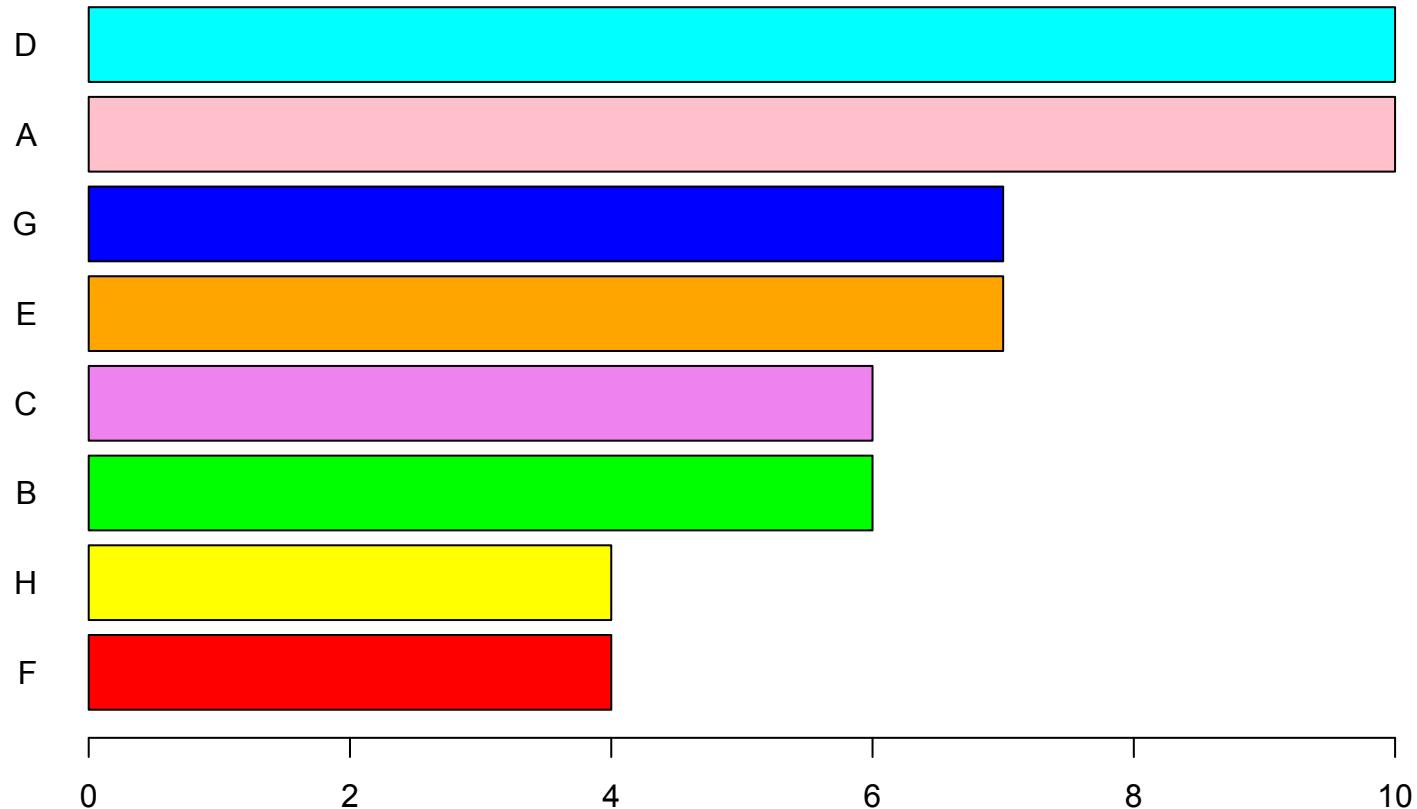
```
> colors = c("red", "yellow", "green", "violet", "orange",
"blue", "pink", "cyan")
> barplot(Freq, col=colors)
```



```
> barplot(sort(Freq), col=colors)
```



```
> barplot(sort(Freq), col=colors, horiz=T, las=1)
```



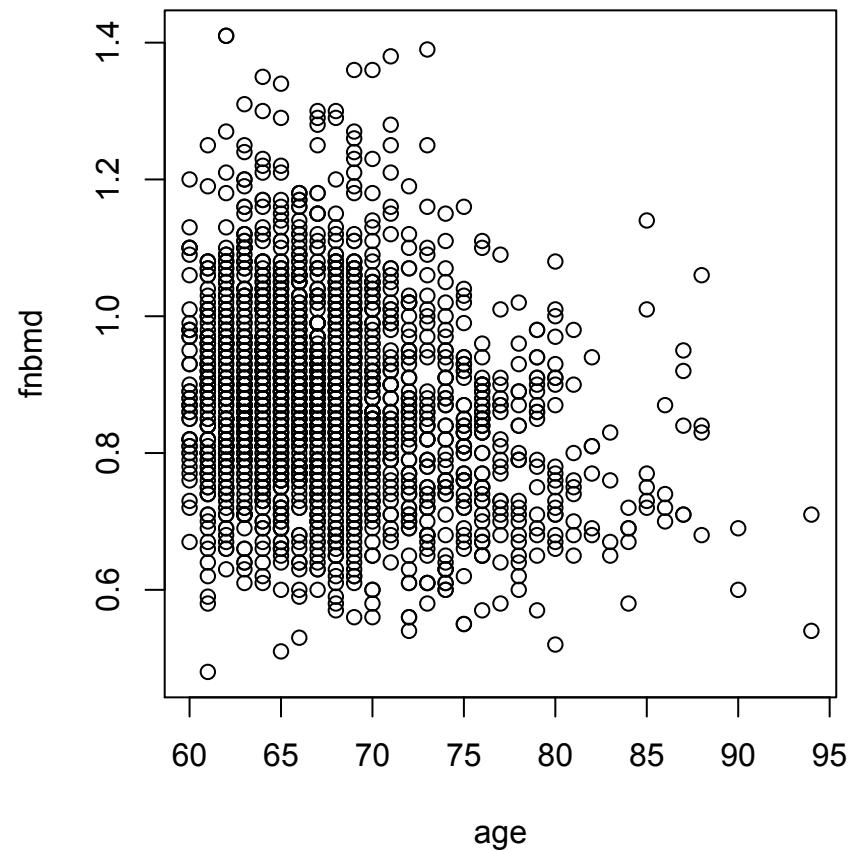
scatterplot trong car
"companion to applied regression"

Biểu đồ tương quan đơn giản

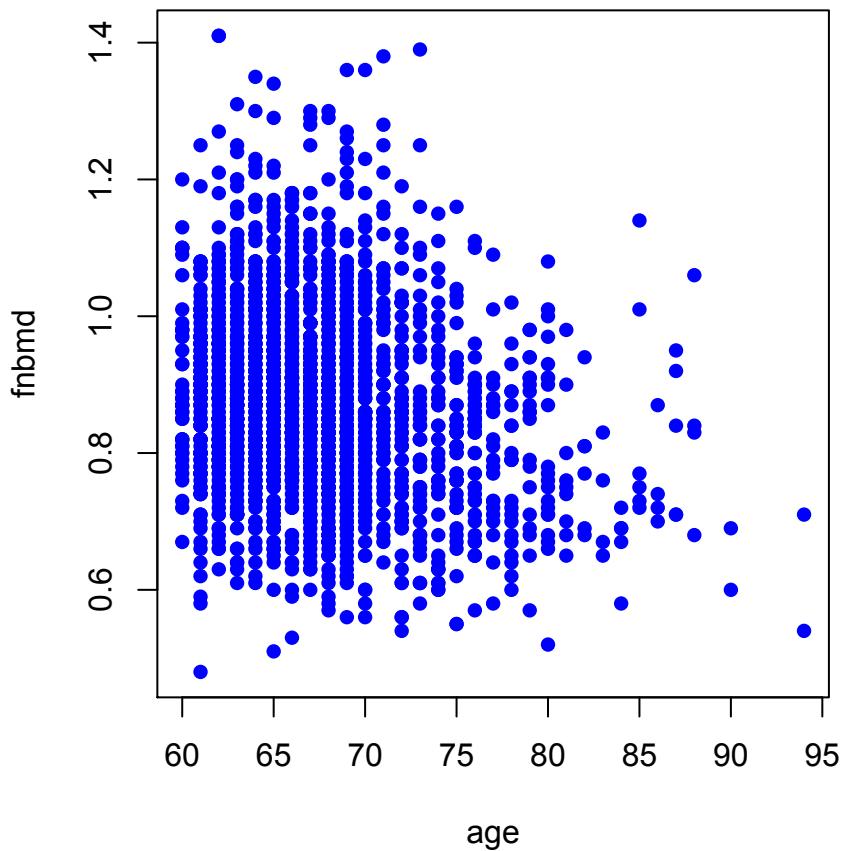
Dữ liệu: osteo

```
osteо = read.table("~/Google Drive/TDT Projects/  
Workshop 12-2015/Datafiles/osteо data.txt",  
header=T, na.strings=".")  
  
osteо = subset(osteо, age>=60)  
  
attach(osteо);
```

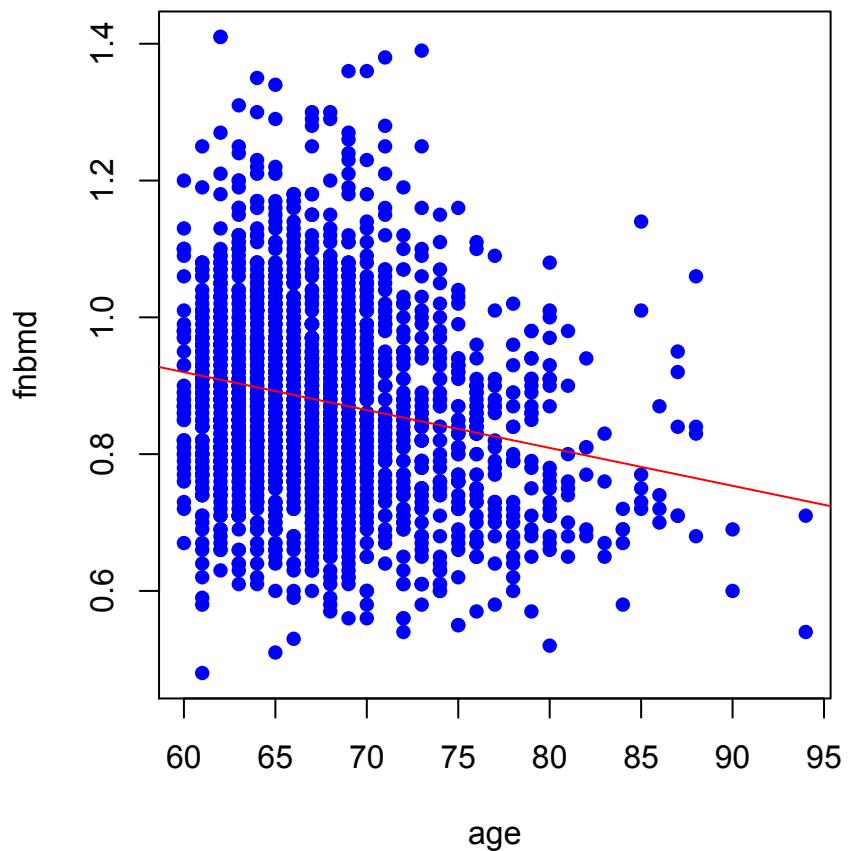
```
plot(fnbmd ~ age)
```



```
plot(fnbmd ~ age,  
      pch=16, col="blue")
```



```
> plot(fnbmd ~ age, pch=16,  
col="blue")  
> abline(lm(fnbmd ~ age),  
col="red")
```



Biểu đồ "tán xạ" (scatter plot)

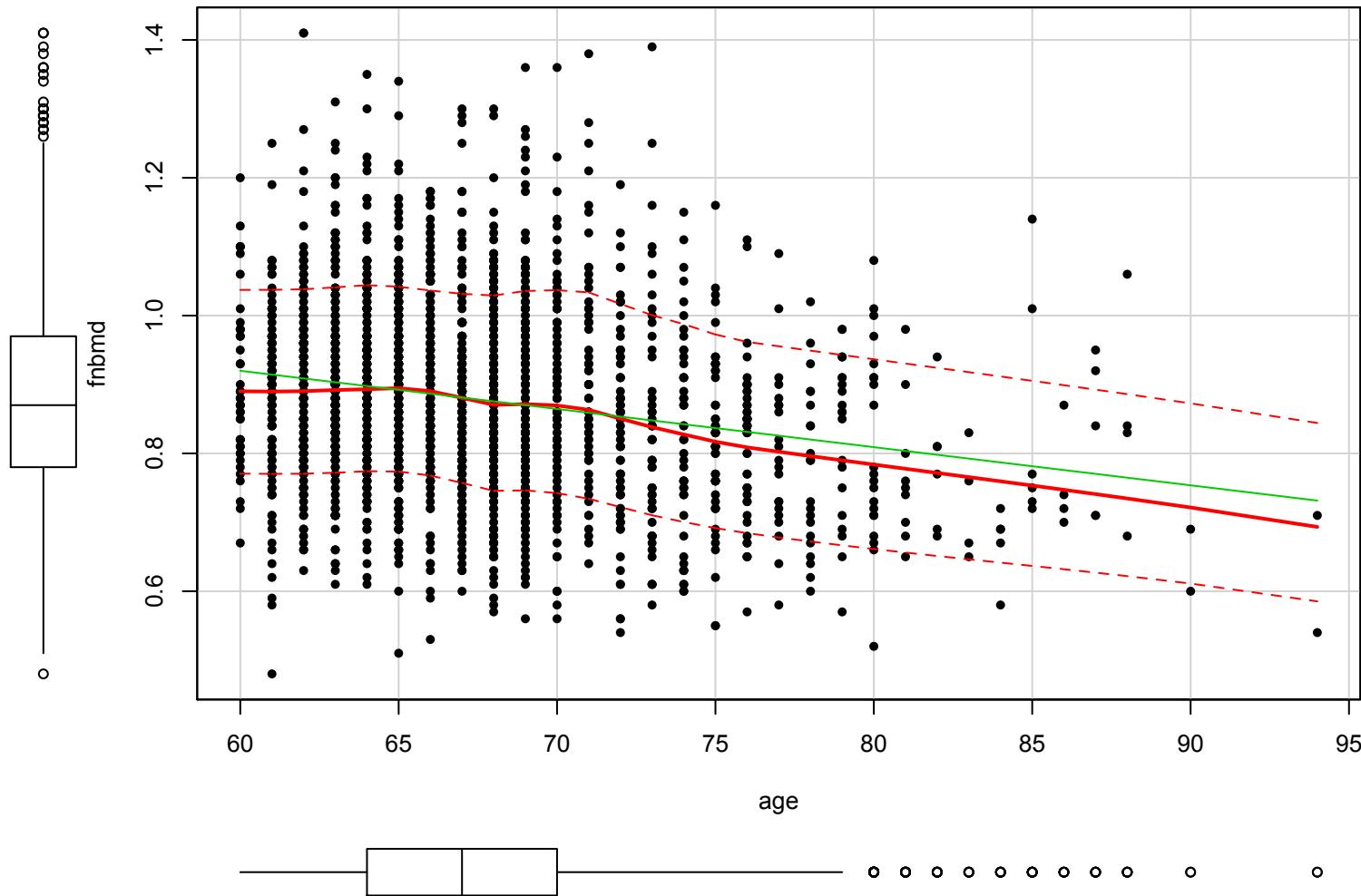
- Package "car"
- Mối liên quan giữa x và y
- scatterplot trong package "car"

scatterplot(y ~ x | group, ...)

Biểu đồ đơn giản dùng package "car"

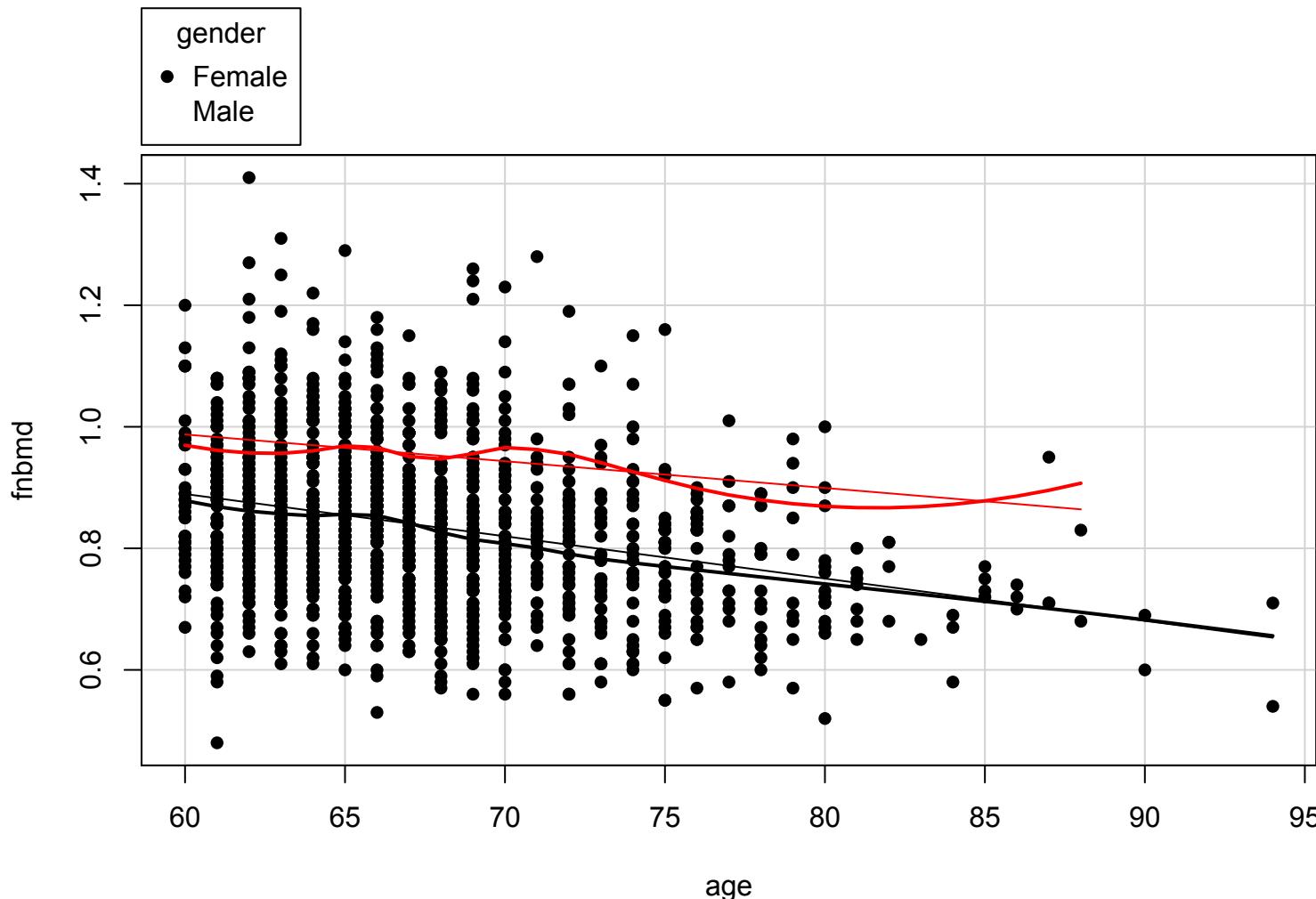
```
library(car)
```

```
scatterplot(fnbmd ~ age)
```



Biểu đồ "tương tác"

scatterplot(fnbmd ~ age | gender)



"Tô điểm" cho biểu đồ

thêm kí hiệu

```
scatterplot(fnbmd ~ age | gender,  
pch=c(1,16))
```

thêm màu

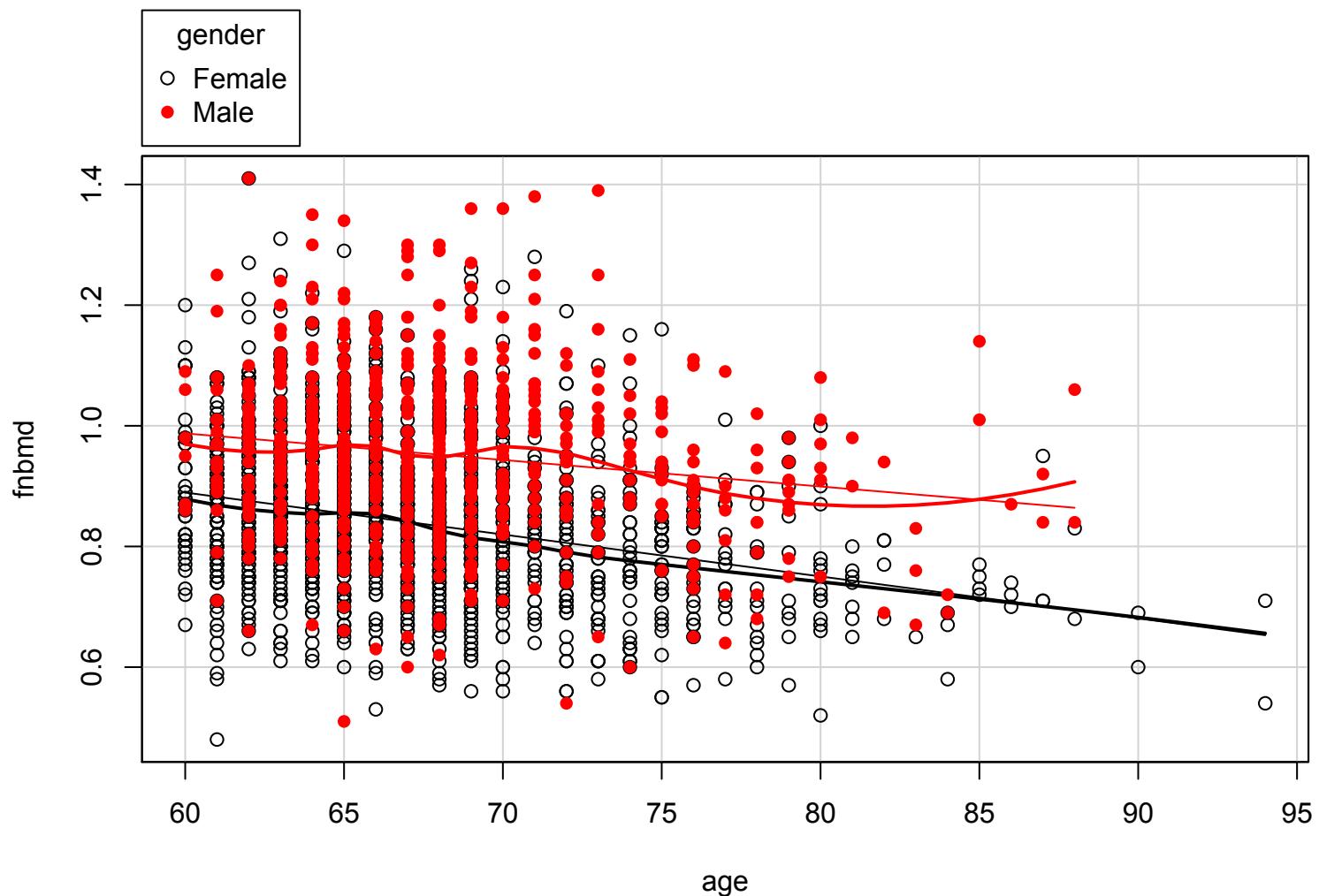
```
scatterplot(fnbmd ~ age | gender,  
pch=c(1,16), col=c("red","black"))
```

thêm tiêu đề, giá trị trực tung và hoành

```
scatterplot(fnbmd ~ age | gender,  
pch=c(1,16), col=c("red","black"),  
xlim=c(55,95), ylim=c(0.2, 1.4), xlab="Age",  
ylab="FNBND")
```

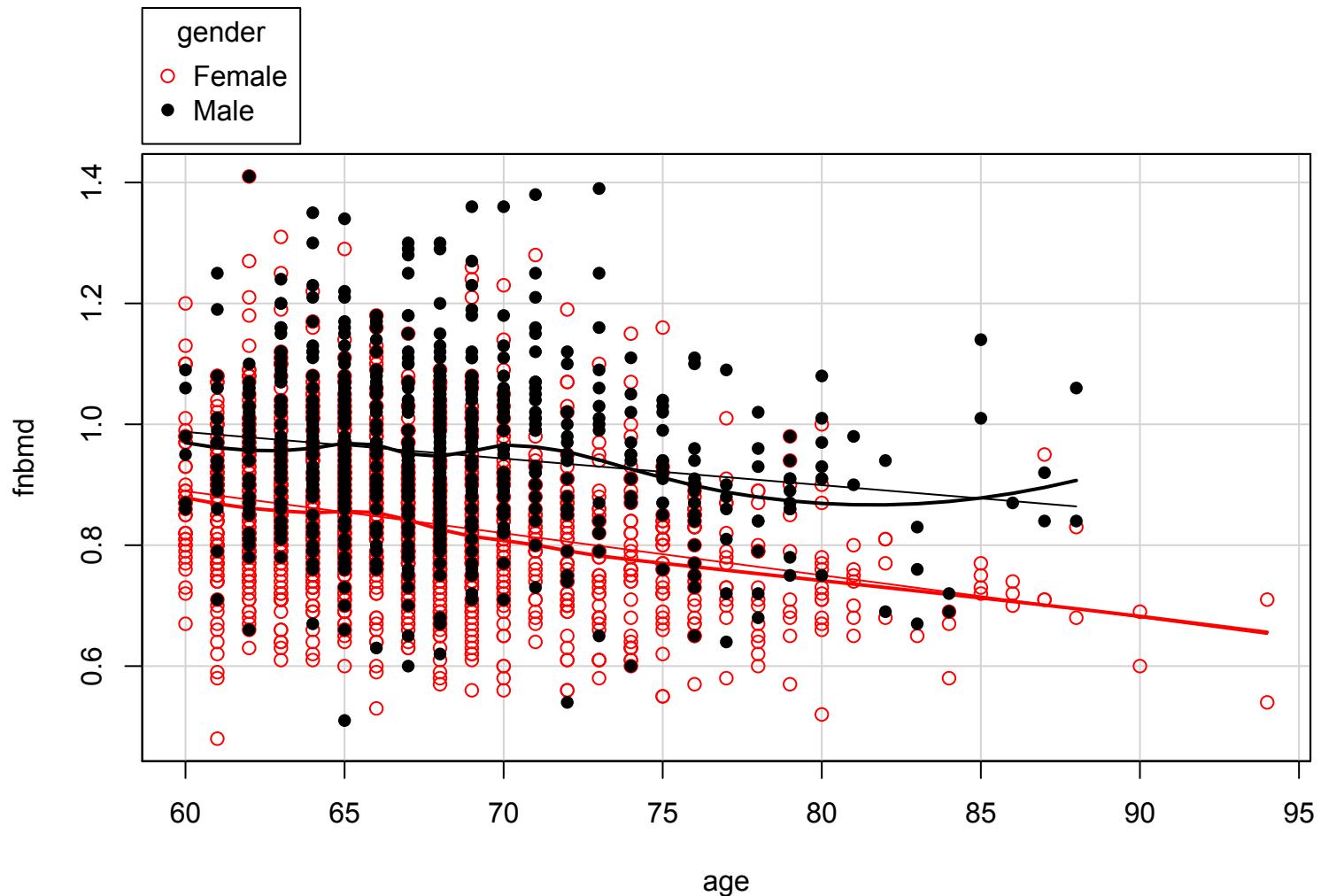
thêm kí hiệu

```
scatterplot(fnbmd ~ age | gender, pch=c(1,16))
```

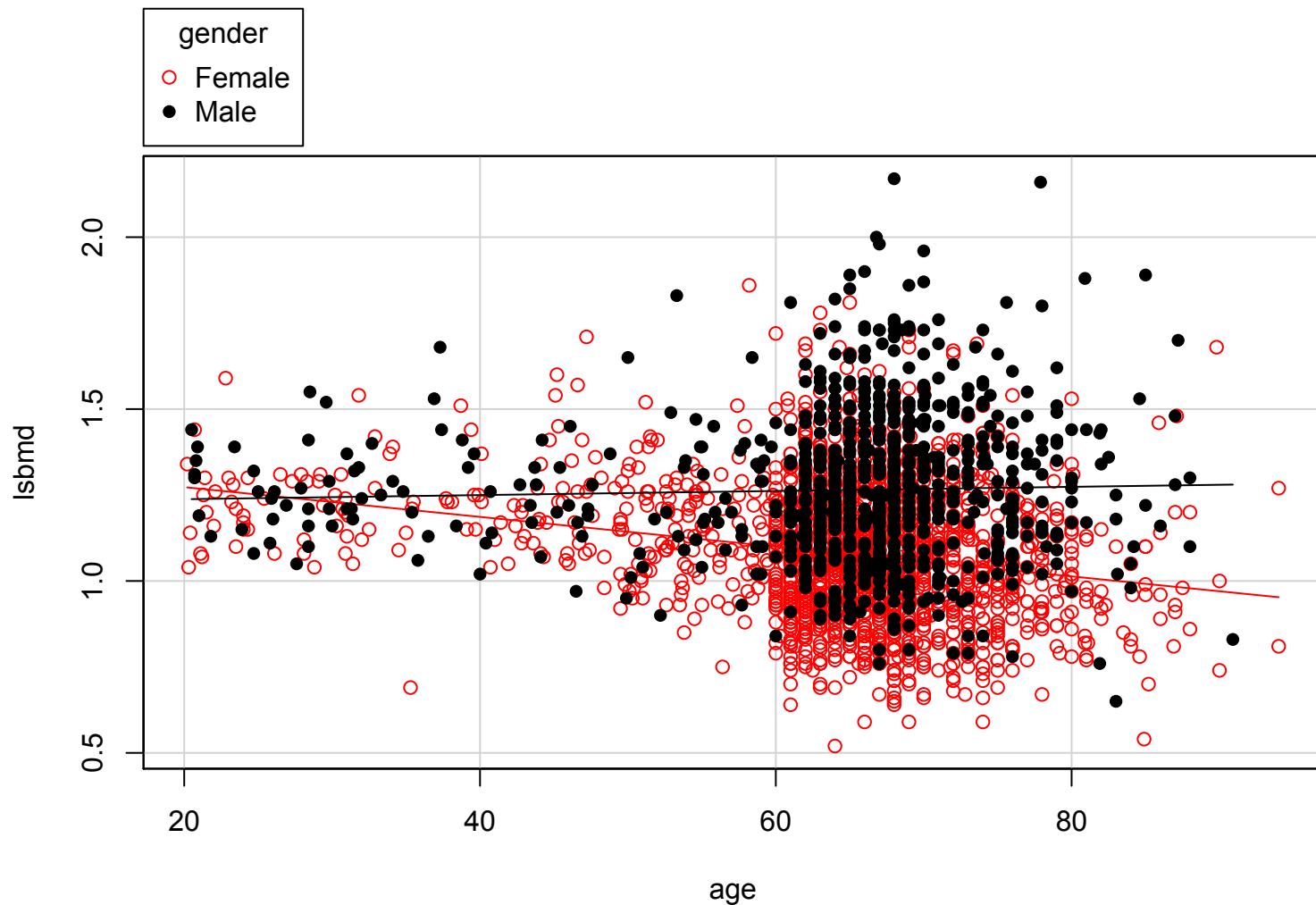


thêm màu

```
scatterplot(fnbmd ~ age | gender, pch=c(1,16),  
col=c("red","black"))
```



```
# thêm màu  
scatterplot(lsbmd ~ age | gender, smooth=F,  
pch=c(1,16), col=c("red","black"))
```



"Tô điểm" cho biểu đồ

```
# bỏ đường biểu diễn smooth  
scatterplot(fnbmd ~ age | gender,  
pch=c(1,16), col=c("red","black"),  
xlim=c(55,95), ylim=c(0.2, 1.4),  
xlab="Age", ylab="FNBND",  
smooth=F)
```

Biểu đồ tương quan đa biến

Bối cảnh

- Có nhiều biến liên tục (continuous variables)
- scatterplot chỉ cung cấp tương quan giữa 2 biến
- Muốn xem mối tương quan giữa các biến *cùng một lúc*

Hàm pairs.panels(data) trong psych

- Package "psych"
- Hàm pairs.panels(data)

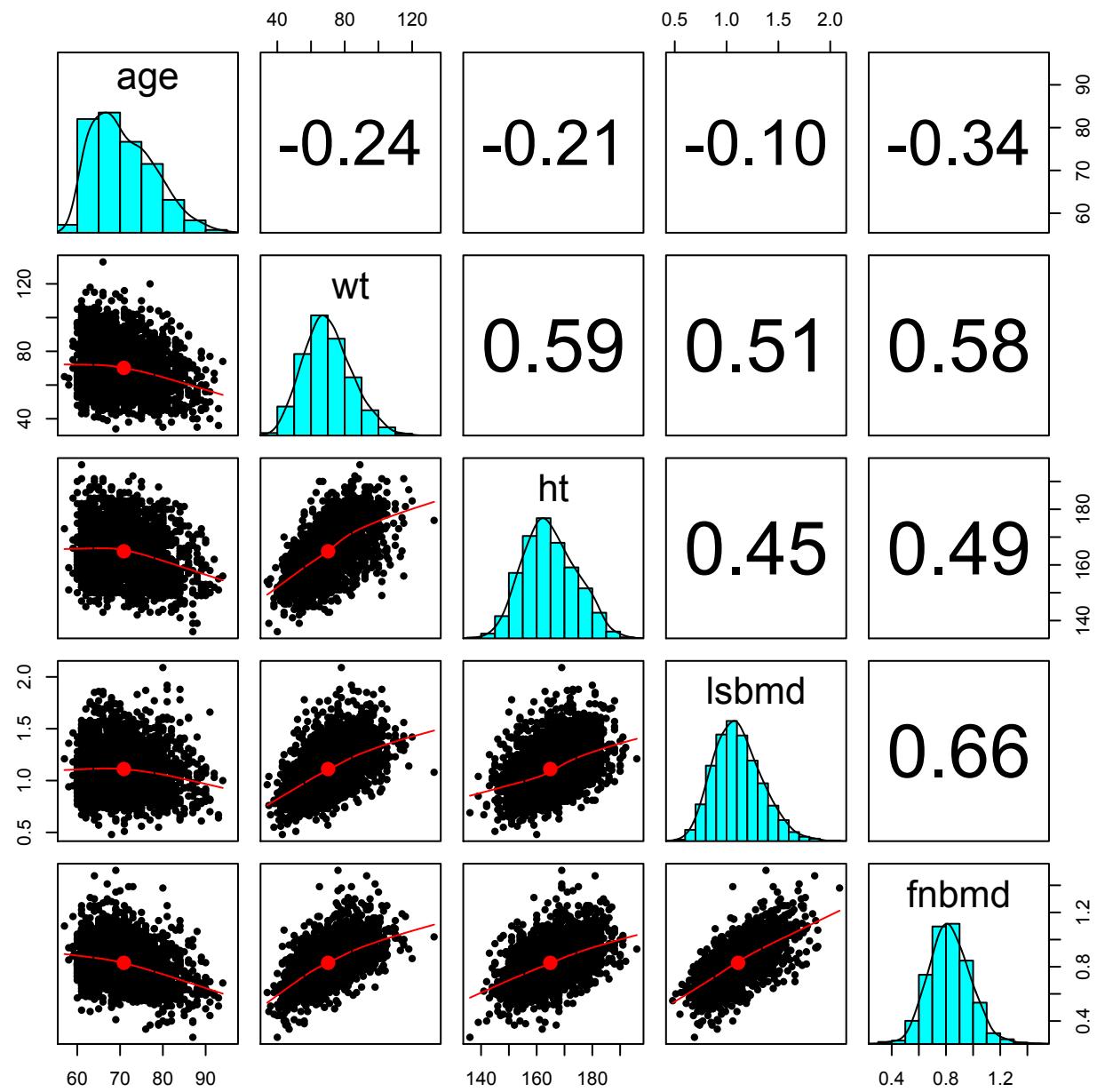
Lệnh pairs.panels in psych

```
# các biến: age, wt, ht, lsbmd, fnbmd
```

```
dat = cbind(age, wt, ht, lsbmd,  
fnbmd)
```

```
require(psych)
```

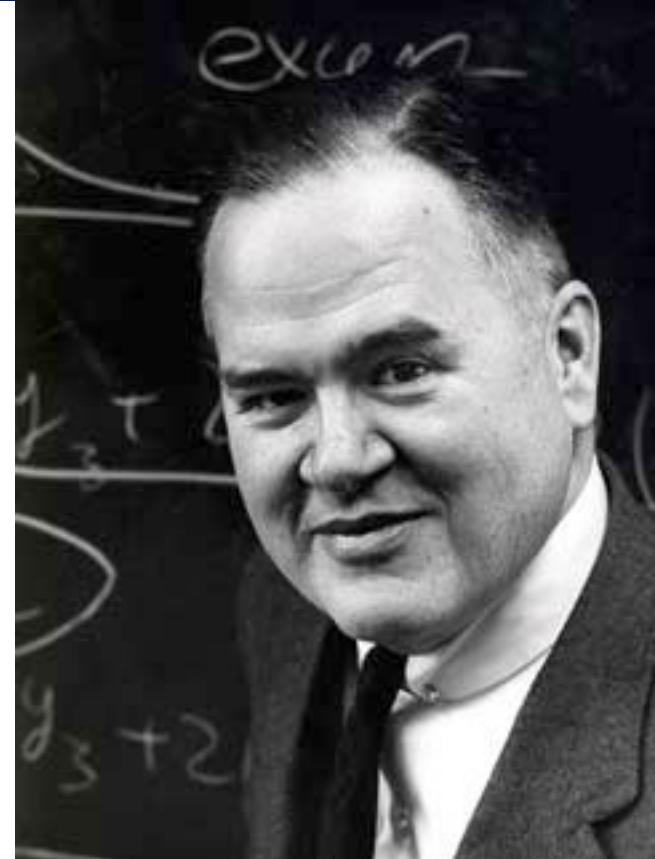
```
pairs.panels(dat)
```



Biểu đồ hộp (box plot)

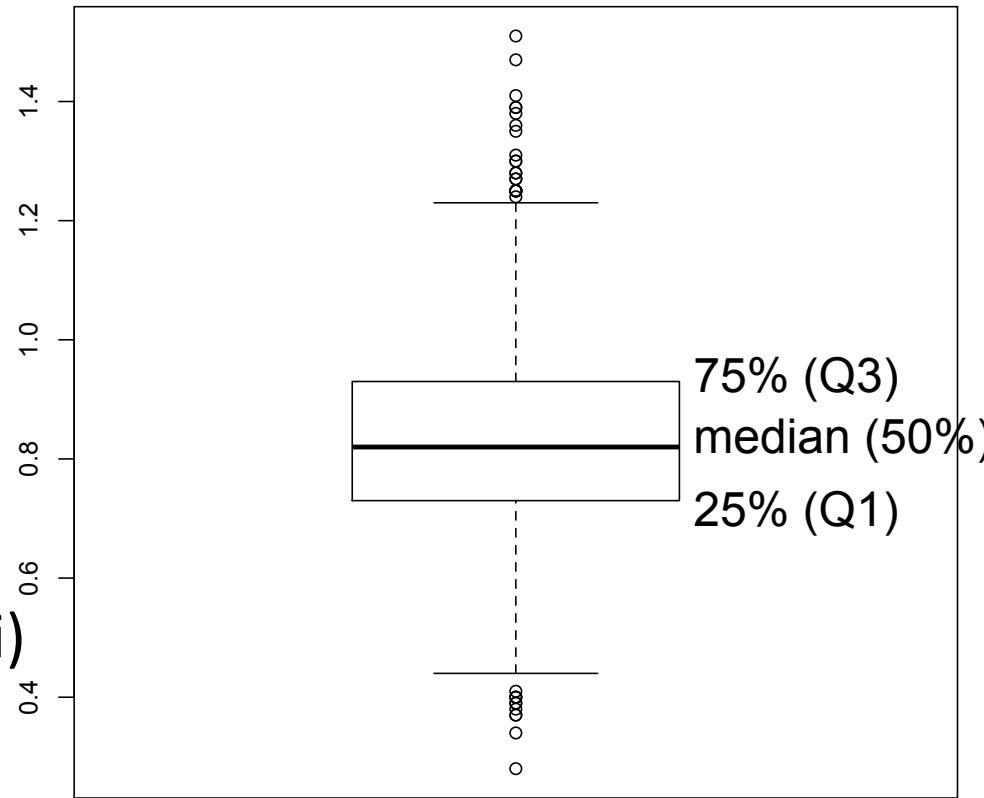
John Tukey (1915 – 2000)

- Nhà hoá học, thống kê học
- Bell labs
- Nổi tiếng với "Exploratory Data Analysis"
- Box plot (1977), jackknife method, Tukey's range test, v.v.
- Sáng tạo ra chữ *bit* (binary digit), *software*



5 yếu tố trong biểu đồ hộp

- median = trung vị
- 2 hinges = 25% và 75% bách phân vị
- fences = $1.5 \times$ interquartile range
- whiskers = nối hai hinges
- Outliers (có thể giá trị ngoại vi)

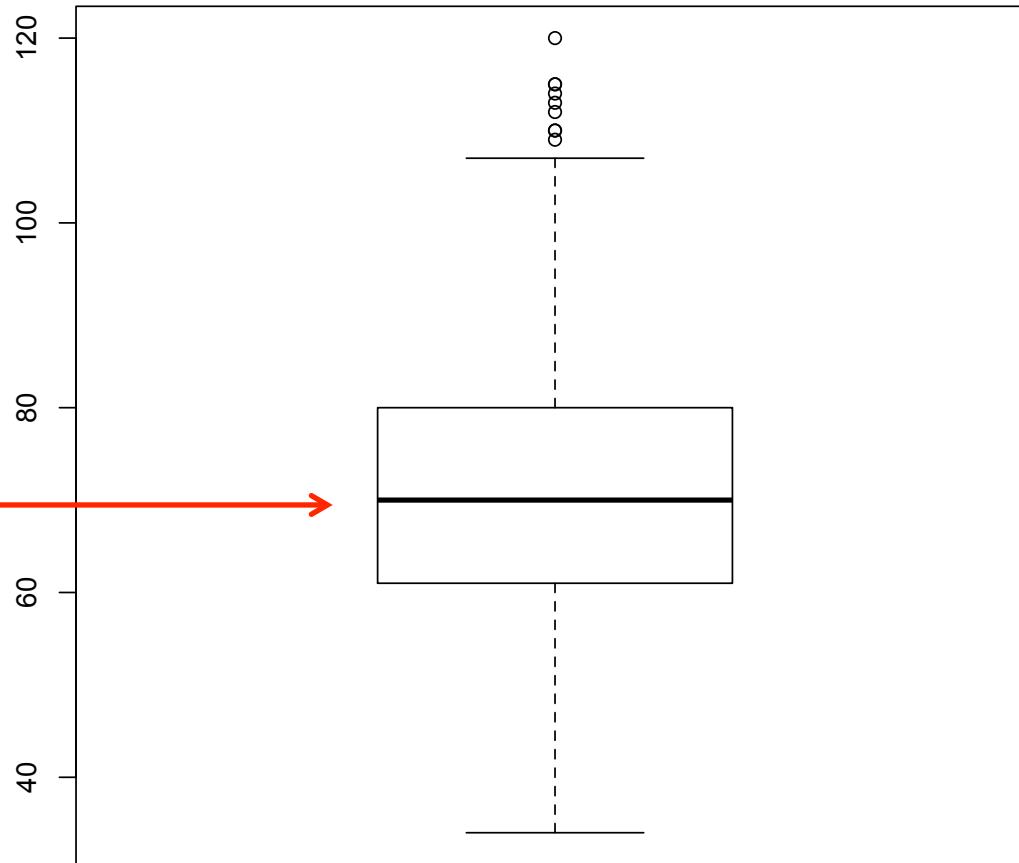


interquartile range = $Q3 - Q1$

Box plot

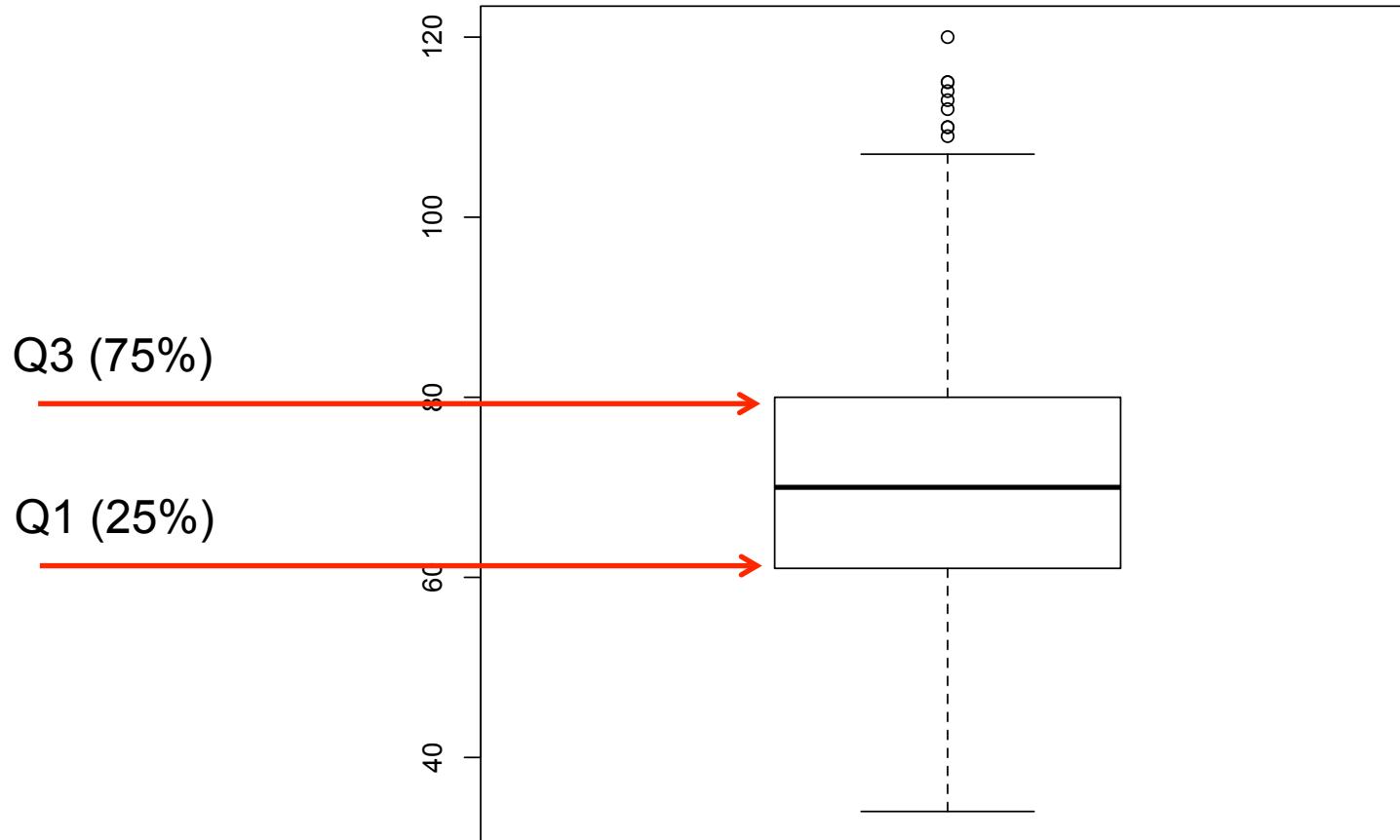
`boxplot(wt)`

Trung vị (50%)



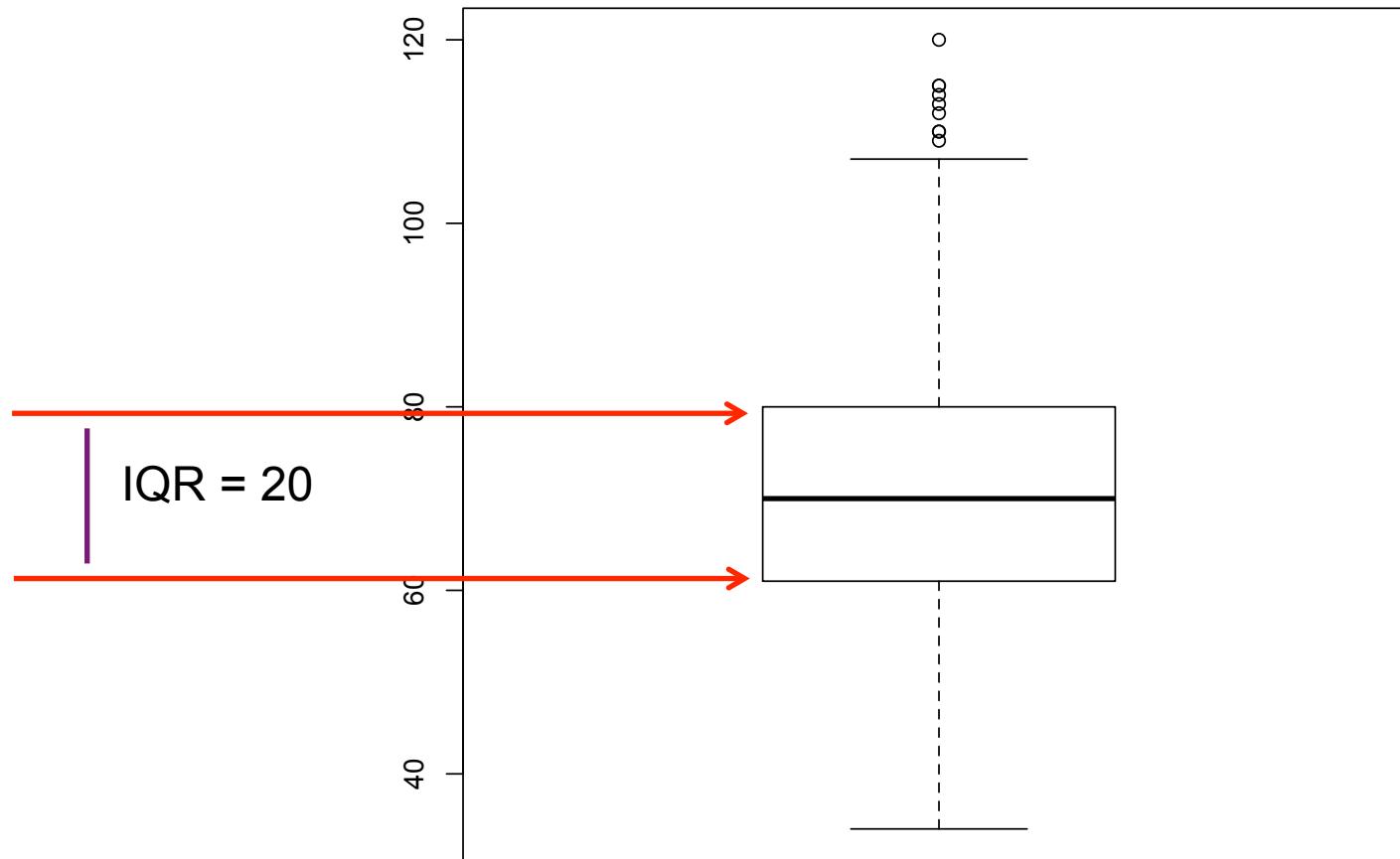
Box plot

`boxplot(wt)`



Box plot

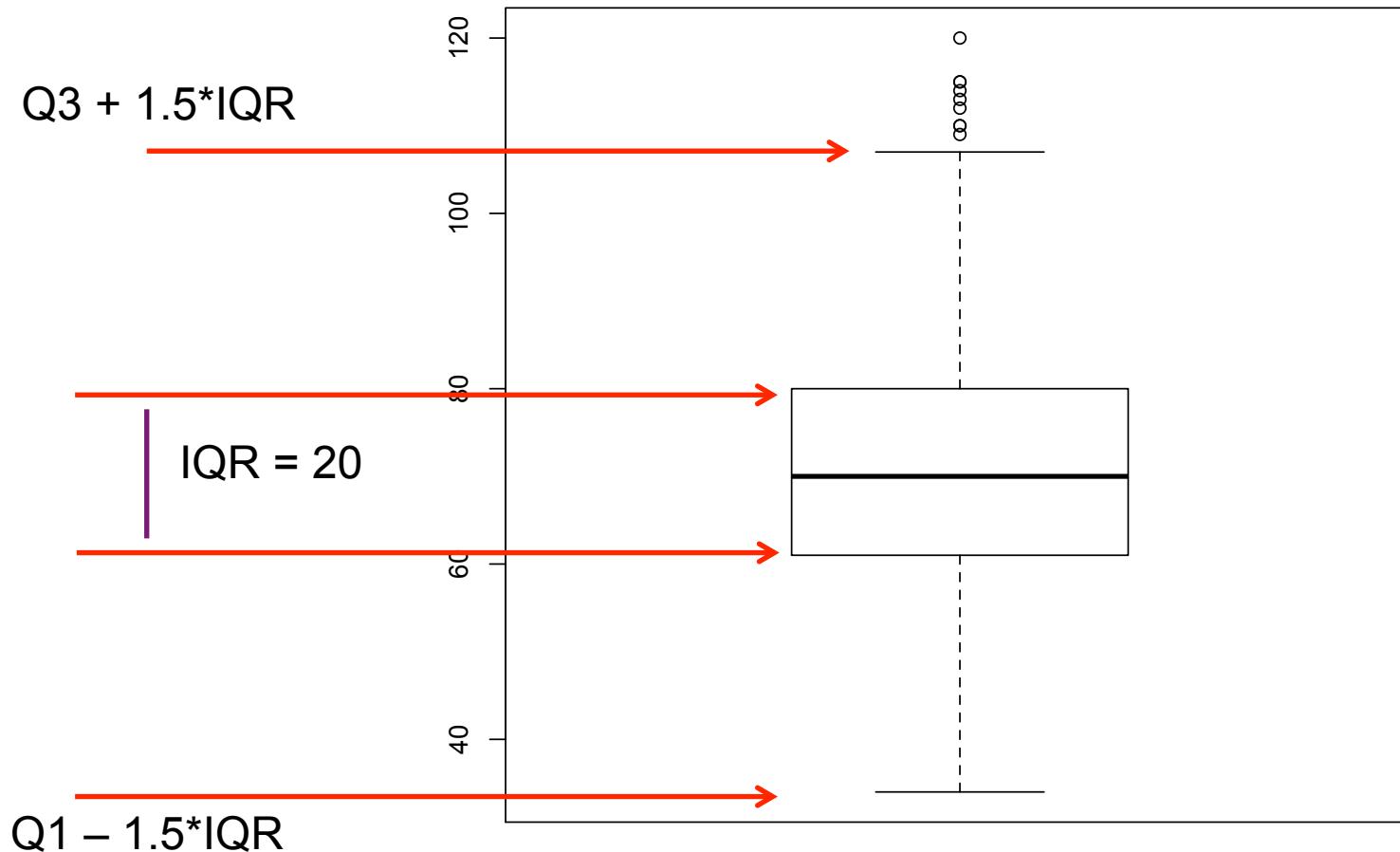
`boxplot(wt)`



IQR = interquartile range

Box plot

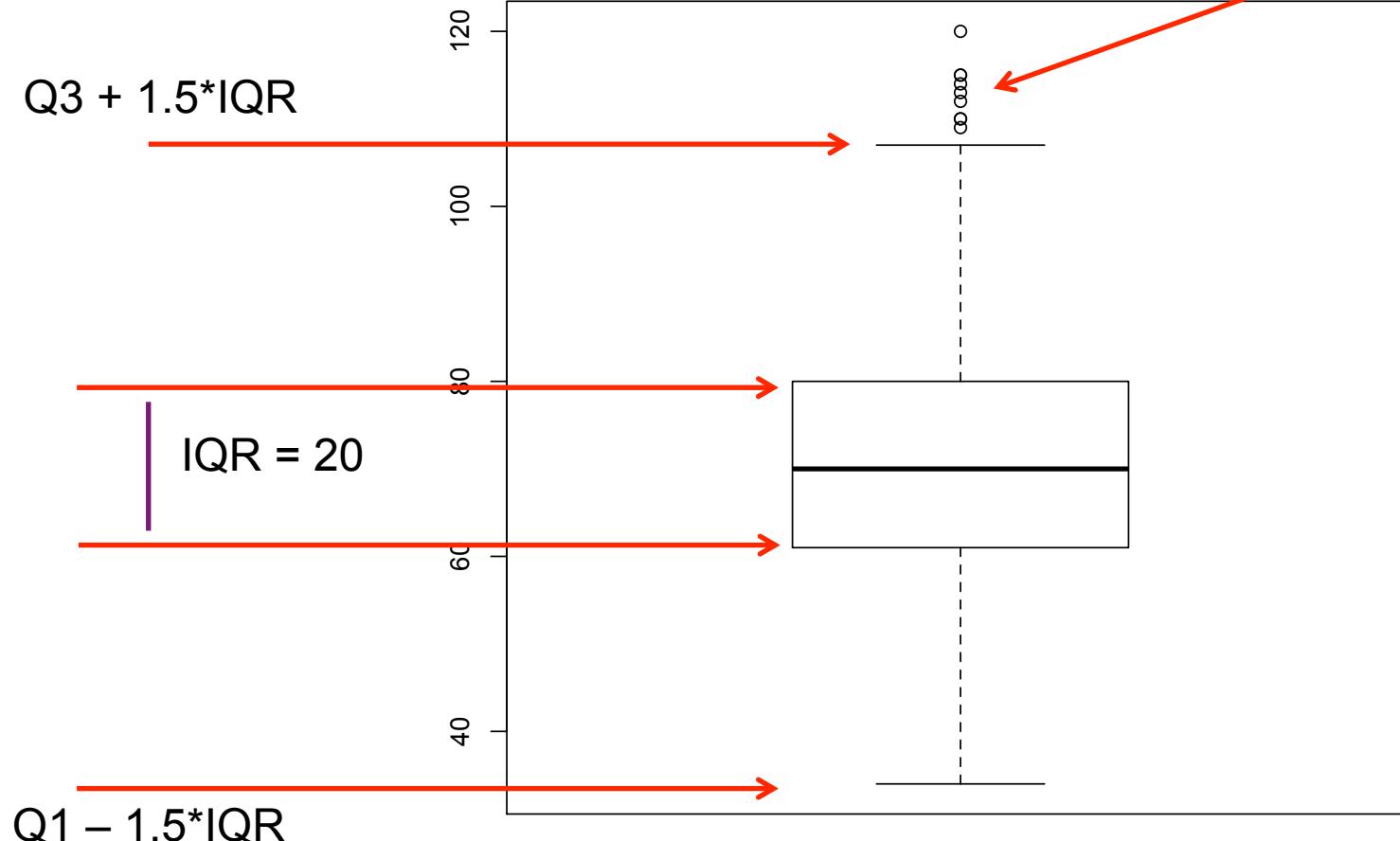
`boxplot(wt)`



IQR = interquartile range

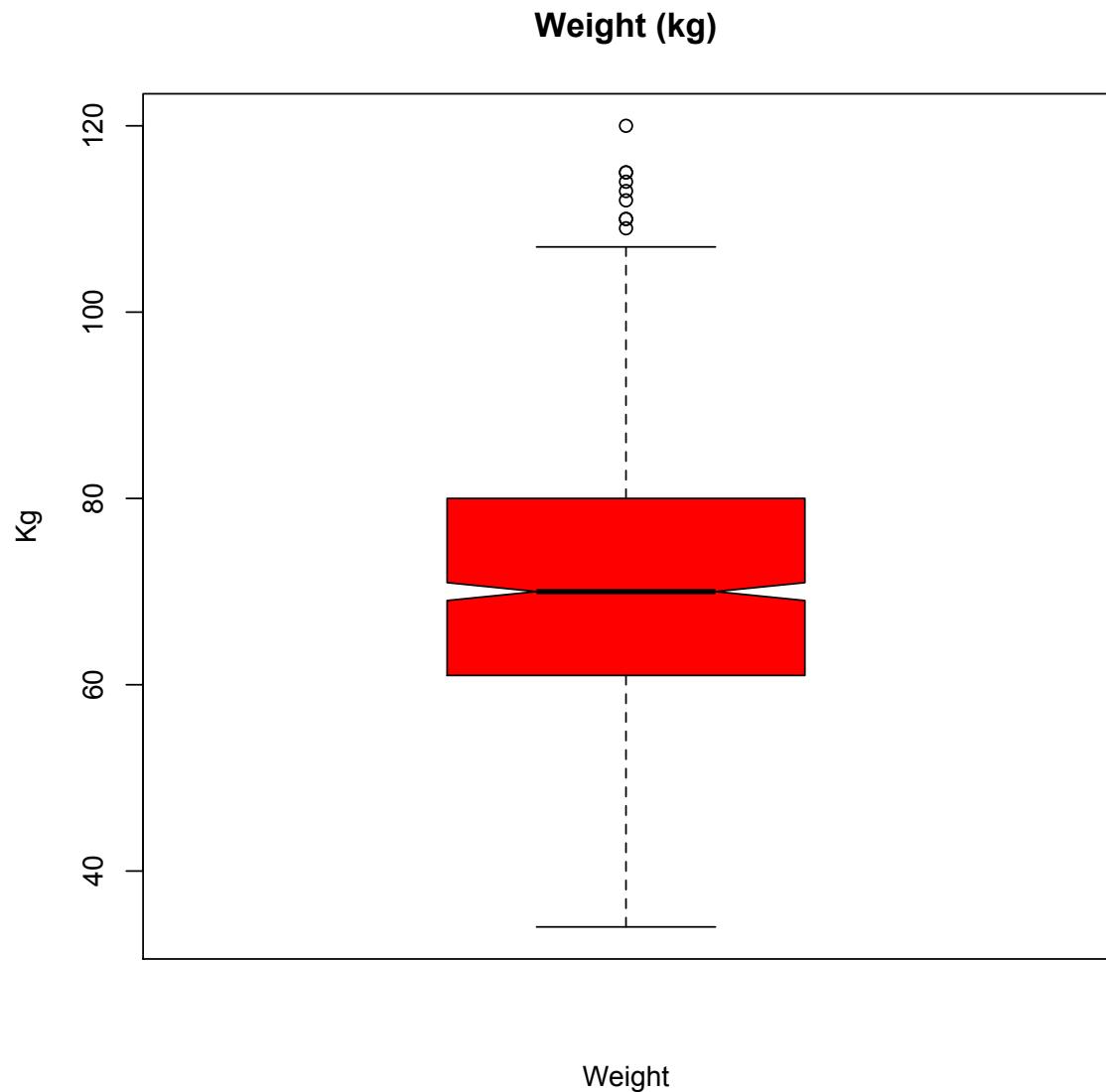
Box plot

boxplot(wt)



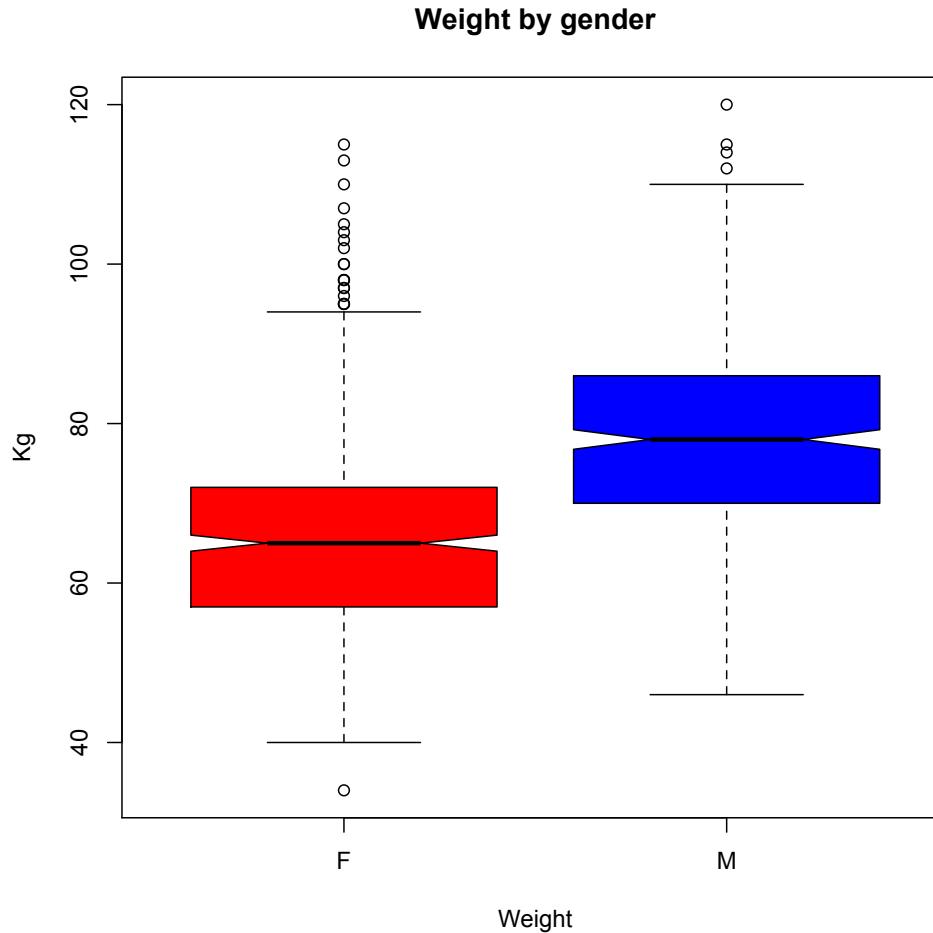
IQR = interquartile range

```
boxplot(wt, main="Weight (kg)",  
xlab="Weight", ylab="Kg", col="Red",  
notch=TRUE)
```



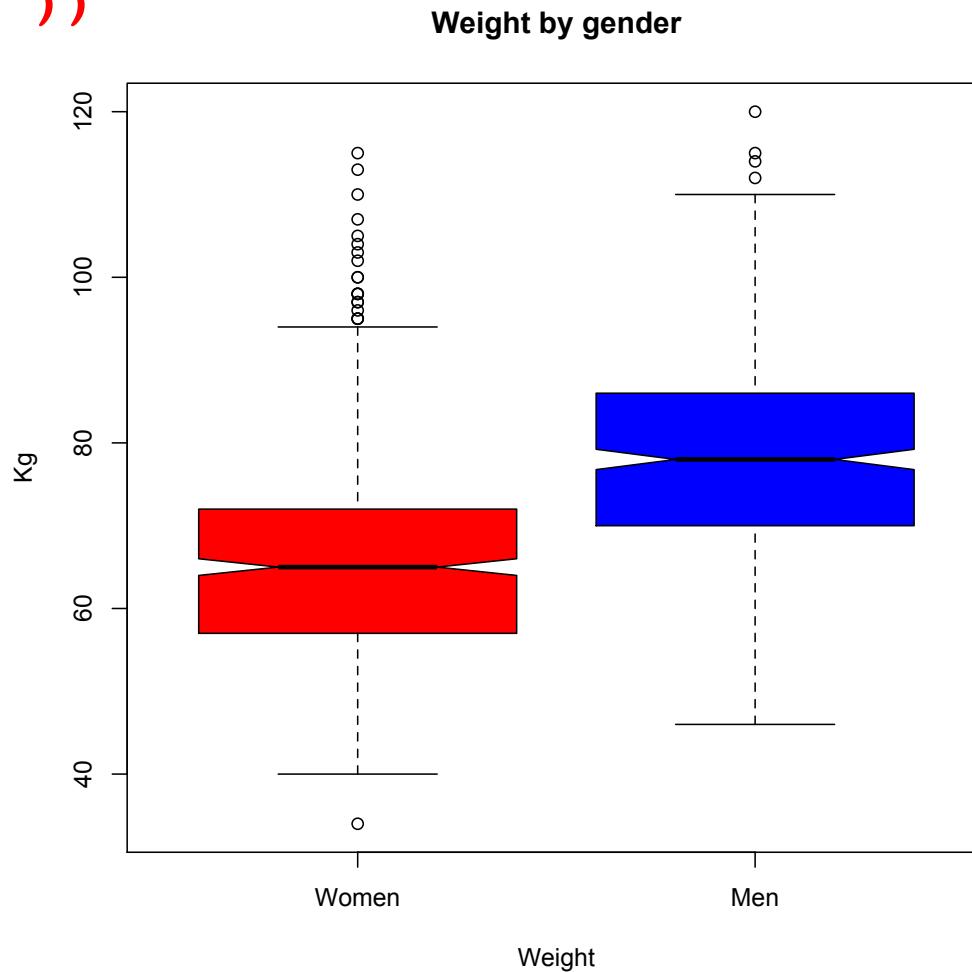
Box plot theo nhóm (giới tính)

```
boxplot(wt~sex, main="Weight by gender", xlab="Weight",  
ylab="Kg", notch=T, col=c("red","blue"))
```

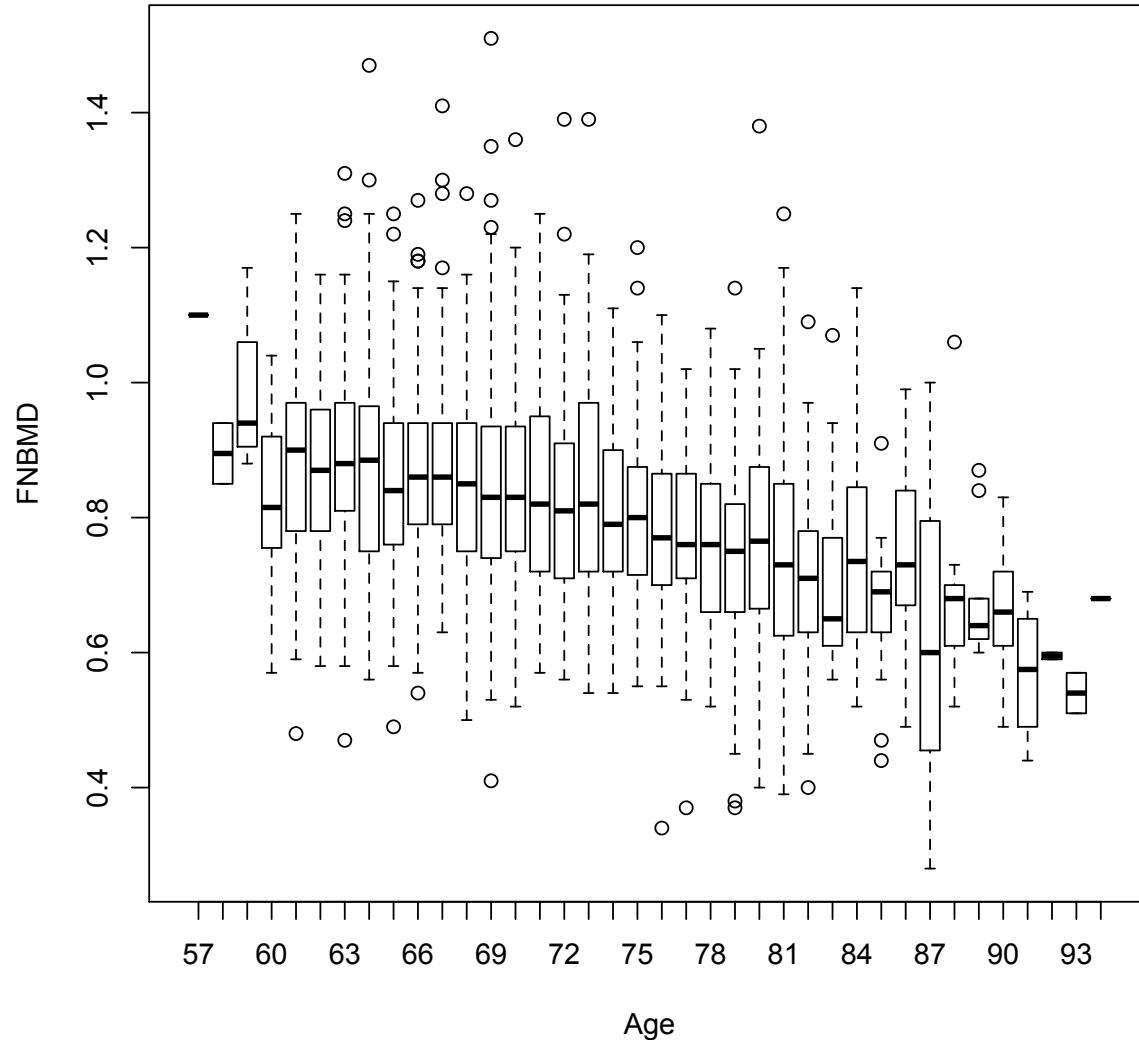


Box plot theo nhóm (giới tính)

```
boxplot(wt~sex, main="Weight by gender", xlab="Weight",  
ylab="Kg", notch=T, col=c("red","blue"),  
names=c("Women", "Men") )
```



```
os=read.csv("http://statistics.vn/data/  
does_vn07.csv",header=T)  
attach(os)  
boxplot(fnbmd ~ age, xlab="Age", ylab="FNBMD")
```

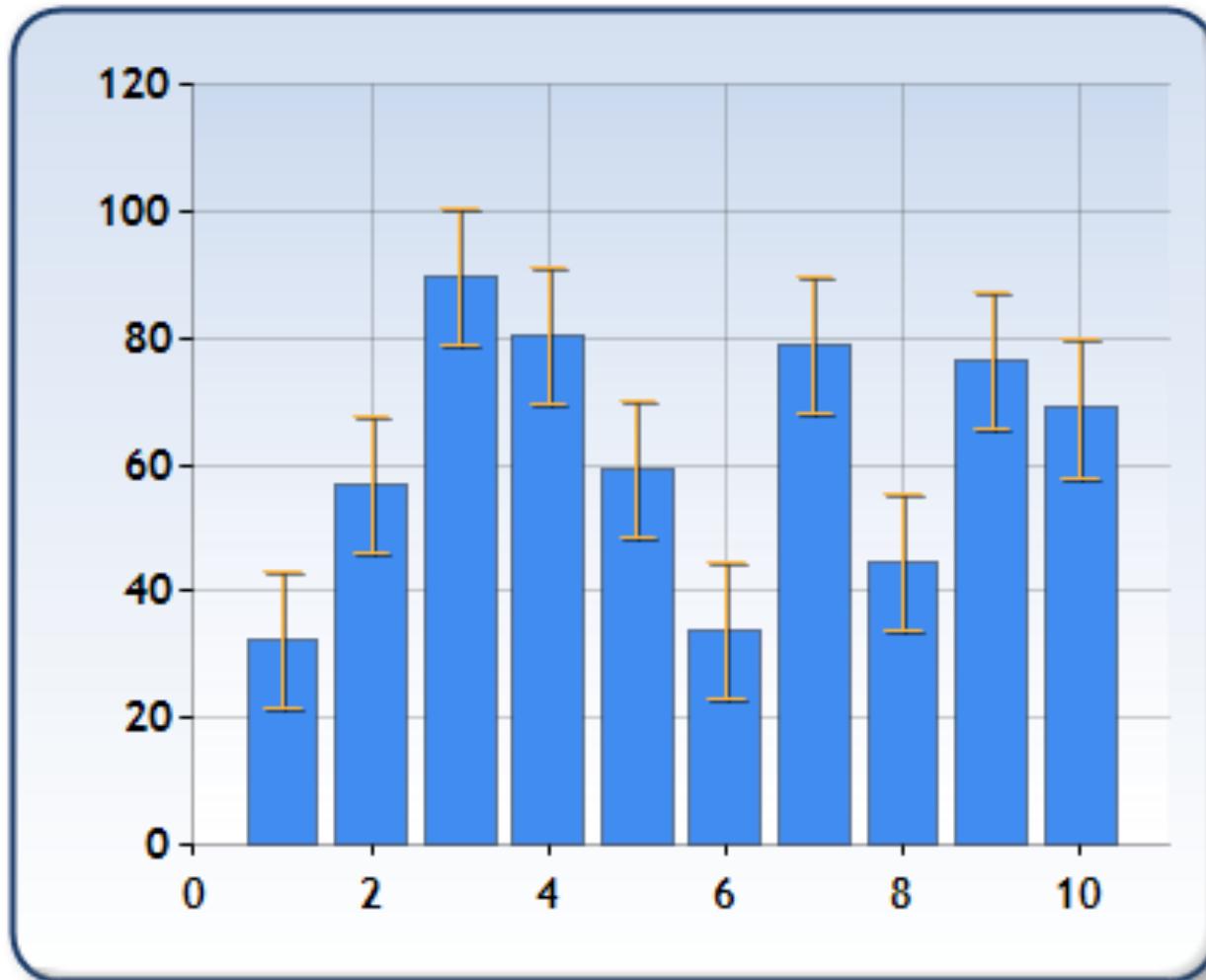


Tóm lược biểu đồ hộp

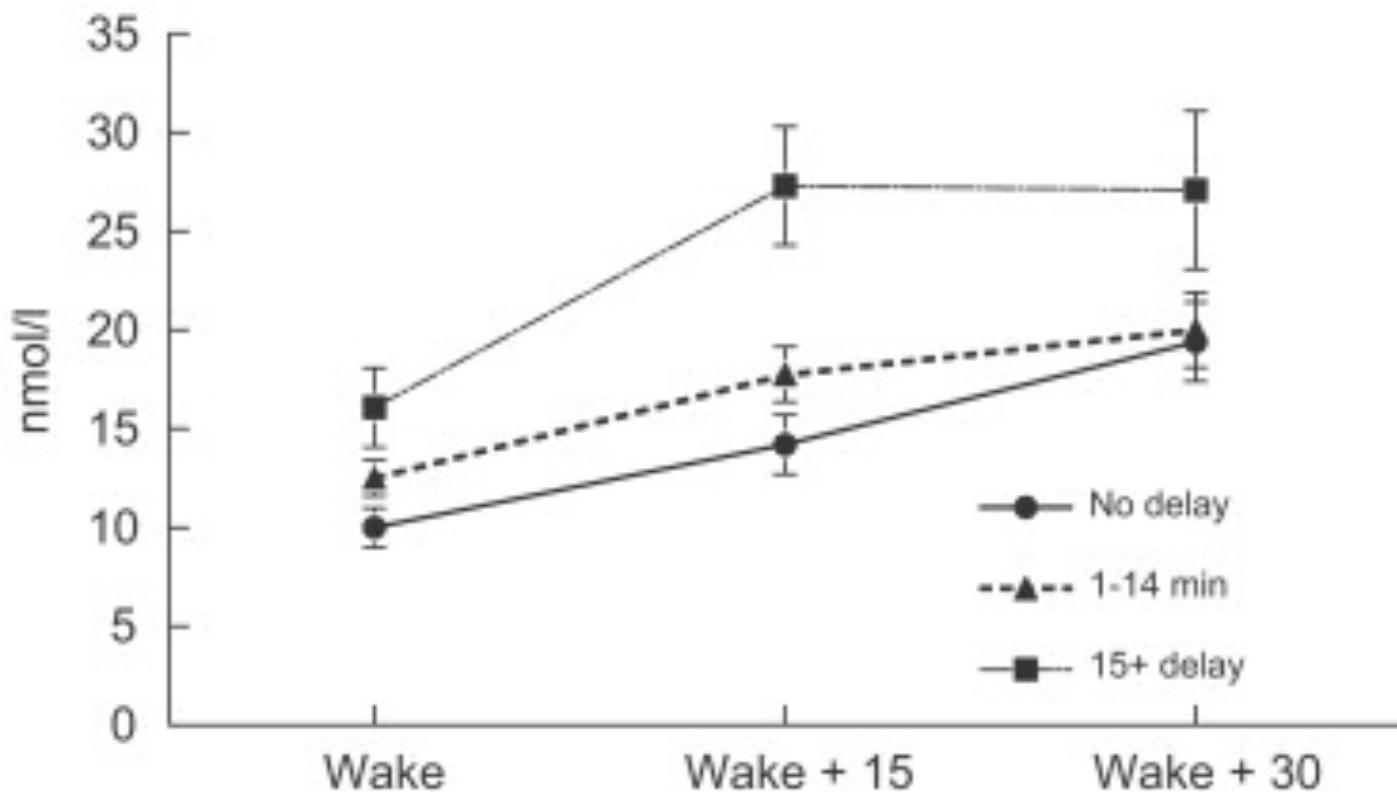
- Biểu đồ hộp (John Tukey)
- Biểu đồ hộp là một phương tiện rất tốt cho mô tả biến liên tục
- Càng ngày càng phổ biến trong bài báo khoa học

Biểu đồ sai số chuẩn (error bar plot)

Biểu đồ "error bar"



Biểu đồ "error bar"

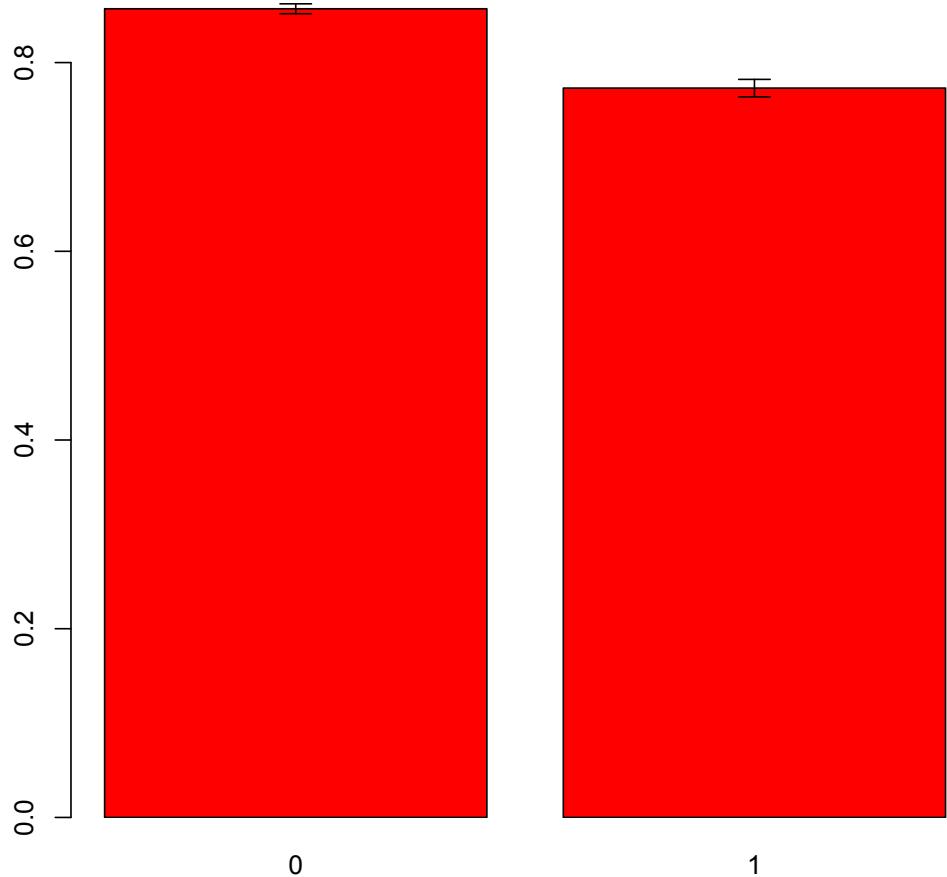


Mục đích

- Mô tả phân bố dữ liệu (distribution of data)
 - So sánh hai nhóm với sai số chuẩn
- cần package "**sciplot**"

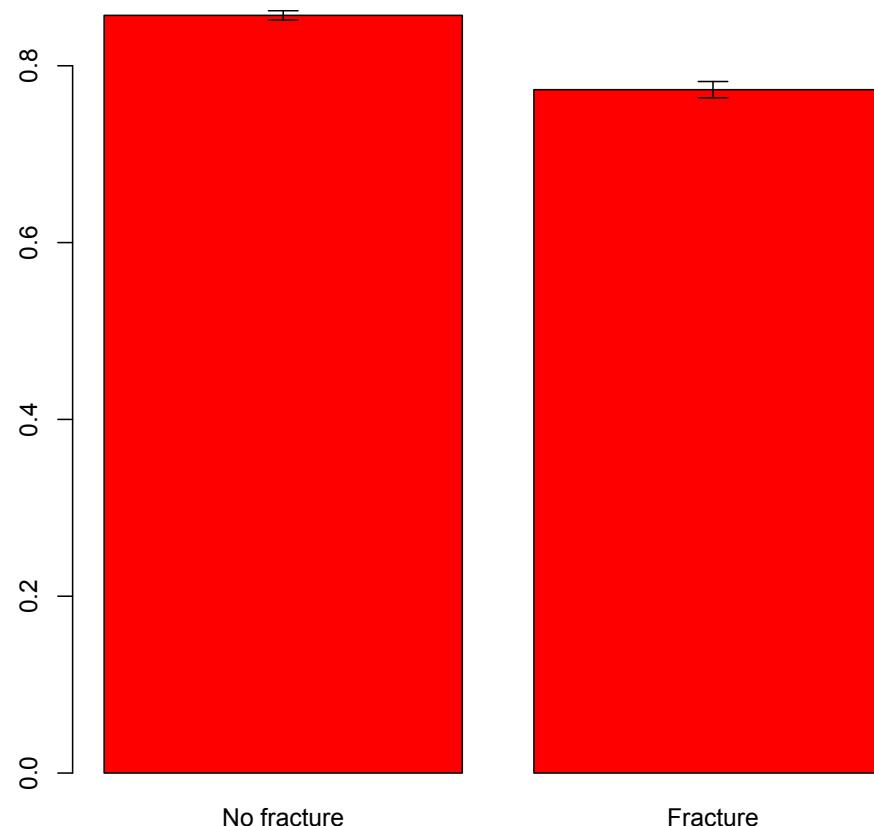
bargraph.CI

```
package sciplot  
library(sciplot)  
bargraph.CI(anyfx, fnbmd, col="red")
```



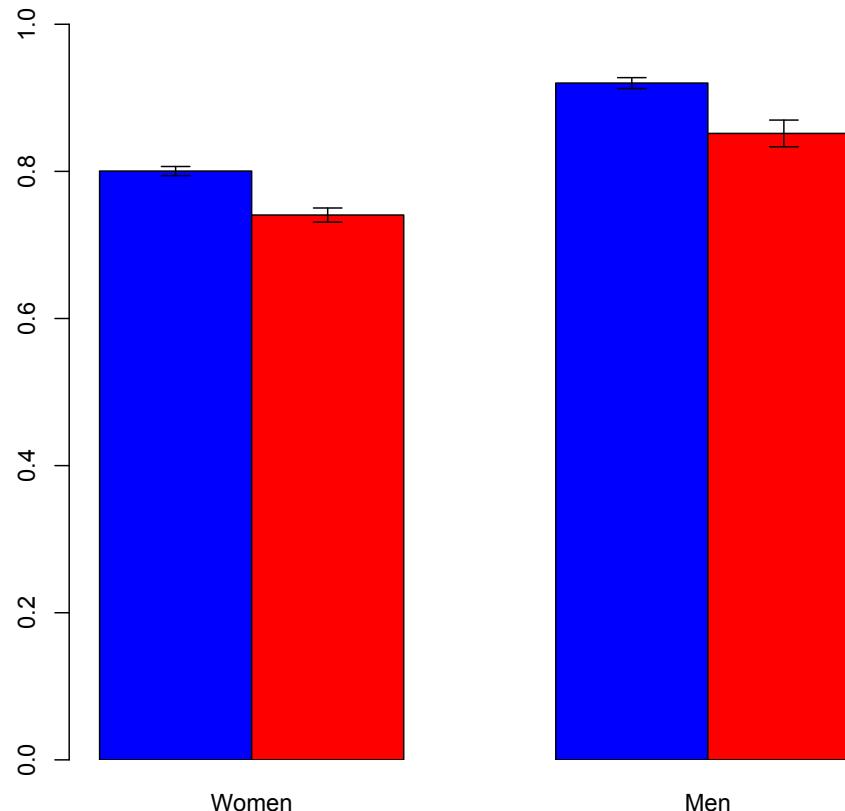
So sánh: bar plot

```
bargraph.CI(gender, fnbmd, anyfx, col="red", names=c("No fracture", "Fracture"))
```



So sánh: bar plot

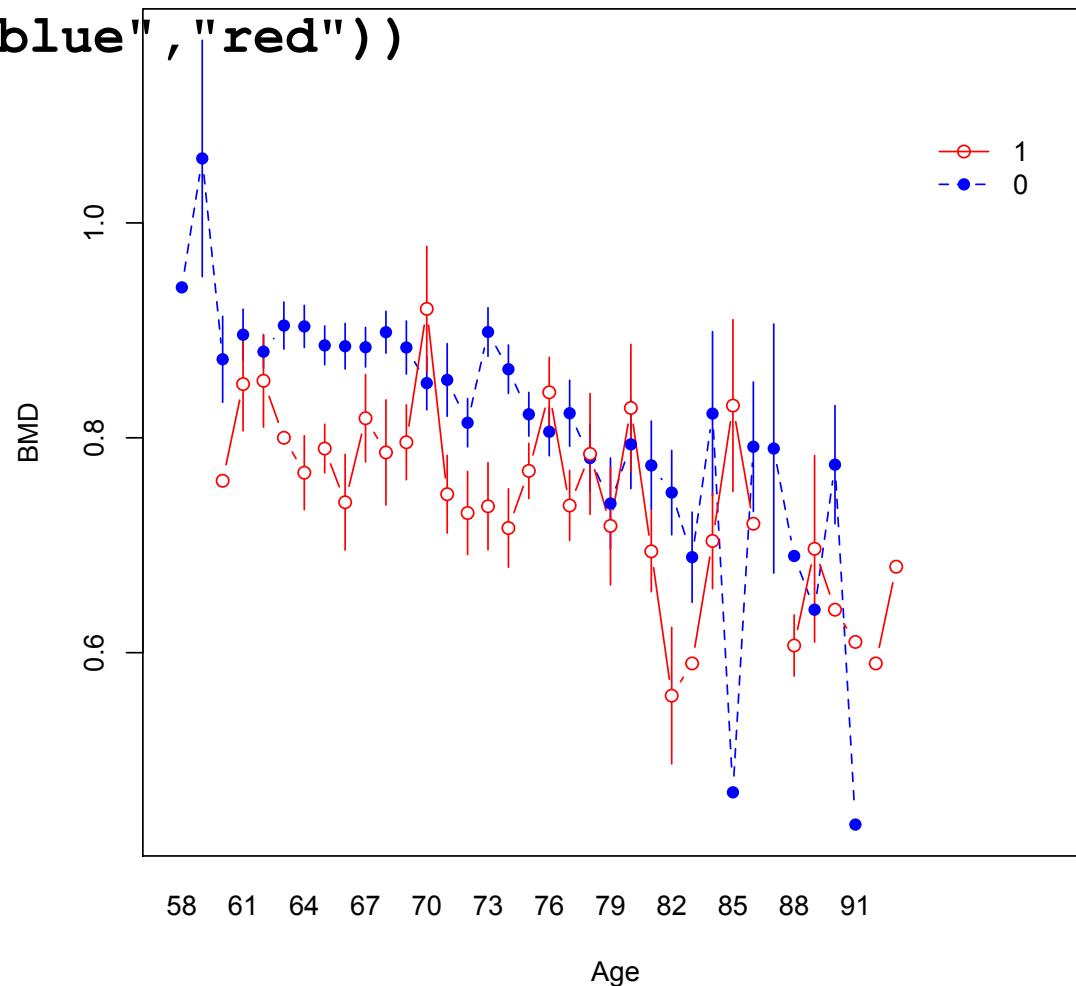
```
bargraph.CI(sex, fnbmd, group=anyfx,  
col=c("blue","red"), names=c("Women", "Men"),  
ylim=c(0,1))
```



So sánh: line plot

```
library(sciplot)

lineplot.CI(age, fnbmd, anyfx, ylab="BMD", xlab="Age",
legend=F, col=c("blue","red"))
```



error.bars.by trong psych

```
error.bars.by(fnbmd, gender,  
bars=T,      ylim=c(0,1.3),  
labels=c("Women", "Men") ,  
col=c("Blue", "Blue") )
```

Tóm lược biểu đồ sai số chuẩn

- Hai giá trị trung bình có thể thể hiện bằng bar plot
- Nhưng phải có sai số chuẩn (tính khoa học!)
- Package "sciplot" → rất có ích!

Tóm lược

- Bốn dạng biểu đồ phổ biến
 - Biểu đồ phân bố: hist()
 - Biểu đồ tương quan: plot(), scatterplot()
 - Biểu đồ hộp: boxplot()
 - Biểu đồ thanh: barplot()