Bài giảng 8a: Phân tích bằng biểu đồ phân bố (histogram)

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Ton Duc Thang University, Vietnam

Biểu đồ phân bố

- Biểu đồ cơ bản nhất trong các biểu đồ
- Mô tả phân bố của một biến số (thường là biến liên tục)
- Có thể rút ra vài chỉ số thống kê từ biểu đồ
- Có thể dùng để kiểm tra các giá trị ngoại vi

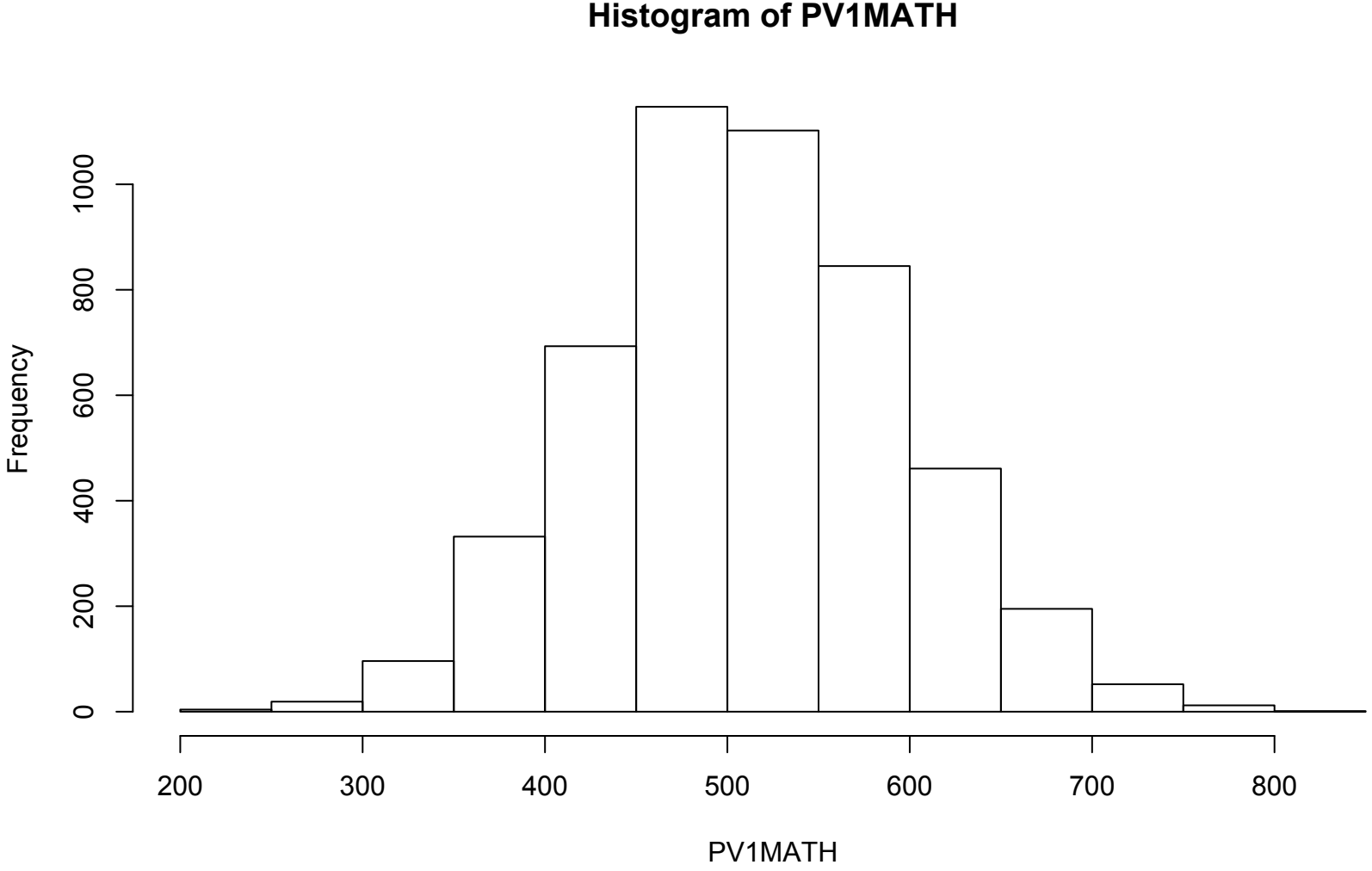
Dữ liệu PISA (Schools)

```
setwd("~/Dropbox/World Bank 2014/Data for 2015
  workshop")

pisa = read.csv("~/Dropbox/World Bank 2014/Data for 2015
  workshop/PISA DATA temp.csv", header=T)

sc = read.csv("~/Dropbox/World Bank 2014/Data for 2015
  workshop/SCHOOL DATA (VN).csv", header=T)

dat = merge(sc, pisa, by="SCHOOLID")

attach(dat)
```

Mô tả biến điểm môn toán

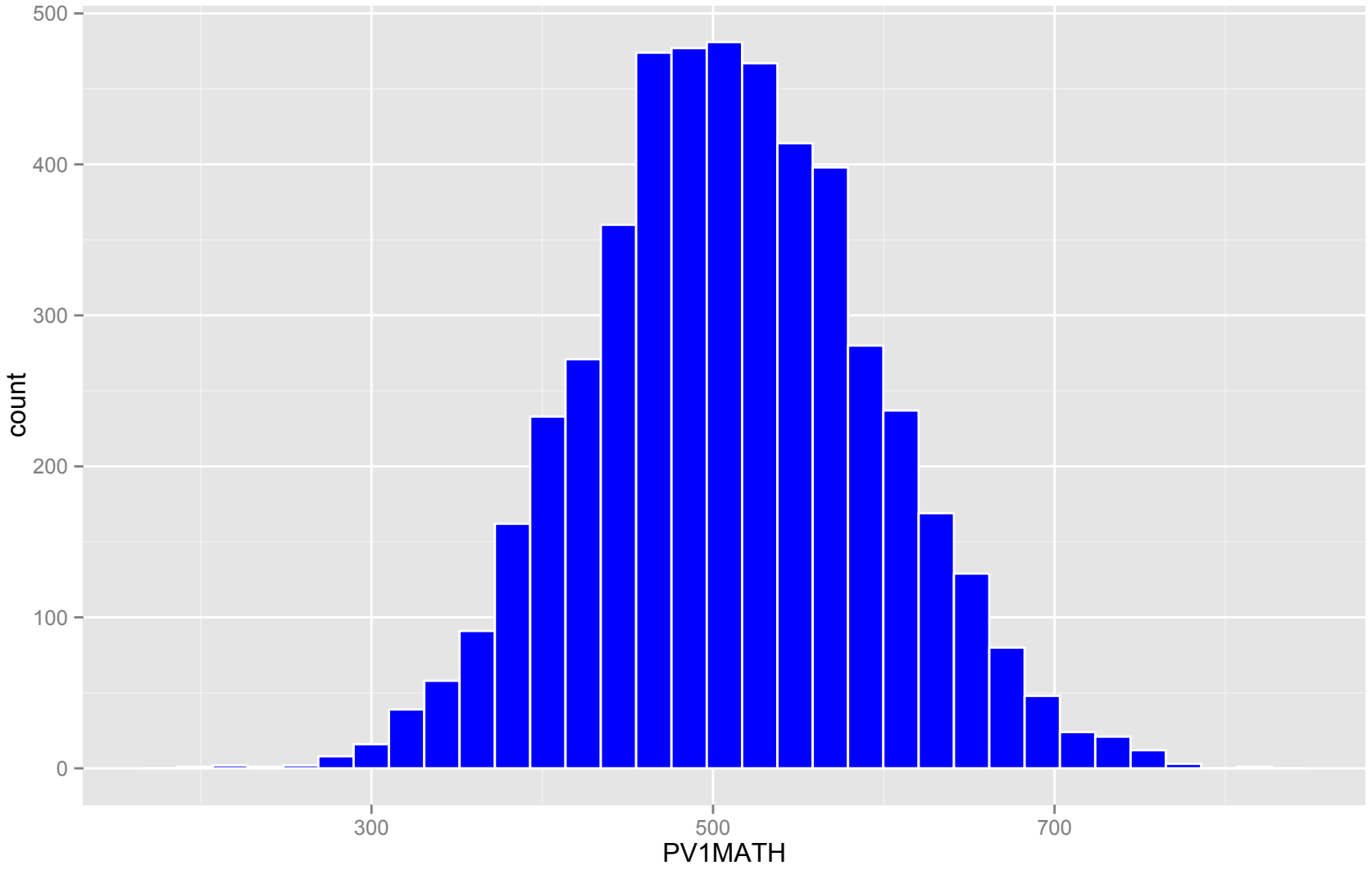
- PV1MATH: điểm môn toán
- Chúng ta muốn mô tả biến này bằng biểu đồ phân bố
- Có thể dùng

```
hist(PV1MATH)
```


Histogram of PV1MATH
Frequency
0
200
400
600
800
1000
200
300
400
500
600
700
800
PV1MATH

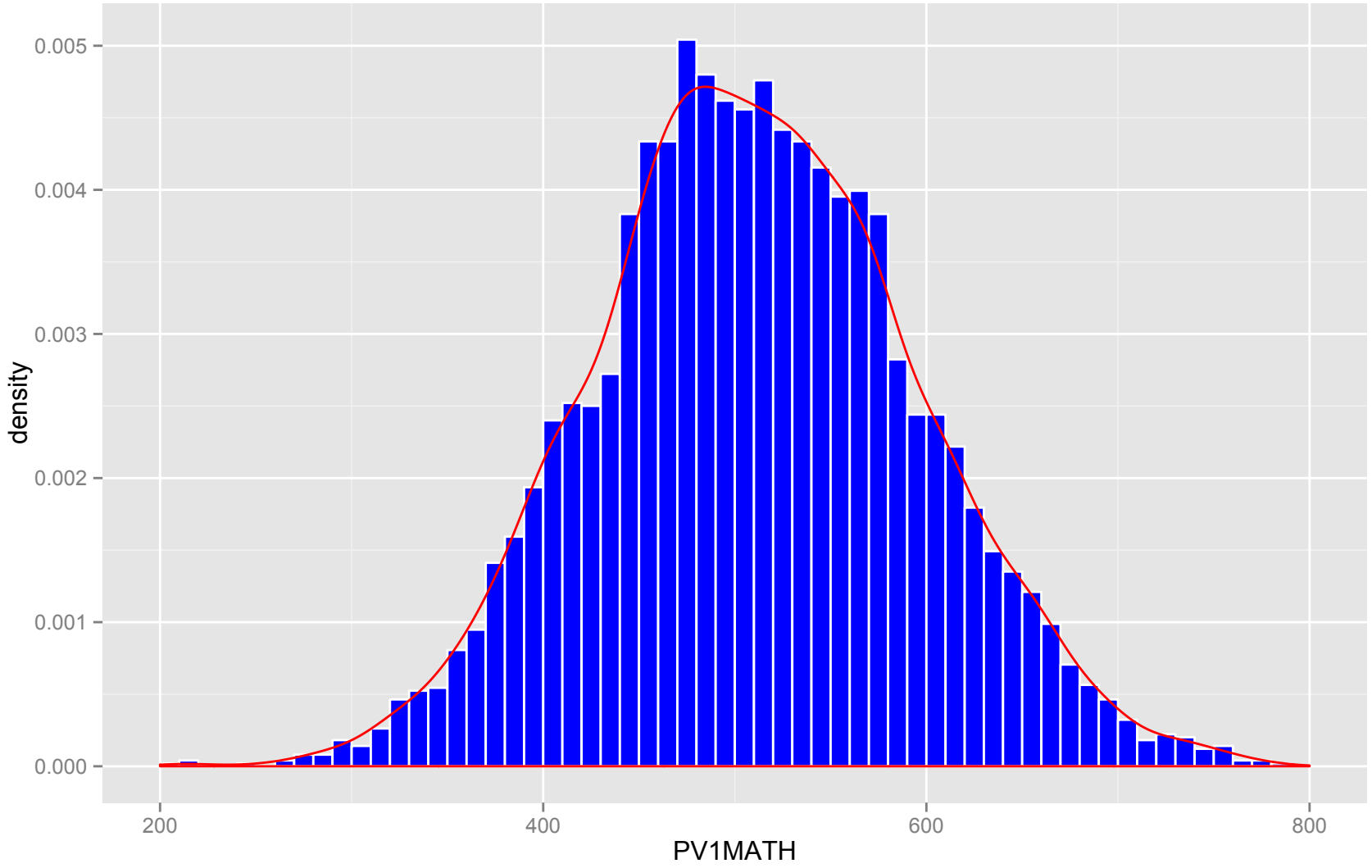
Dùng ggplot2 để vẽ biểu đồ phân bố

```
attach(dat)

library(ggplot2); library(gridExtra)

p = ggplot(dat, aes(x=PV1MATH))

p1 = p + geom_histogram(color="white",
fill="blue")
```


500
400
300
200
100
0
count
300
500
700
PV1MATH

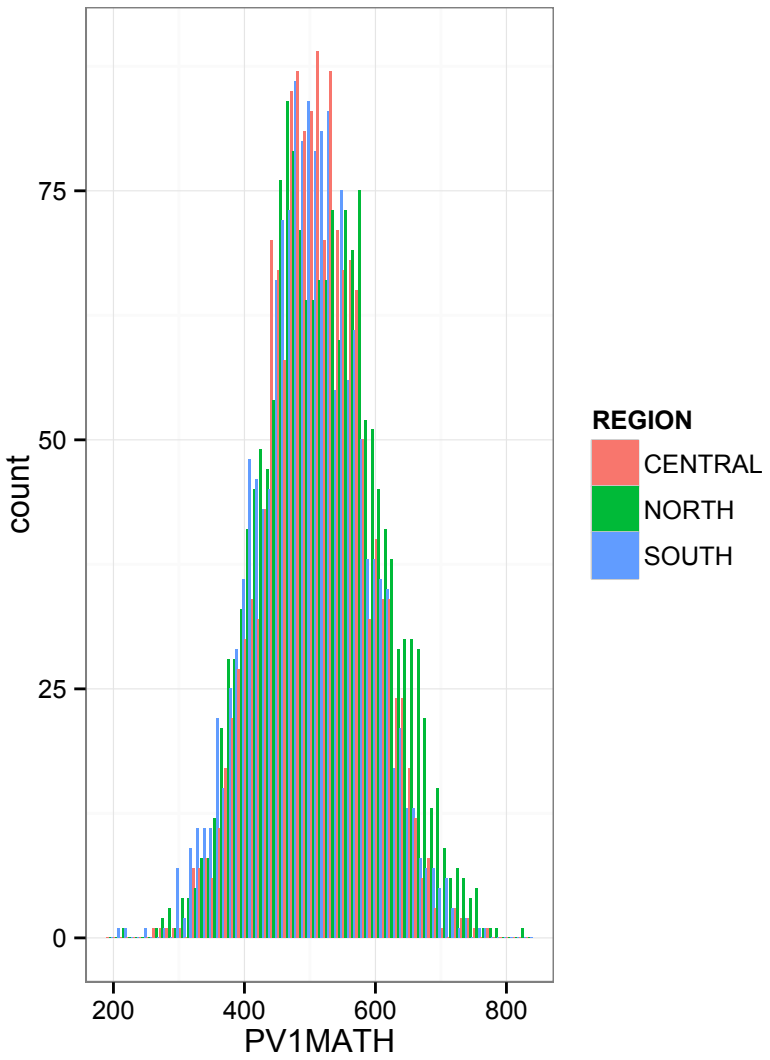
Thêm đường density

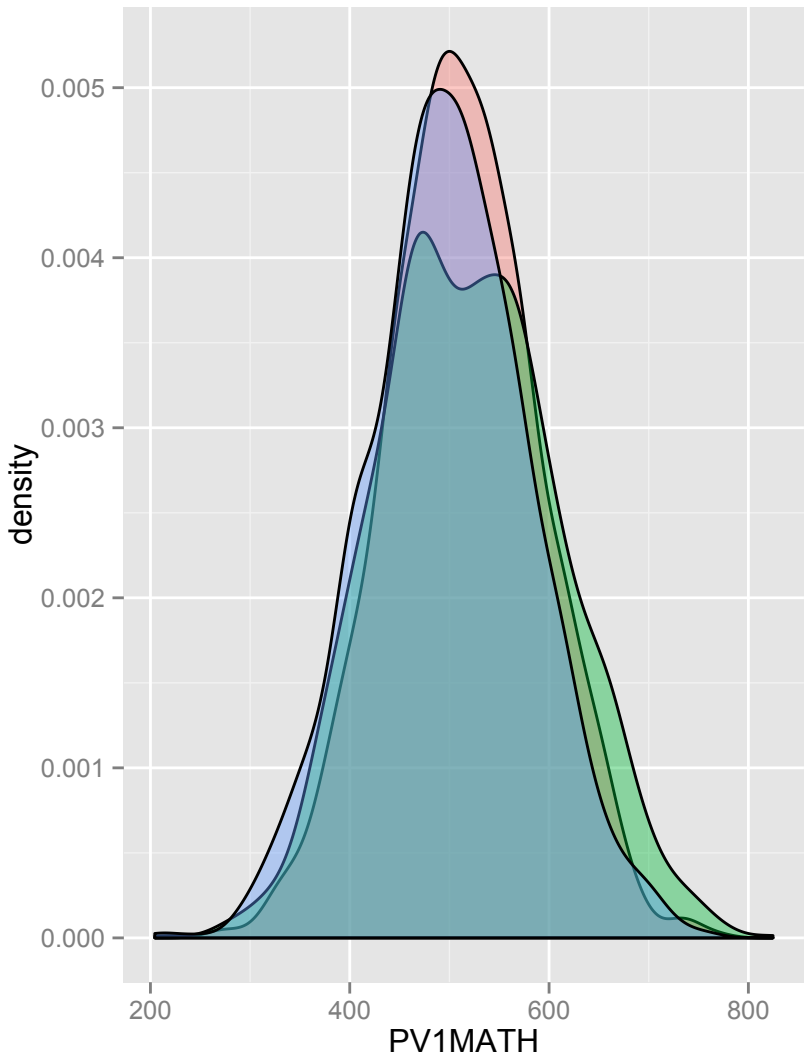
```
attach(dat)

library(ggplot2); library(gridExtra)

p = ggplot(dat, aes(x=PV1MATH))

p2 = p + geom_histogram(aes(y=..density..),
binwidth=10, color="white", fill="blue") +
geom_density(alpha=0.5, col="red") +
xlim(200,800)

p2
```


0.005
0.004
0.003
0.002
0.001
0.000
density
200
400
600
800
PV1MATH

Phân bố theo nhóm (REGION)

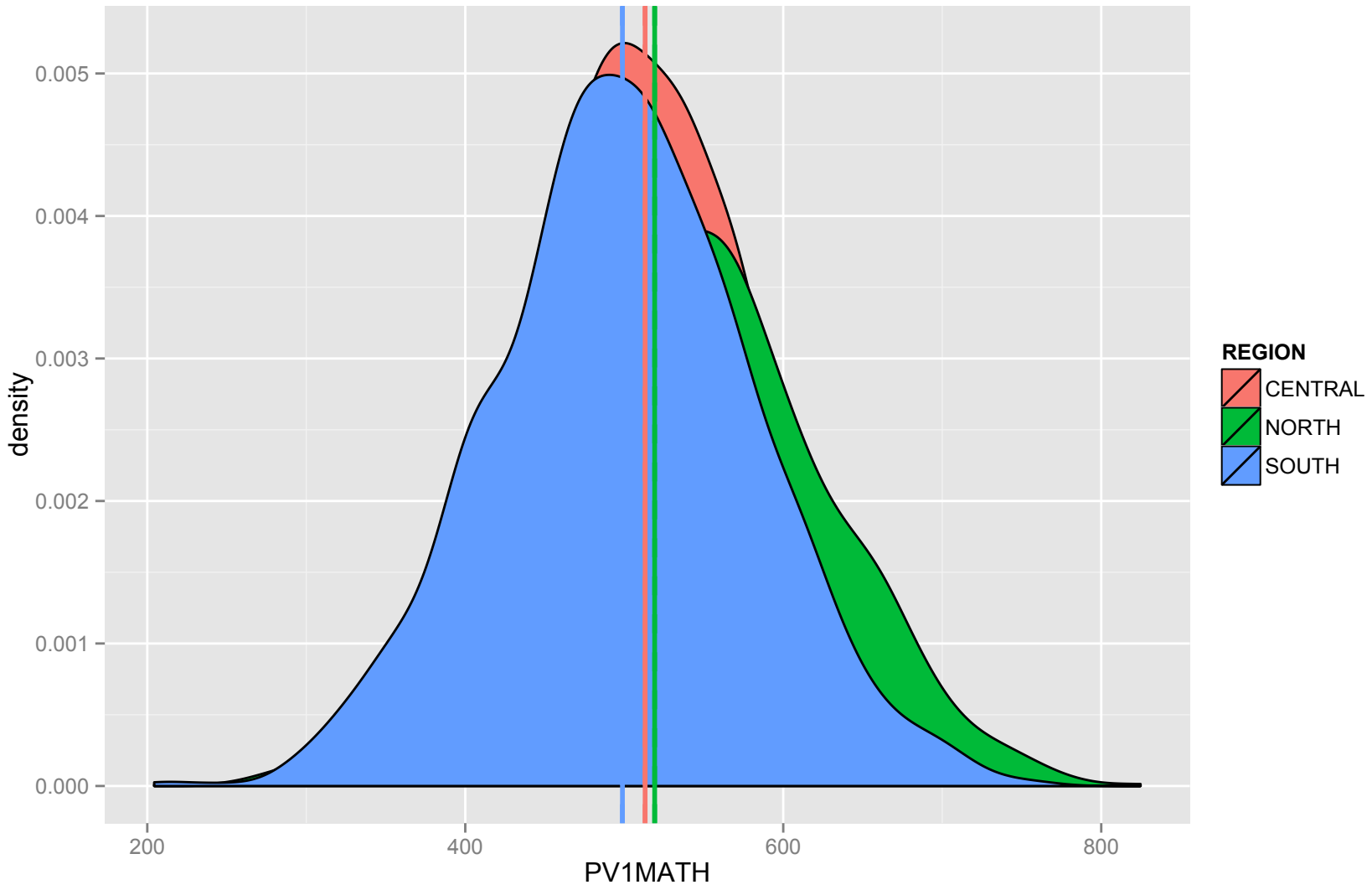
```
p = ggplot(dat, aes(x=PV1MATH, fill=REGION))
p1 = p + geom_histogram(binwidth=10,
position="dodge") +
theme(legend.position="top") + theme_bw()
p2 = p + geom_density(alpha=0.4) +
theme(legend.position="none")


library(gridExtra)
grid.arrange(p1, p2, ncol=2)
```


count
75
50
25
0
200
400
600
800
PV1MATH
REGION
CENTRAL
NORTH
SOUTH


density
0.005
0.004
0.003
0.002
0.001
0.000
200
400
600
800
PV1MATH

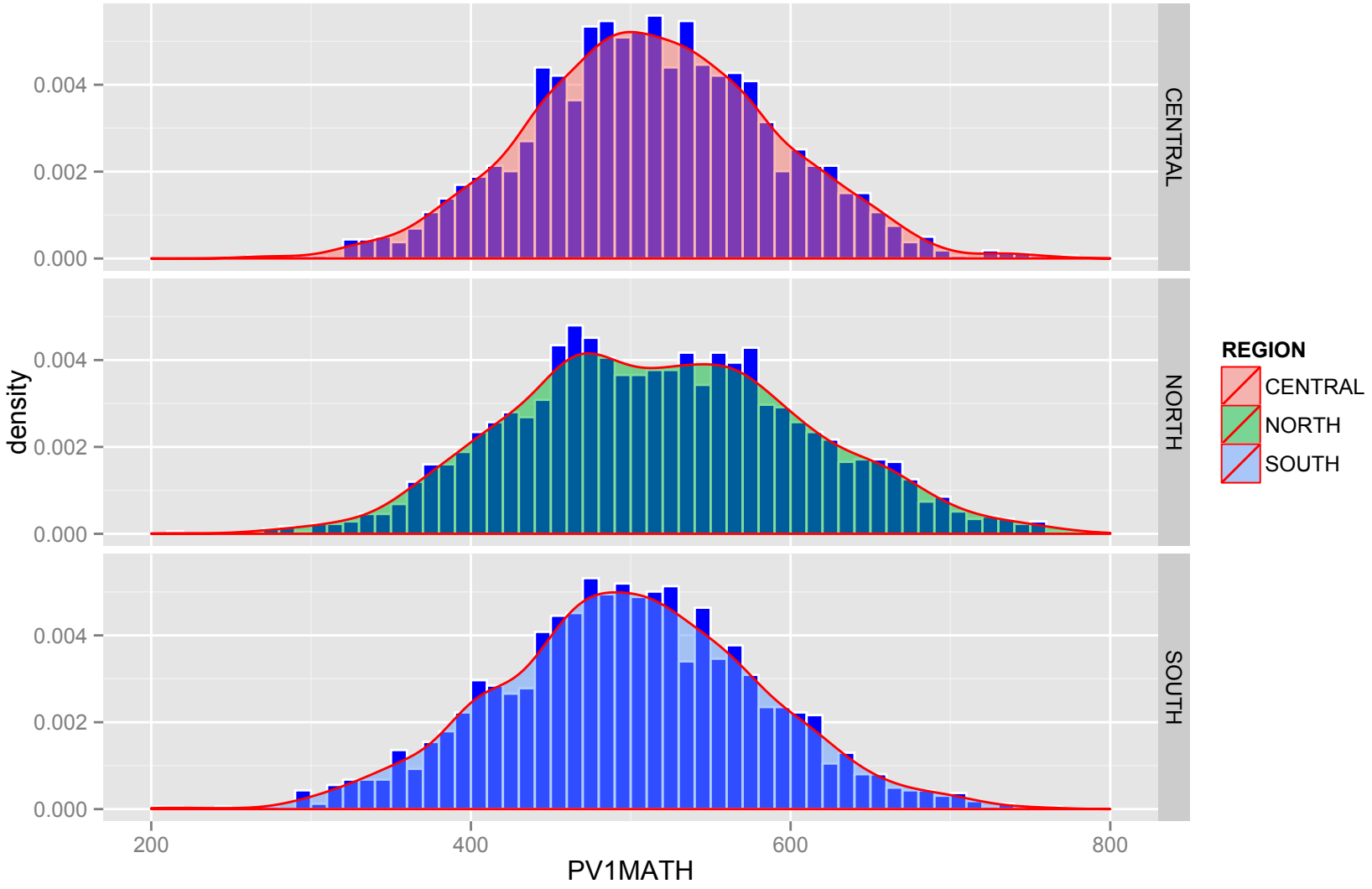
Vẽ đường trung bình

```
sdat = aggregate(dat$PV1MATH, by=list(dat
$REGION), FUN=mean)

colnames(sdat) = c("Group", "Mean")

p = ggplot(dat, aes(x=PV1MATH, fill=REGION))

p = p + geom_density()

p = p + geom_vline(data=sdat,
aes(xintercept=Mean, colour=Group), size=1)
```


0.005
0.004
0.003
0.002
0.001
0.000
density
200
400
600
800
PV1MATH
REGION
CENTRAL
NORTH
SOUTH

Tách ra thành nhiều panel

```
p = ggplot(dat, aes(x=PV1MATH, fill=REGION))
p2 = p + geom_histogram(aes(y=..density..),
binwidth=10, color="white", fill="blue") +
geom_density(alpha=0.5, col="red") +
xlim(200,800) + facet_grid(REGION ~.)
```


density
0.004
0.002
0.000
CENTRAL
NORTH
SOUTH
200
400
600
800
PV1MATH
REGION
CENTRAL
NORTH
SOUTH