

# boxplot

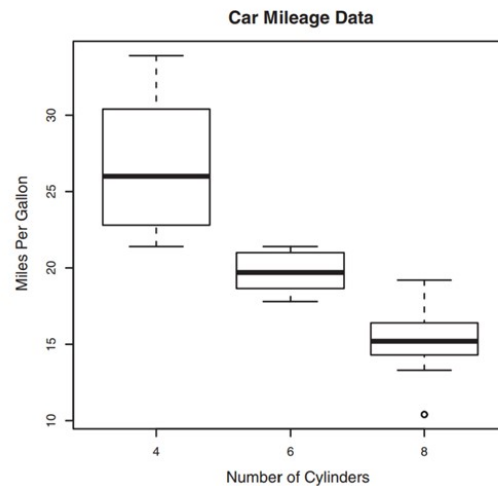
连续+分类

比直方图数据更多 (?)

## Boxplot (箱线图)

箱线图可以分类来做

```
boxplot(mpg ~ cyl, data=mtcars,
        main="Car Mileage Data",
        xlab="Number of Cylinders",
        ylab="Miles Per Gallon")
```



清华大学统计学研究中心

可看出每个组的 **平均** (用中位表示 (?))



- 没有outlier
  - 数据量太小
  - 数据比较集中



中位数以上的数据都一样

# spineplot

分类+分类 (两个, 都用频数表示)

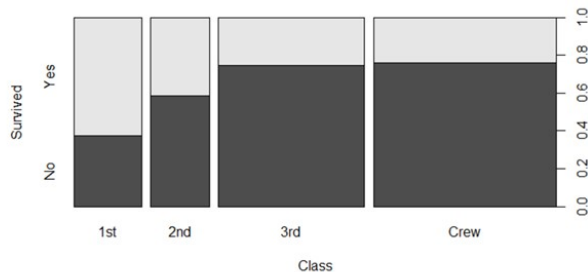
一种特殊的条形图 (堆叠)

# Spineplot (棘状图)

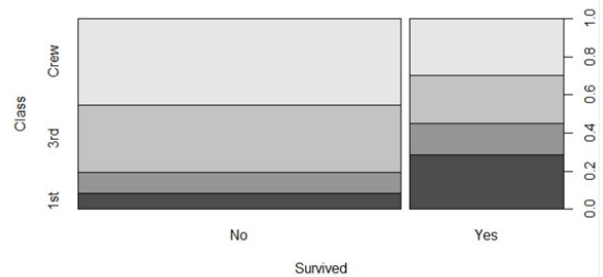
28

Class	Survived	
	No	Yes
1st	122	203
2nd	167	118
3rd	528	178
Crew	673	212

spineplot(dat)



spineplot(t(dat))



清华大学统计学研究中心

高度相同，通过宽度表示

左图中Yes/No（响应变量）所占比例一目了然

## dot plot

连续型

## scatter plot

两个连续变量之间的关系，可能线性可能非线性

每一个点都是单独个体的数据

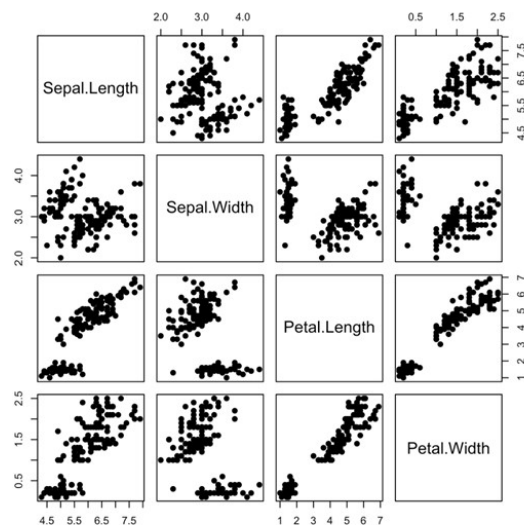
## scatter matrix

p个变量的两两关系，以矩阵形式排列，有 $p^2$ 个窗格

## Scatter matrix (散点图矩阵)

- 散点图的高维扩展
- 基本构成是多个变量的两两散点图以矩阵的形式排列起来
- p个变量通常有 $p \times p$ 个窗格，便于查看变量间两两的关系

```
pairs(iris[,1:4], pch = 19)
```



清华大学统计学研究中心

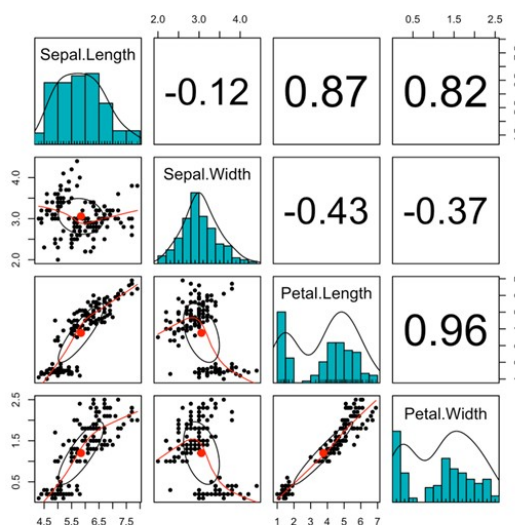
有对称性↑

可以只保留上半

## Scatter matrix (散点图矩阵)

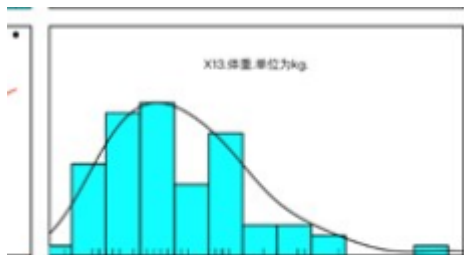
```
library(psych)
```

```
pairs.panels(iris[, -5], method = "pearson",  
# correlation method hist.col = "#00AFBB",  
density = TRUE, # show density plots  
ellipses = TRUE # show correlation ellipses)
```



清华大学统计学研究中心

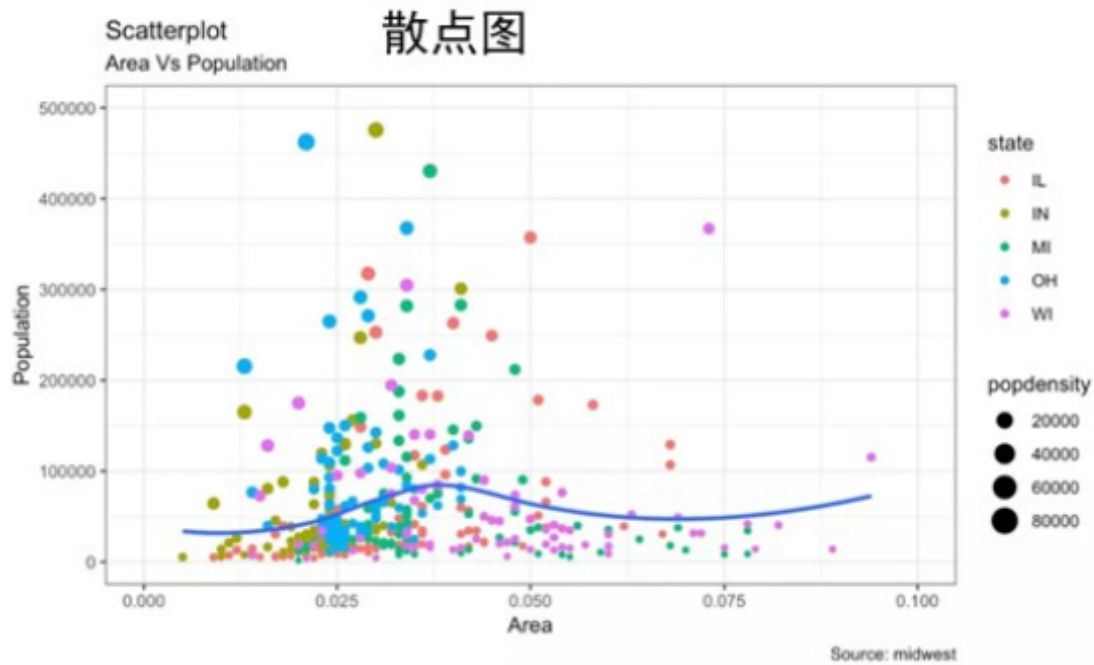
具体数据 (调用分析包)



直方图的拟合曲线转90°+对称→小提琴图

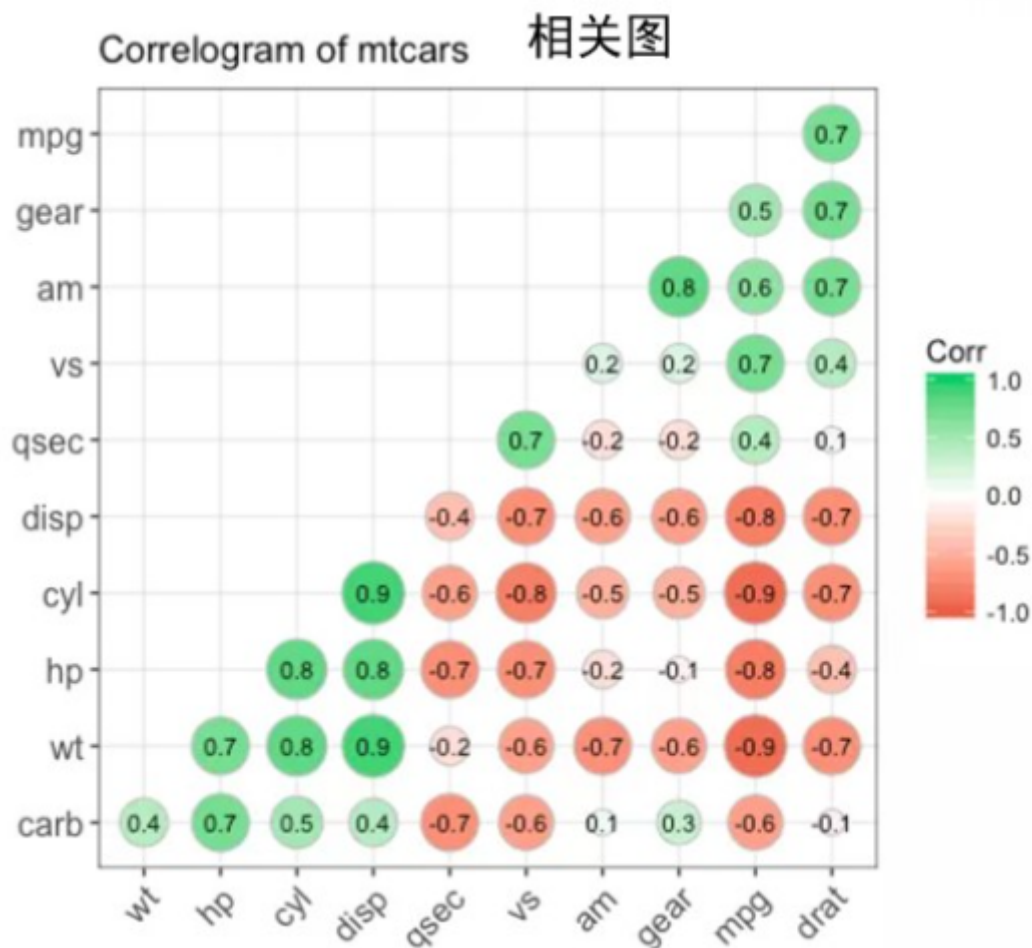
## how to use graphs

### 相关性



用圈的大小表示数量多少

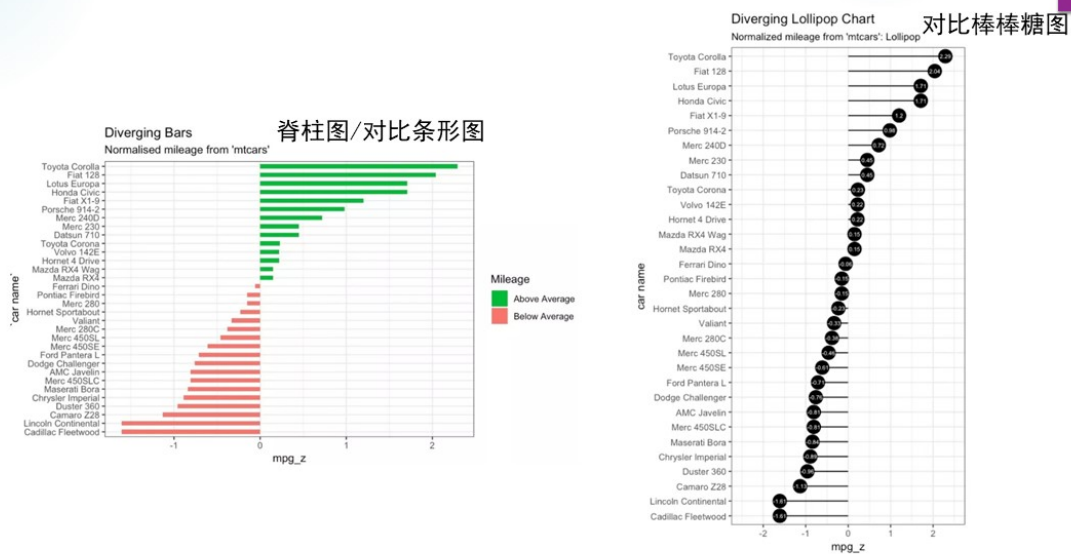
相关图 数据为零表示没有相关性没有线性相关性



## 偏差

### 2.2 数据呈现之图形 类型二：偏差

46

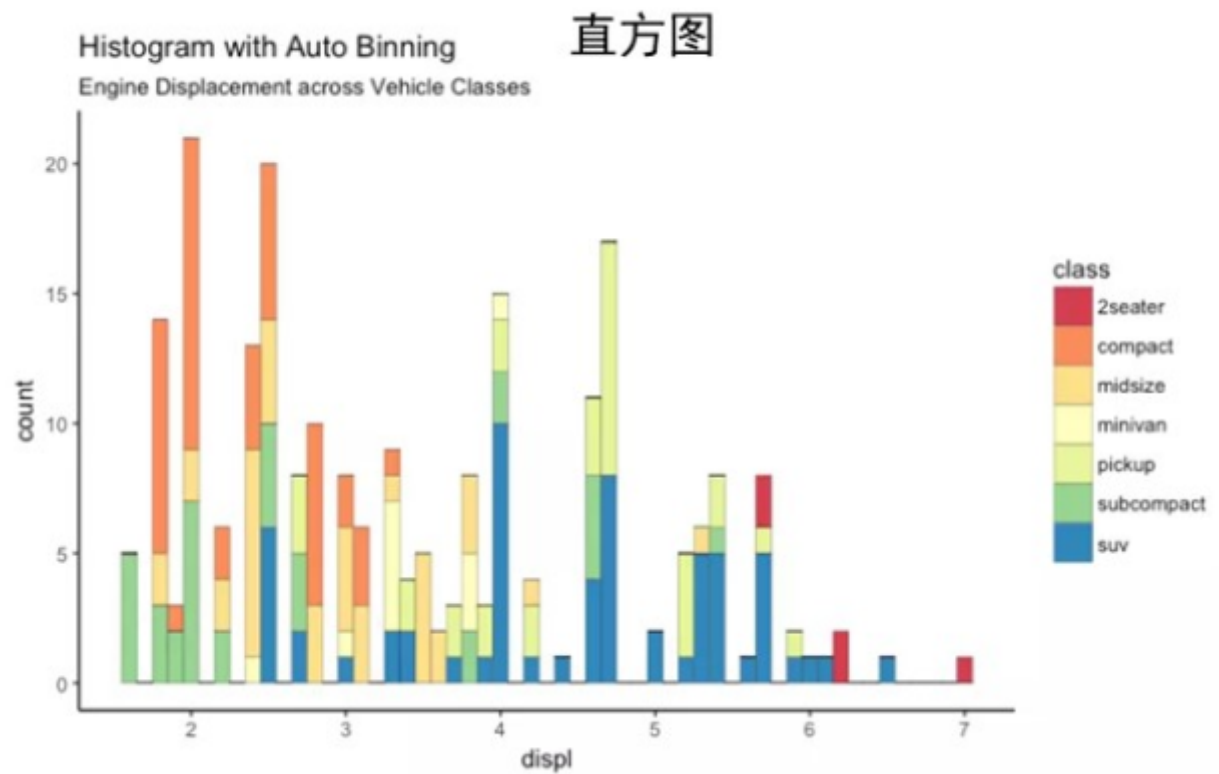


清华大学统计学研究中心

条的长度不代表数值大小

## 分布

直方图（比较乱不方便读）



密度图

箱线图

小提琴图

## 组成

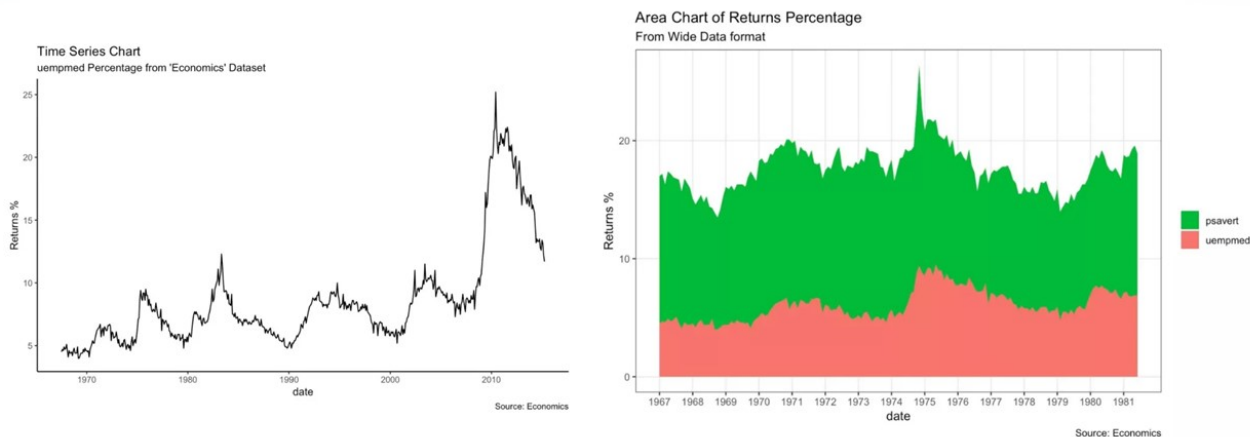
饼图 太细的不方便看

条形图 优选

## 时间序列

## 2.5 数据呈现之图形 类型五：时间序列

50



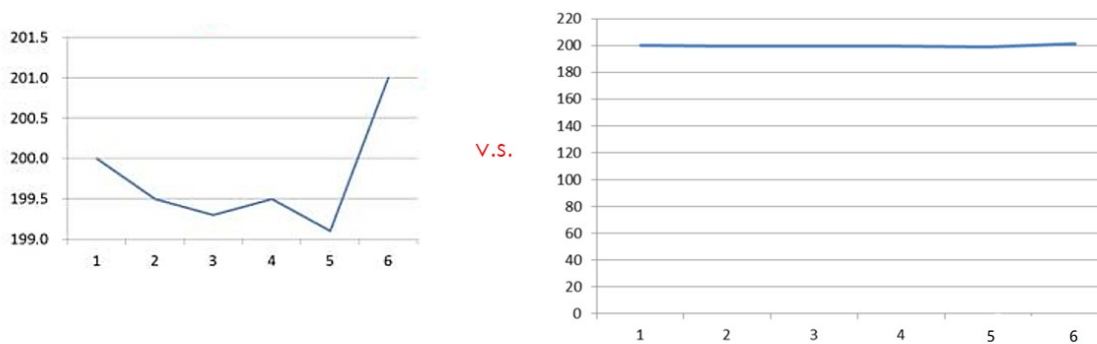
清华大学统计学研究中心

3D图有很多冗余的数据，高度也不方便看（角度）  
纵坐标一定要从零开始，避免夸大效果

## 2.6 你被图形骗到了吗

52

纵坐标一定要从零开始

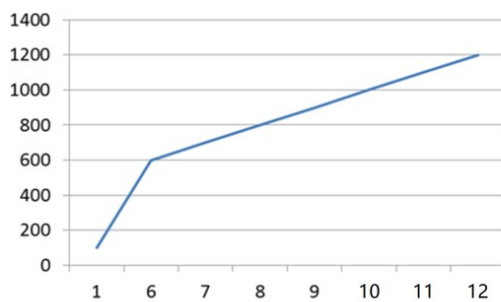


清华大学统计学研究中心

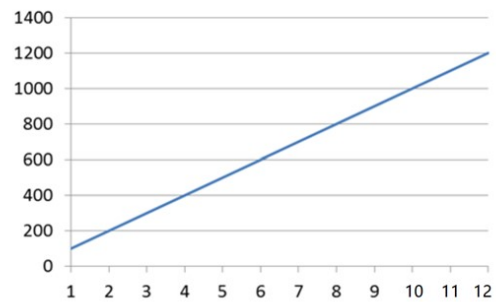
## 2.6 你被图形骗到了吗

53

横坐标不要随便压缩



V.S.



清华大学统计学研究中心

变化其实是稳定的（横坐标只能 等比例压缩）

## probabilities and distribution

### basic definition

- **outcome**  
1,2,3,4,5,6
- **event**  
set of outcomes eg 2 or 4 or 6
- **probability**  
无穷多次试验下的相对频数

$$P(E) = \frac{m}{N}.$$

### rules



## 1.2 Rules of Probability

5

### ► Basic:

(1)  $P(E) \geq 0$

(2)  $P(\Omega) = P(\omega_1 \cup \omega_2 \cup \dots \cup \omega_n) = 1$

(3)  $P(E \cup F) = P(E) + P(F)$  if E and F are **mutually exclusive**

### ► Useful:

Inclusion-Exclusion:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



清华大学统计学研究中心

## ★条件概率

### 全概率

## 1.3 Conditional Probability

6


- The probability of an event (A) occurring conditional (or given) that the event B occurs is:
$$P(A|B) = P(A \cap B) / P(B)$$

where  $P(B) \neq 0$

- Note that
$$Pr(B) = Pr(B|A)Pr(A) + Pr(B|\bar{A})Pr(\bar{A})$$

**Cancer** Let A and B be defined as in Example 3.19, and suppose that 7% of the general population of women will have a positive mammogram. What is the probability of developing breast cancer over the next 2 years among women in the general population?

$$\begin{aligned} Pr(B) &= Pr(\text{breast cancer}) \\ &= Pr(\text{breast cancer} | \text{mammogram}^+) \times Pr(\text{mammogram}^+) \\ &\quad + Pr(\text{breast cancer} | \text{mammogram}^-) \times Pr(\text{mammogram}^-) \\ &= .1(.07) + .0002(.93) = .00719 = 719 / 10^5 \end{aligned}$$



清华大学统计学研究中心

检验报告  $P(\text{Test}^+ | \text{Disease}^+) \rightarrow \text{计算} \rightarrow P(D^+ | T^+)$

## independence独立 & mutually exclusive 互斥

抛两次硬币，这两次硬币的结果是相互独立，互不影响

“结果为正面”与“结果为反面”是相互排斥的

## 1.4 Independence 独立 and Mutually Exclusive 互斥

- In general:  $P(A \cap B) = P(B) \times P(A|B)$

- If two events A and B are statistically **independent**,

$$P(A|B) = P(A)$$

$$P(B|A) = P(B)$$

$$P(A \cap B) = P(B) \times P(A)$$

**Example:** Rolling a dice twice. Let A = getting a 6, B = getting the 6 again.  $P(A \text{ and } B)$  = probability of rolling a pair of the same 6 =  $1/36$

- Independence is **different** from **mutually exclusive (disjoint)** events where

$$P(A \cap B) = 0$$

### 贝叶斯定律

## 1.5 Bayes Rule

- Useful for computing conditional probability based on  $P(B|A)$  or  $P(A|B)$

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}$$

- If  $P(A)$  is unknown

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)}$$

where  $B^c$  denotes "the complement of B" or "not B"

```
setwd("D:/Tennedar/THU/MS") #路径用"/"
getwd()
dat = read.csv("dat.csv",encoding = "UTF-8",header=T) #第一行header不用读
# dat <- read.table("dat.txt");

table(dat$X2.年龄)
table(dat$X1.性别)
table(dat$X6.最喜欢的食堂)
#frequency table

counts=table(dat$X1.性别,dat$X10.是否爱吃甜食)
counts
#列联表 两个维度
barplot(counts,
        main='Stacked Bar Plot',
        xlab='是否爱吃甜食',
        ylab='性别',
```

```

col=c('blue','gold'),
legend=rownames(counts))

hist(dat$X12.身高.单位为m.)
hist(dat$X13.体重.单位为kg.)
hist(dat$X13.体重.单位为kg.,breaks=6,main='break6')
hist(dat$X13.体重.单位为kg.,breaks=20,main='break6')
stem(dat$X12.身高.单位为m.)
stem(dat$X13.体重.单位为kg.)

boxplot(dat$X5.本学期选的学分总数,main='本学期选的学分总数')
boxplot(dat$X13.体重.单位为kg.,main='体重')
boxplot(dat$X12.身高.单位为m.~dat$X1.性别)
boxplot(dat$X13.体重.单位为kg.~dat$X1.性别)

head(dat) #仅前几行
counts2=table(dat$X6.最喜欢的食堂)
pie(counts2,main='test')
barplot(counts2,col="purple",border='green')
barplot(counts2,col="pink",horiz=T)
boxplot(dat$X12.身高.单位为m.~dat$X1.性别)
cols=c("red","blue","pink","purple","yellow","green","gray","black")
head(mtcars)
c=table(mtcars$cyl,mtcars$vs)
barplot(c,
        main='test1',
        xlab='x',
        ylab='y',
        col=cols,
        legend=rownames(c),
        beside=F)
barplot(c,
        main='test1',
        xlab='x',
        ylab='y',
        col=cols,
        legend=rownames(c),
        beside=T)

boxplot(mtcars$mpg~mtcars$cyl,main="boxplot",ylab="y",xlab="x",col=cols)

mtcars$cyl.f =factor(mtcars$cyl,levels=c(4,6,8),labels=c("4","6","8"))
mtcars$am.f=factor(mtcars$am,levels=c(0,1),labels=c("auto","standard"))
boxplot(mpg~am.f*cyl.f,data=mtcars,varwidth=TRUE,col=cols,main="test",xlab="x",ylab="y")

spineplot(counts,col=cols)
plot(mtcars$wt,mtcars$mpg,main="test",pch=19,frame=F,col=cols)
abline(lm(mtcars$mpg~mtcars$wt),col="blue")

```

```

library(scatterplot3d)

x=iris$Sepal.Length
y=iris$Sepal.Width
z=iris$Petal.Length
grps=as.factor(iris$Species)
scatterplot3d(x,y,z,pch=16,,grid=T,box=F)

pairs(iris[,1:4],pch=19,lower.panel= NULL)
# cex 直径
library(psych)
pairs.panels(iris[, -5],method="pearson",density=T,ellipses=T)

```

