第四讲

R的基本分析和统计图形
---单变量分析

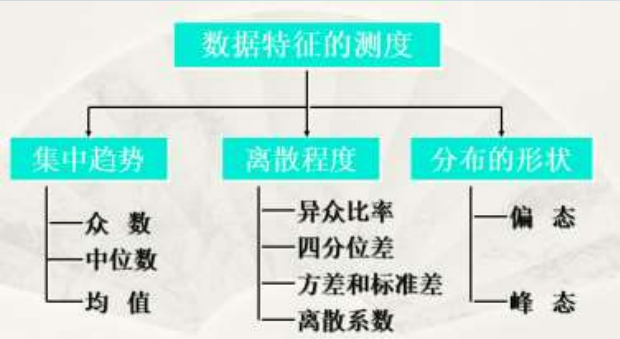
第4讲 R的基本分析和统计图形

- 数据基本分析是从单个变量的描述统计开始的
  - 描述统计的目的是揭示变量的分布特点
  - 描述统计的基本工具两类:
    - 第一，计算描述统计量
    - 第二，数据可视化

  - 不同类型变量的基本分析方法有所不同，需要采用不同的描述统计量和不同的可视化图形。

第4讲 R的基本分析和统计图形

- 从大数据分析案例看数据基本分析
  - 美食餐馆食客点评数据的基本分析
    - 数据：五道口和北太平庄两个区域600余家餐馆最受欢迎的10种菜系的大众点评数据
      - 口味、就餐环境、服务质量满分40分；平均综合得分满分5分
    - 菜系的餐馆分布有怎样的特点？
      - 单个分类型变量的描述统计
    - 食客评分、人均消费金额有怎样的统计特征？
      - 单个数值型变量的描述统计
  - 单个分类型变量的基本分析：频数分布表
  - 数值型变量的基本分析：描述统计量



shop_ID	region	food_type	review_n	taste	environment	service	score_avg	cost_avg	heat
508022	北太平庄	火锅	571	21	19	17	3.57	50	
508241	五道口	咖啡厅	260	24	26	26	4.04	37	
508272	五道口	咖啡厅	58	19	23	20	3.64	32	
508302	北太平庄	小吃	339	19	13	12	3.34	31	
508452	北太平庄	北京菜	901	24	21	19	3.77	96	
508491	北太平庄	火锅	571	19	16	17	3.62	42	
508739	北太平庄	川菜	694	26	14	15	3.73	69	
509126	五道口	川菜	374	18	16	14	3.25	46	
509198	五道口	咖啡厅	456	21	25	21	3.63	40	
509479	五道口	小吃	492	20	17	16	3.42	28	
510020	五道口	韩国料理	1153	23	17	17	3.64	48	
510090	北太平庄	韩国料理	977	24	21	22	3.98	60	
510125	北太平庄	火锅	231	21	15	18	3.45	40	
510502	北太平庄	川菜	108	21	16	18	3.54	36	
510548	北太平庄	快餐	136	19	15	15	3.18	15	
510895	北太平庄	韩国料理	1213	26	22	23	4.03	63	

- 该案例分析还涉及，例如：
  - 食客评分及人均消费金额是否具有相关性？
    - 两数值型变量取值相关性
  - 餐馆的区域分布与菜系分布是否具有相关性？
    - 两分类型变量的相关性研究

第4讲 R的基本分析和统计图形

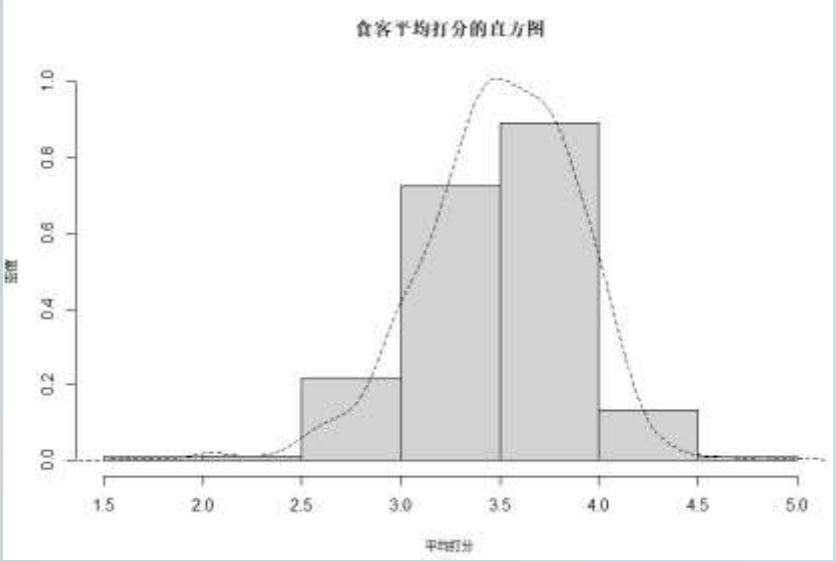
- 数据可视化工具
  - 数据可视化的基本工具是统计图形
  - R的绘图基础
    - 图形设备和图形文件
      - 图形窗口是一种图形设备
      - 图形文件也是一种图形设备

函数	功能
win.graph()	手工创建打开一个图形设备，该设备为当前图形设备
dev.cur()	显示当前图形设备的编号
dev.list()	显示当前已有几个图形设备被创建打开
dev.set(*n*)	指定编号为 *n* 的图形设备为当前图形设备
dev.off()	关闭当前图形设备，即关闭当前图形窗口
dev.off(*n*)	关闭编号为 *n* 的图形设备

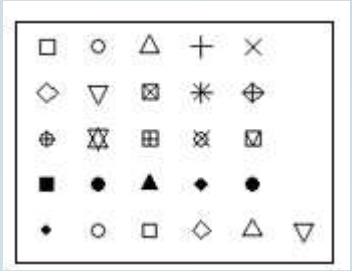
函数	功能
pdf("文件名.pdf")	指定某 PDF 格式文件为当前图形设备
win.metafile("文件名.wmf")	指定某 WMF 波形格式文件为当前图形设备
png("文件名.png")	指定某 PNG 格式文件为当前图形设备
jpeg("文件名.gpg")	指定某 JPEG 格式文件为当前图形设备
bmp("文件名.bmp")	指定某 BMP 格式文件为当前图形设备
postscript("文件名.ps")	指定某 PS 格式文件为当前图形设备

第4讲 R的基本分析和统计图形

- R的绘图基础
  - 图形组成和图形参数



类别	特征	英文缩写
符号	种类	pch
	大小	cex
	填充色	bg
线条	线型	lty
	宽度	lwd
颜色	颜色	col



类别	特征	英文缩写
标题内容	横坐标内容	xlab
	纵坐标内容	ylab
标题文字	文字颜色	col.lab
	文字大小	cex.lab
	文字字体	font.lab

类别	特征	英文缩写
标题内容	主标题内容	main
	副标题内容	sub
主标题文字	文字颜色	col.main
	文字大小	cex.main
	文字字体	font.main
副标题文字	文字颜色	col.sub
	文字大小	cex.sub
	文字字体	font.sub

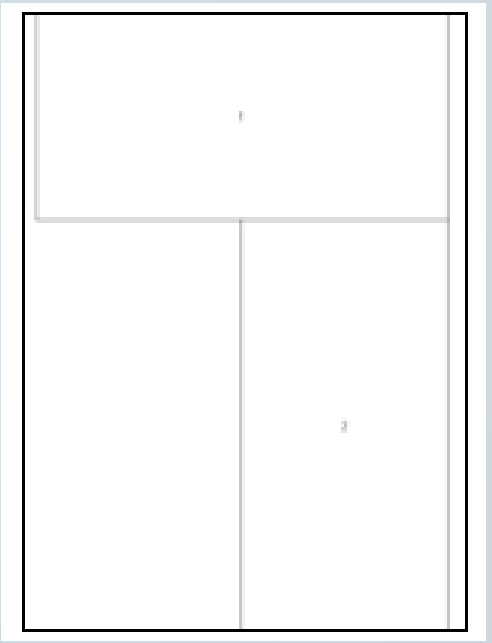
类别	特征	英文缩写
刻度	位置	at
	长度和方向	tcl
刻度范围	横坐标范围	xlim
	纵坐标范围	ylim
刻度文字	文字内容	label
	文字颜色	col.axis
	文字大小	cex.axis
	文字字体	font.axis

```
hist(MyData$score_avg,xlab="平均打分",ylab="密度",main="食客平均打分的直方图",cex.lab=0.7,freq=FALSE,ylim=c(0,1))
lines(density(MyData$score_avg),lty=2,col=1)
```

第4讲 R的基本分析和统计图形

- R的绘图基础
  - 图形布局：
    - 标准布局： par(mfrow=c(行数,列数),mar=c(*n*1,*n*2,*n*3,*n*4))
      - 例如： par(mfrow=c(2,1))
    - 自由布局：

```
#####################图形布局
MyLayout<-matrix(c(1,1,0,2),nrow=2,ncol=2,byrow=TRUE)
DrawLayout<-layout(MyLayout,widths=c(1,1),heights=c(1,2),respect=TRUE)
layout.show(DrawLayout)
```



第4讲 R的基本分析和统计图形

- 菜系的餐馆分布有怎样的特点分类型—单变量的基本分析
  - 编制频数分布表：用于展示单个分类型变量的分布特征
    - 编制频数分布表：
      - table()；prop.table(表名)

shop_ID	region	food_type	review_n	taste	environment	service	score_avg	cost_avg	heat
508022	北太平庄	火锅	571	21	19	17	3.57	50	
508241	五道口	咖啡厅	260	24	26	26	4.04	37	
508272	五道口	咖啡厅	58	19	23	20	3.64	32	
508302	北太平庄	小吃	339	19	13	12	3.34	31	
508452	北太平庄	北京菜	901	24	21	19	3.77	96	
508491	北太平庄	火锅	571	19	16	17	3.62	42	
508739	北太平庄	川菜	694	26	14	15	3.73	69	
			374	18	16	14	3.25	46	
			456	21	25	21	3.63	40	
			492	20	17	16	3.42	28	
		理	1153	23	17	17	3.64	48	
		理	977	24	21	22	3.98	60	
			231	21	15	18	3.45	40	
			108	21	16	18	3.54	36	
			136	19	15	15	3.18	15	
		理	1213	26	22	23	4.03	63	

```
MyData<-read.table(file="美食餐馆食客评分数据.txt",header=TRUE,sep=" ",stringsAsFactors=FALSE)
(freqT<-table(MyData$food_type))
addmargins(freqT)      #在频数分布表上增加合计
prop.table(freqT)*100  #计算百分比
```

```
> (freqT<-table(MyData$food_type))

  北京菜     川菜 韩国料理     火锅   咖啡厅     快餐     面包 其他小吃     甜点     小吃
      79       76       47       64       46       76       46       50       36      161
> addmargins(freqT)      #在频数分布表上增加合计

  北京菜     川菜 韩国料理     火锅   咖啡厅     快餐     面包 其他小吃     甜点     小吃      Sum
      79       76       47       64       46       76       46       50       36      161      681
> prop.table(freqT)*100  #计算百分比

   北京菜      川菜  韩国料理      火锅    咖啡厅      快餐      面包  其他小吃      甜点      小吃
11.600587 11.160059  6.901615  9.397944  6.754772 11.160059  6.754772  7.342144  5.286344 23.641703
```

第4讲 R的基本分析和统计图形

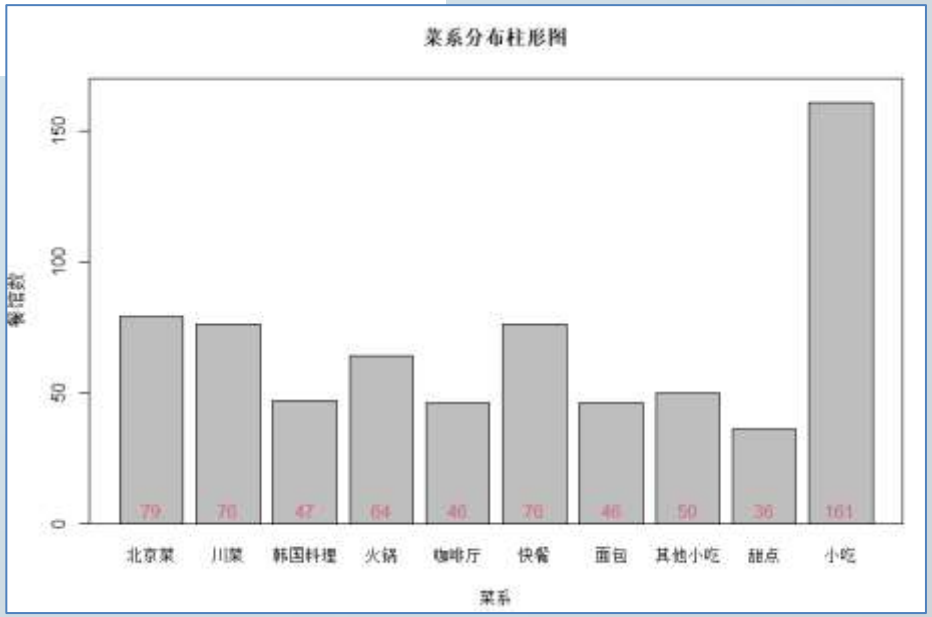
- 菜系的餐馆分布有怎样的特点分类型—单变量的基本分析
  - 可视化：直观展示单个分类型变量分布特征
    - 常用统计图形有柱形图或条形图、饼图、扇形图

barplot(数值型向量名,horiz=TRUE/FALSE,names.arg=条形的标签向量)

```
T<-barplot(freqT,xlab="菜系",ylab="餐馆数",ylim=c(0,170),main="菜系分布柱形图")
box()
text(T,5,freqT,col=2)
```

T向量中记录了各柱形图的横坐标位置

```
> T
      [,1]
 [1,]  0.7
 [2,]  1.9
 [3,]  3.1
 [4,]  4.3
 [5,]  5.5
 [6,]  6.7
 [7,]  7.9
 [8,]  9.1
 [9,] 10.3
[10,] 11.5
```

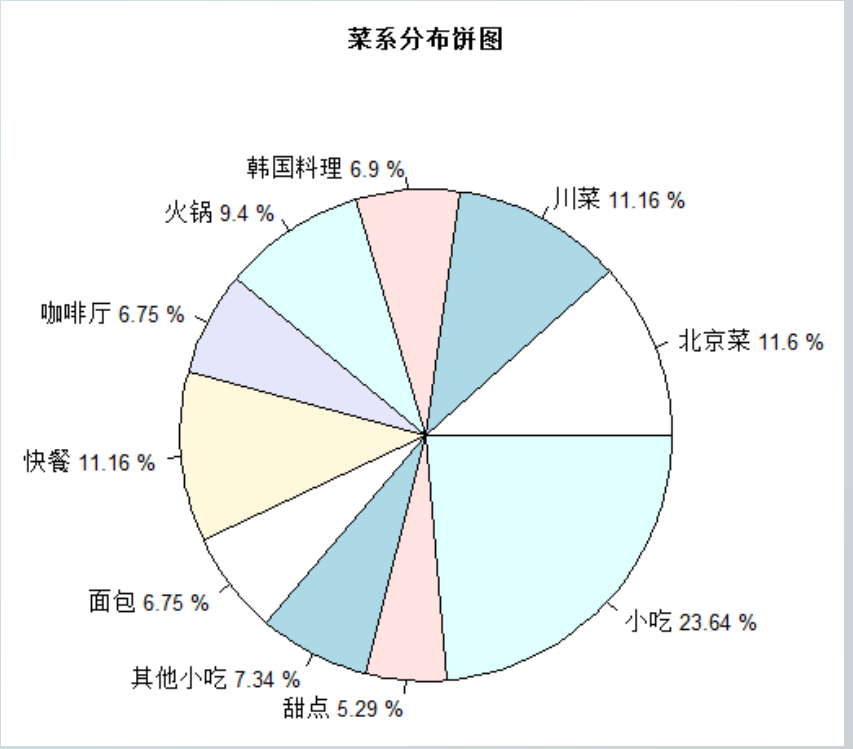


第4讲 R的基本分析和统计图形

- 菜系的餐馆分布有怎样的特点分类型—单变量的基本分析
  - 可视化：直观展示单个分类型变量分布特征
    - 常用统计图形有柱形图或条形图、饼图、扇形图

pie(数值型向量名,labels=切片标签向量,clockwise=TRUT/FALSE)

```
Pct<-round(freqT/length(MyData$food_type)*100,2)
GLabs<-sapply(dimnames(freqT),FUN=function(x) paste(x,Pct,"%",sep=" "))
pie(freqT,cex=0.8,labels=GLabs,main="菜系分布饼图",cex.main=0.8)
```

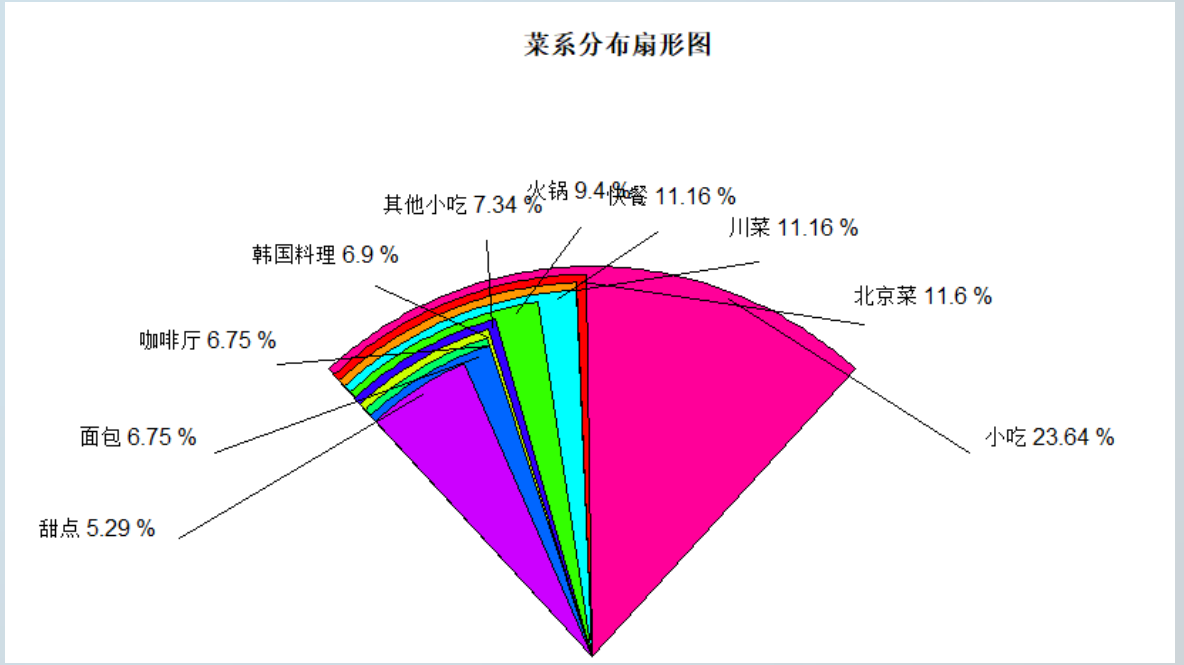


第4讲 R的基本分析和统计图形

- 菜系的餐馆分布有怎样的特点分类型—单变量的基本分析
  - 可视化：直观展示单个分类型变量分布特征
    - 常用统计图形有柱形图或条形图、饼图、扇形图

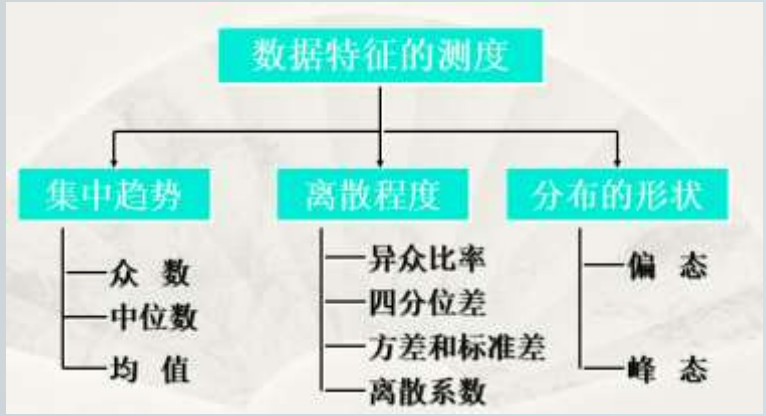
fan.plot(数值型向量名,labels=切片标签向量)

```
library("plotrix")
fan.plot(freqT,labels=GLabs)
title(main="菜系分布扇形图")
```



第4讲 R的基本分析和统计图形

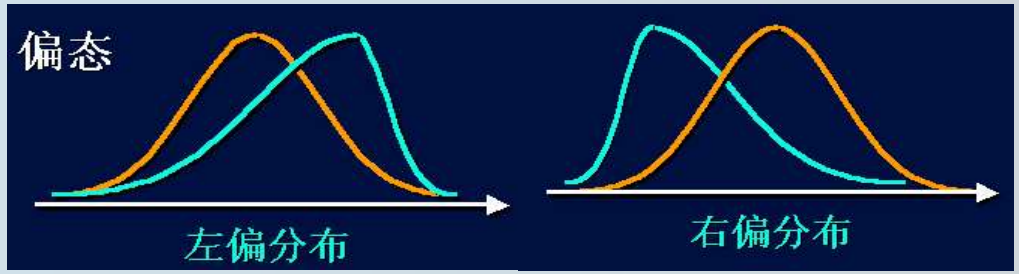
- 食客评分及人均消费金额有怎样的统计特征—单个数值型变量的描述统计
  - 计算基本描述统计量：刻画单个数值型变量分布特征
  - 分布特征和描述统计量：



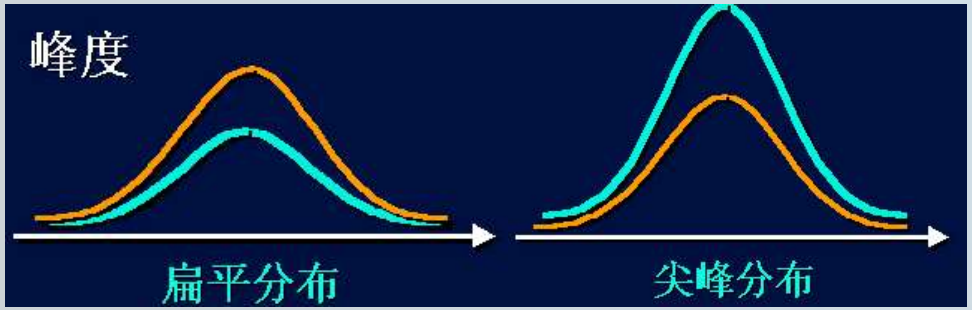
- 偏态系数：

$$\text{Skewness} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^3}{S^3}$$

- 系数=0为对称分布；系数>0为右偏分布；系数<0为左偏分布



- 峰度系数：



$$\text{Kurtosis} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\bar{X})^4}{S^4} - 3$$

- 系数=0扁平峰度适中;系数>0为尖峰分布;系数<0为扁平分布

第4讲 R的基本分析和统计图形

- 食客评分及人均消费金额有怎样的统计特征—单个数值型变量的描述统计
  - 计算基本描述统计量：刻画单个数值型变量分布特征

shop_ID	region	food_type	review_n	taste	environment	service	score_avg	cost_avg	heat
508022	北太平庄	火锅	571	21	19	17	3.57	50	5
508241	五道口	咖啡厅	260	24	26	26	4.04	37	4
508272	五道口	咖啡厅	58	19	23	20	3.64	32	3
508302	北太平庄	小吃	339	19	13	12	3.34	31	5
508452	北太平庄	北京菜	901	24	21	19	3.77	96	5
508491	北太平庄	火锅	571	19	16	17	3.62	42	5
508739	北太平庄	川菜	694	26	14	15	3.73	69	5
509126	五道口	川菜	374	18	16	14	3.25	46	5
509198	五道口	咖啡厅	456	21	25	21	3.63	40	5
509479	五道口	小吃	492	20	17	16	3.42	28	5
510020	五道口	韩国料理	1153	23	17	17	3.64	48	5
510090	北太平庄	韩国料理	977	24	21	22	3.98	60	5
510125	北太平庄	火锅	231	21	15	18	3.45	40	4
510502	北太平庄	川菜	108	21	16	18	3.54	36	3
510548	北太平庄	快餐	136	19	15	15	3.18	15	4
510895	北太平庄	韩国料理	1213	26	22	23	4.03	63	5

```
apply(MyData[,5:9],MARGIN = 2,mean)
    taste environment     service   score_avg    cost_avg
21.985316   18.861968   18.989721    3.500808   47.555066
```

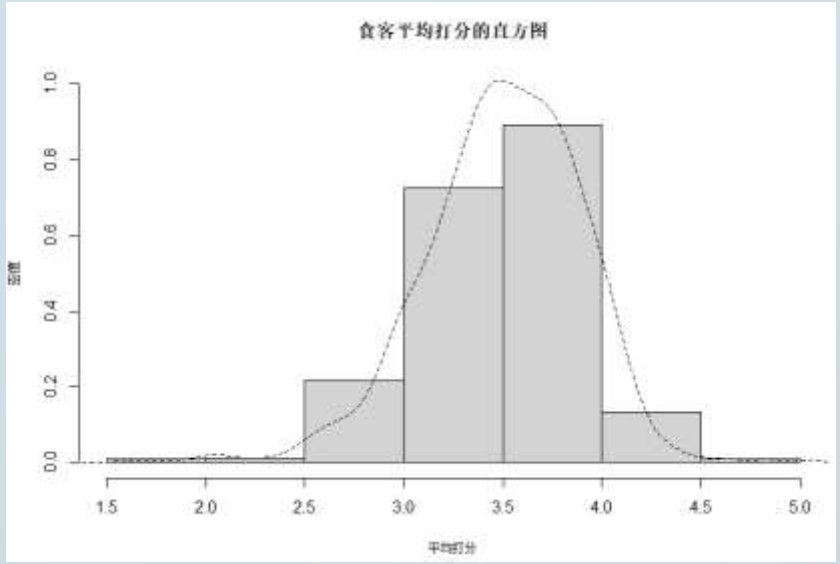
```
Des.Fun<-function(x,...){
  Av<-mean(x,...)
  Sd<-sd(x,...)
  N<-length(x[!is.na(x)])
  Sk<-sum((x[!is.na(x)]-Av)^3/Sd^3)/N
  Ku<-sum((x[!is.na(x)]-Av)^4/Sd^4)/N-3
  result<-list(avg=Av,sd=Sd,skew=Sk,kurt=Ku)
  return(result)
}
```

```
       taste     environment service   score_avg  cost_avg
avg   21.98532  18.86197     18.98972  3.500808   47.55507
sd    2.505832  2.881234     2.569884  0.3945881  388.7081
skew  0.3189935 0.5942442    0.6888209 -0.5322249 25.75078
kurt  0.1785143 0.7111549    1.348758  1.290654   665.9261
```

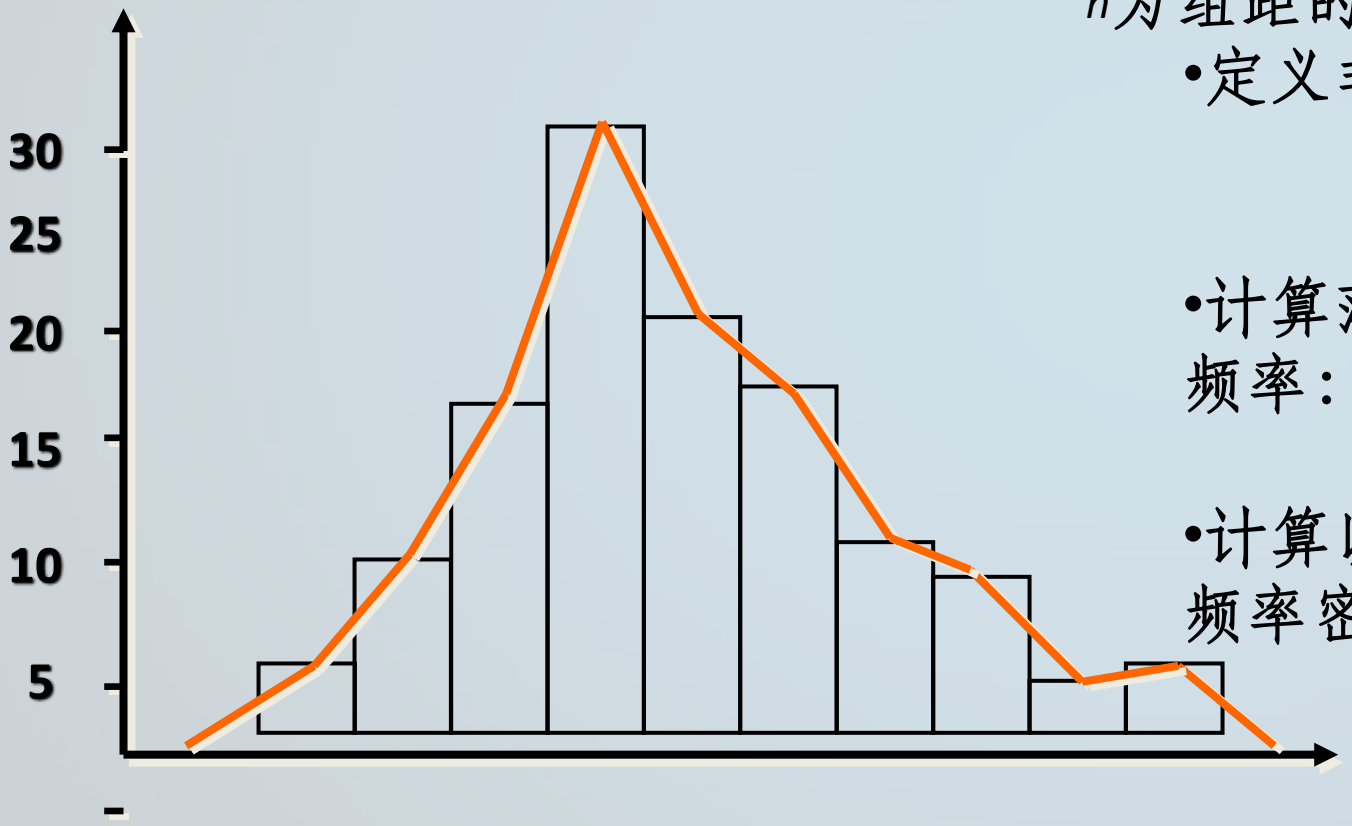
第4讲 R的基本分析和统计图形

- 食客评分及人均消费金额有怎样的统计特征—单个数值型变量的描述统计
  - 可视化：直观展示单个数值型变量分布特征
  - 常用图形：直方图，核密度图，箱线图，小提琴图等等
    - 直方图：**hist(数值型向量名或域名,freq=TRUE/FALSE)**
      - 添加核密度估计曲线—核密度图：**density(数值型向量)**

```
hist(MyData$score_avg,xlab="平均打分",ylab="密度",main="食客平均打分的直方图",cex.lab=0.7,freq=FALSE,ylim=c(0,1))
lines(density(MyData$score_avg),lty=2,col=1)
```



•核密度估计：核密度曲线可视为，将直方图各组的组中值及对应的密度值为坐标确定的点，做连线后形成的折线图



•设有n个观测，计算落入以x_0为中心（组中值）h为组距的“直方桶”区间R中的观测个数

•定义非负的距离函数：

$$k(||x_0-x_i||)=\begin{cases}1, |x_0-x_i| \le \frac{h}{2}; \ i=1,2,...,n \\ 0, |x_0-x_i| > \frac{h}{2}; \ i=1,2,...,n\end{cases}$$

•计算落入以x_0为中心的R中的观测频数和频率：

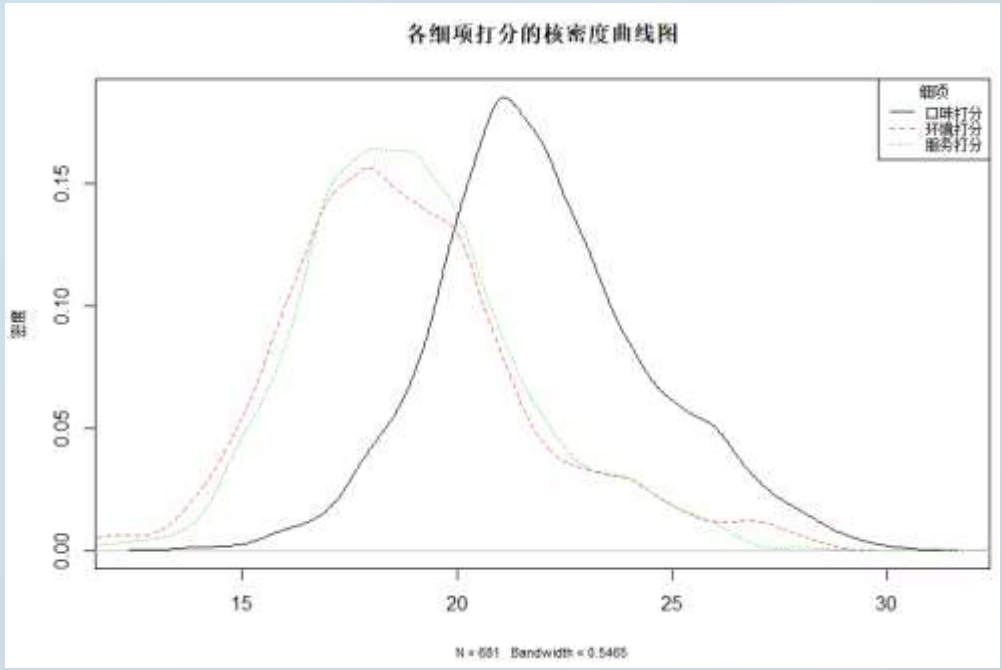
$$\frac{1}{n}\sum_{i=1}^{n} k(||x_0-x_i||)$$

•计算以x_0为中心h范围内观测点频率的函数：频率密度=频率/组距h

$$f(x_0)=\frac{1}{hn}\sum_{i=1}^{n} k(||x_0-x_i||)$$

•核密度估计：核密度曲线可视为，将直方图各组的组中值及对应的密度值为坐标确定的点，做连线后形成的折线图

```
plot(density(MyData$taste),main="各细项打分的核密度曲线图",ylab="密度",cex.lab=0.7)
lines(density(MyData$environment),lty=2,col=2)
lines(density(MyData$service),lty=3,col=3)
legend("topright",title="细项",c("口味打分","环境打分","服务打分"),lty=c(1,2,3),col=c(1,2,3),cex=0.7)
```



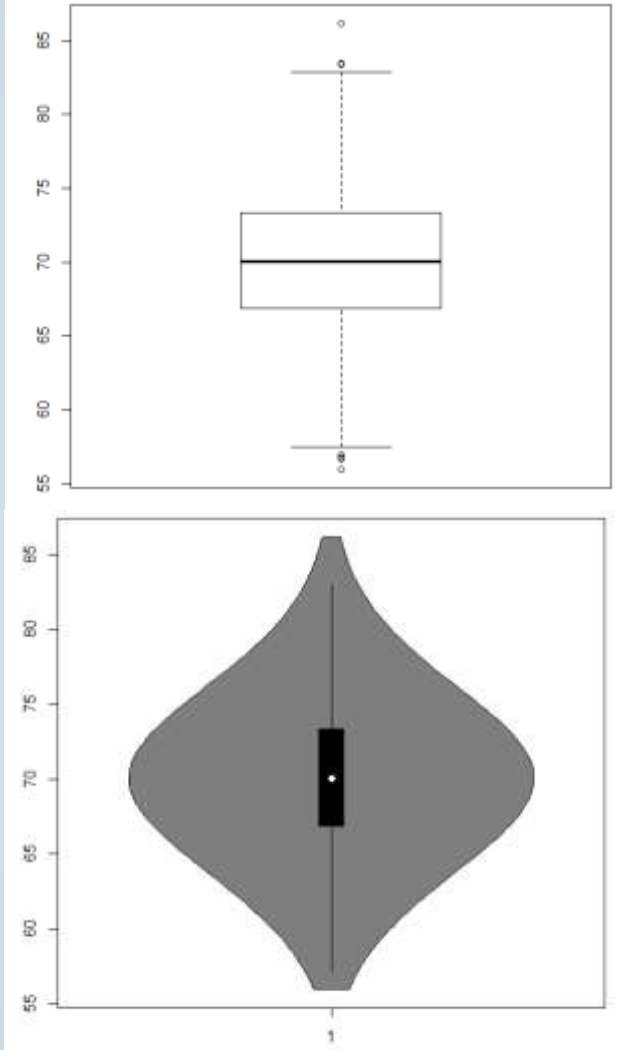
第4讲 R的基本分析和统计图形

- 常用图形
  - 箱线图：

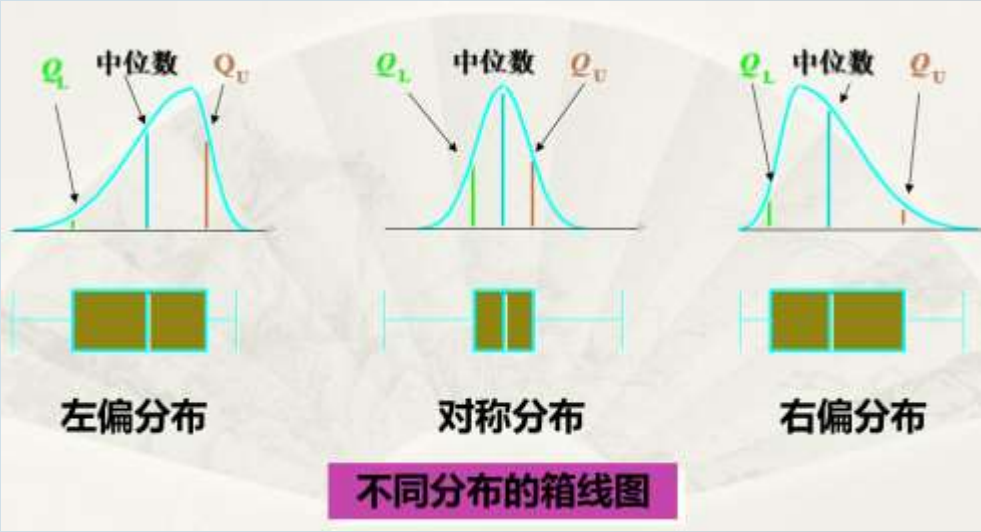
boxplot(数值型向量名或域名, horizontal=TRUE/FALSE, axes=TRUE/FALSE)

  - 小提琴图：

vioplot(数值型向量名或域名列表, names=横坐标轴标题向量)

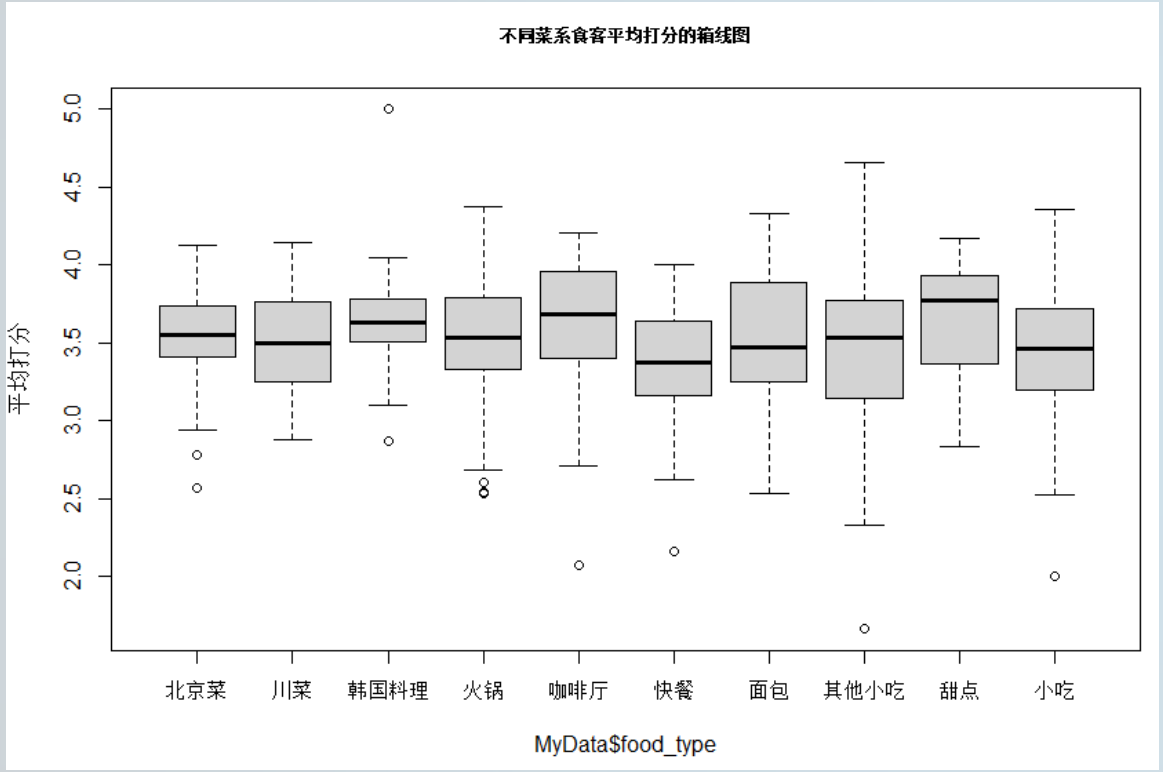


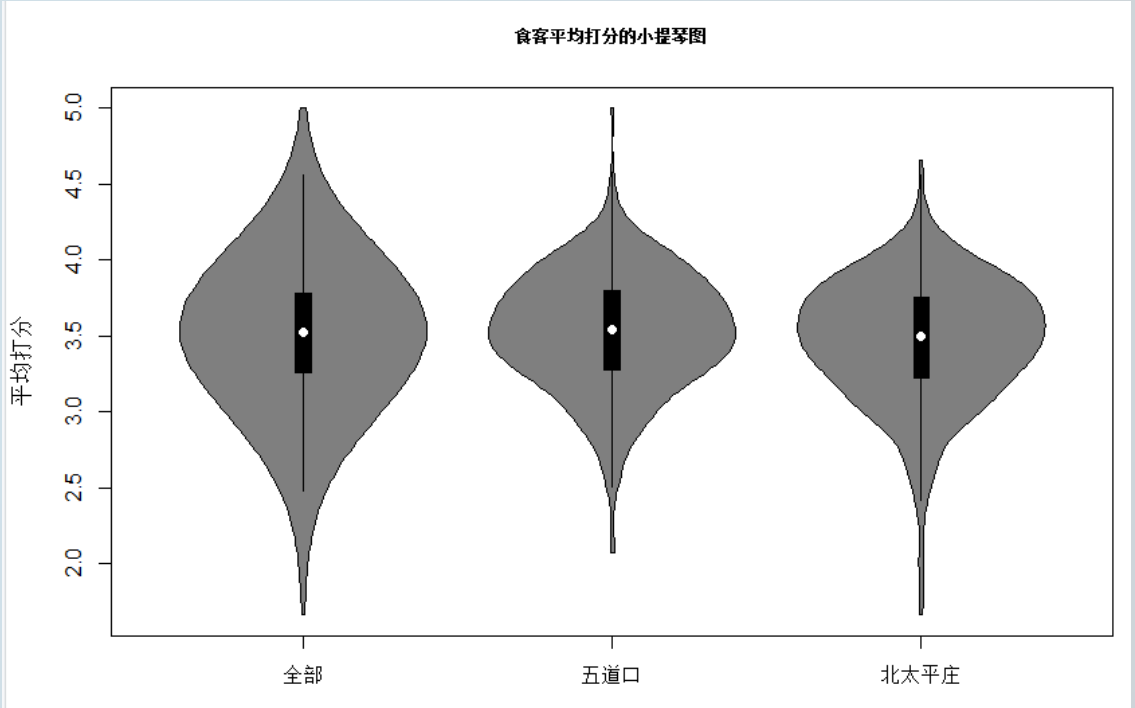
- 离群点：以四分位差的1.5倍为标准
- 从箱线图看分布对称性



第4讲 R的基本分析和统计图形

boxplot(MyData$score_avg~MyData$food_type,main="不同菜系食客平均打分的箱线图",ylab="平均打分",cex.main=0.7)






```
library("vioplot")
vioplot(MyData[,"score_avg"],MyData[MyData$region=="五道口","score_avg"],MyData[MyData$region=="北太平庄","score_avg"],
        names=c("全部","五道口","北太平庄"))
title(main="食客平均打分的小提琴图",cex.main=0.7,ylab="平均打分")
```

第4讲 R的基本分析和统计图形：综合应用

- 北京市空气质量监测数据基本分析
  - 原始数据：2016全年(12.28止)北京市35个空气监测点的实时(小时)监测数据

- 四类监测点：
  - 城区环境评价点；
  - 郊区环境评估点；
  - 对照点及区域点；
  - 交通污染监控点
- 经/纬度

北京空气质量分析.R × MyData ×

Filter

	SiteName	date	PM2.5	AQI	CO	NO2	O3	SO2	SiteTypes	SiteX	SiteY
1	奥体中心	20160101	164.958333	154.58333	3.9291667	122.625000	10.666667	45.666667	城区环境评价点	116.397	39.982
379	八达岭	20160101	91.681818	92.33333	1.7083333	83.375000	4.333333	55.625000	对照点及区域点	115.988	40.365
1083	北部新区	20160101	190.458333	196.66667	4.7041667	93.375000	2.000000	17.125000	城区环境评价点	116.174	40.090
1218	昌平	20160101	128.750000	123.20833	3.0541667	112.708333	3.750000	19.833333	郊区环境评价点	116.230	40.217
1706	大兴	20160101	230.250000	230.33333	NA	125.750000	12.166667	50.750000	郊区环境评价点	116.404	39.718
2072	定陵	20160101	130.500000	121.25000	3.7833333	81.041667	4.625000	32.375000	对照点及区域点	116.220	40.292
2477	东高村	20160101	285.083333	224.00000	3.2291667	99.000000	3.875000	28.541667	对照点及区域点	117.120	40.100
2598	东四	20160101	178.833333	166.66667	3.2000000	100.625000	3.833333	35.458333	城区环境评价点	116.417	39.929
2923	东四环	20160101	234.708333	209.00000	5.1541667	106.291667	2.208333	38.083333	交通污染监控点	116.483	39.939
3368	房山	20160101	215.541667	196.25000	4.2708333	117.416667	3.833333	56.083333	郊区环境评价点	116.136	39.742
3776	丰台花园	20160101	212.291667	186.12500	4.4708333	120.166667	3.541667	39.166667	城区环境评价点	116.279	39.863
4140	古城	20160101	184.208333	165.75000	3.4416667	102.416667	6.916667	40.541667	城区环境评价点	116.184	39.914
4671	官园	20160101	176.125000	151.54167	3.3458333	121.583333	5.250000	49.083333	城区环境评价点	116.339	39.929
5061	怀柔	20160101	136.250000	116.16667	1.9500000	57.826087	4.750000	14.583333	郊区环境评价点	116.628	40.328
5239	琉璃河	20160101	281.416667	277.79167	4.4666667	109.708333	2.000000	17.333333	对照点及区域点	116.000	39.580
5603	门头沟	20160101	122.736842	113.75000	2.3916667	90.608696	8.625000	21.041667	郊区环境评价点	116.106	39.937
5847	密云	20160101	161.625000	134.58333	2.7291667	67.958333	10.208333	28.166667	郊区环境评价点	116.832	40.370
6333	密云水库	20160101	119.166667	94.37500	1.5375000	39.458333	13.791667	15.083333	对照点及区域点	116.911	40.499
6699	南三环	20160101	196.875000	181.50000	3.5958333	112.916667	8.458333	35.916667	交通污染监控点	116.368	39.856
6942	农展馆	20160101	199.916667	180.20833	3.8333333	119.875000	30.750000	35.375000	城区环境评价点	116.461	39.937

Showing 1 to 20 of 12,705 entries

- 分析目标：
  - 供暖季PM2.5浓度有怎样的分布特征？
  - 供暖季是否存在PM2.5浓度“爆表”的情况？哪些监测点出现了几天“爆表”？

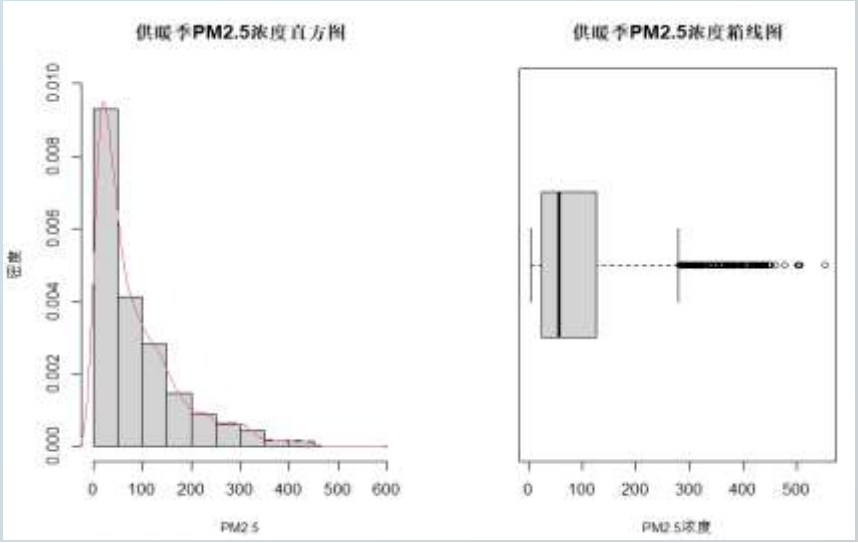
第4讲 R的基本分析和统计图形：综合应用

- 北京市空气质量监测数据的基本分析
  - 供暖季PM2.5浓度有怎样的分布特征---数值型单变量基本分析

```
############################北京市空气质量监测数据基本分析
MyData<-read.table(file="空气质量.csv",header=TRUE,sep=",",stringsAsFactors=FALSE)
Data<-subset(MyData,(MyData$date<=20160315|MyData$date>=20161115))    #仅分析供暖季的空气质量数据
Data<-na.omit(Data)    #完整观测
library("psych")
describe(Data$PM2.5,IQR=TRUE)
par(mfrow=c(1,2))
hist(Data$PM2.5,xlab="PM2.5",ylab="密度",main="供暖季PM2.5浓度直方图",cex.lab=0.8,freq=FALSE,ylim=c(0,0.01))
lines(density(Data$PM2.5,na.rm=TRUE),col=2)
boxplot(Data$PM2.5,horizontal =TRUE,main="供暖季PM2.5浓度箱线图",xlab="PM2.5浓度",cex.lab=0.8)
```

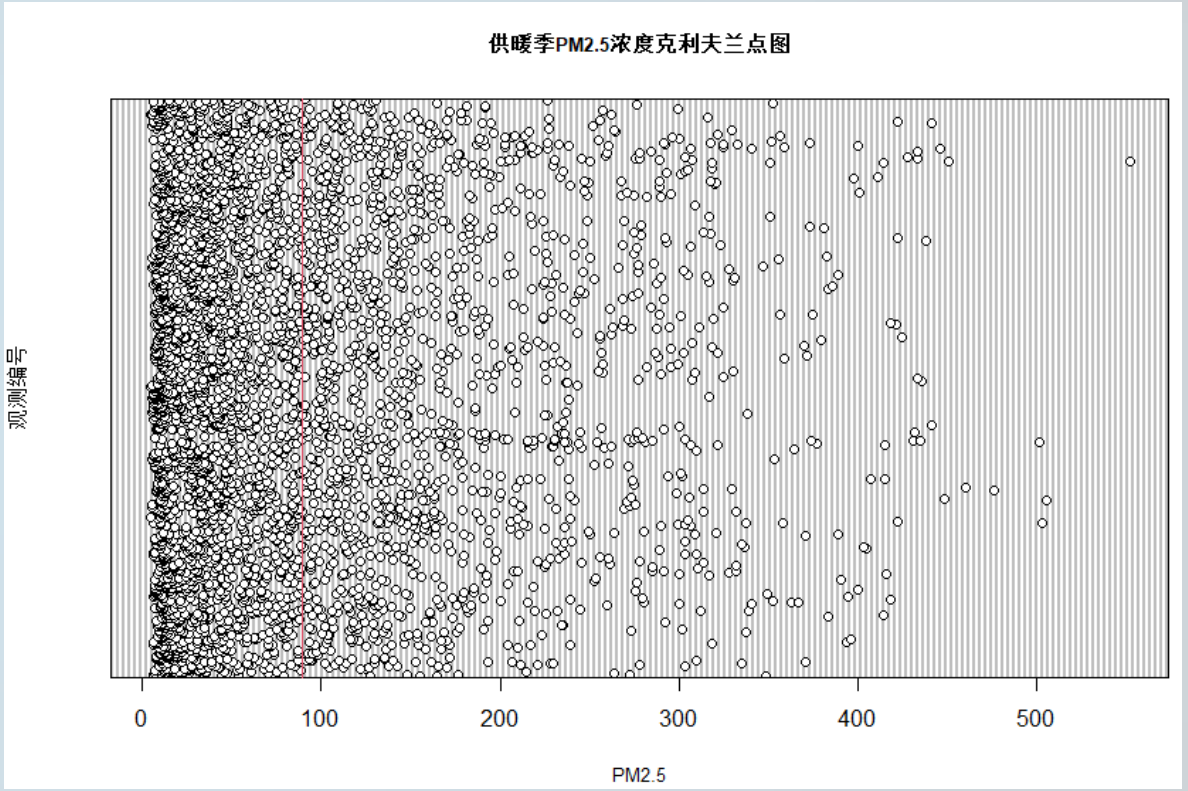
```
   vars    n  mean    sd median trimmed  mad  min   max  range skew kurtosis   se    IQR
X1    1 4112 88.96 88.03  56.71   73.15 61.9 4.38 551.9 547.53 1.61     2.53 1.37 102.59
```

- trimmed：剔除10%(最大和最小取值段各5%)数据后的均值
- mad：中位数绝对离差
- IQR：四分位差



第4讲 R的基本分析和统计图形：综合应用

- 供暖季是否存在PM2.5浓度爆表的情况？哪些监测点出现了几天爆表？
  - 图形工具：克利夫兰点图
    - dotchar(数值型向量名或域名)
    - abline(h=纵坐标值)
    - abline(v=横坐标值)




```
dotchart(Data$PM2.5,main="供暖季PM2.5浓度克利夫兰点图",cex.main=0.8,xlab="PM2.5",ylab="观测编号",cex.lab=0.8)
abline(v=mean(Data$PM2.5),col=2)
```

第4讲 R的基本分析和统计图形：综合应用

- 供暖季是否存在PM2.5浓度爆表的情况？哪些监测点出现了几天爆表？
  - 数值研判：从统计意义判断--极端值(以四分位差的1.5倍为标准)

```
(InDiff<-quantile(Data$PM2.5,0.75)-quantile(Data$PM2.5,0.25))
(threshold<-(quantile(Data$PM2.5,0.75)+1.5*InDiff))   #爆表临界值
exData<-Data[Data$PM2.5>threshold,]   #爆表数据
length(unique(exData$date))    #爆表天数
sort(table(exData[,"SiteName"]),decreasing = TRUE)   #各站点的爆表情况
unique(exData[,"SiteName"])
```

北京空气质量分析.R × | MyData ×

Filter

	SiteName	date	PM2.5	AQI	CO	NO2	O3	SO2	SiteTypes	SiteX	SiteY
1	奥体中心	20160101	164.958333	154.58333	3.9291667	122.625000	10.666667	45.666667	城区环境评价点	116.397	39.982
379	八达岭	20160101	91.681818	92.33333	1.7083333	83.375000	4.333333	55.625000	对照点及区域点	115.988	40.365
1083	北部新区	20160101	190.458333	196.66667	4.7041667	93.375000	2.000000	17.125000	城区环境评价点	116.174	40.090
1218	昌平	20160101	128.750000	123.20833	3.0541667	112.708333	3.750000	19.833333	郊区环境评价点	116.230	40.217
1706	大兴	20160101	230.250000	230.33333	NA	125.750000	12.166667	50.750000	郊区环境评价点	116.404	39.718
2072	定陵	20160101	130.500000	121.25000	3.7833333	81.041667	4.625000	32.375000	对照点及区域点	116.220	40.292
2477	东高村	20160101	285.083333	224.00000	3.2291667	99.000000	3.875000	28.541667	对照点及区域点	117.120	40.100
2598	东四	20160101	178.833333	166.66667	3.2000000	100.625000	3.833333	35.458333	城区环境评价点	116.417	39.929
2923	东四环	20160101	234.708333	209.00000	5.1541667	106.291667	2.208333	38.083333	交通污染监控点	116.483	39.939
3368	房山	20160101	215.541667	196.25000	4.2708333	117.416667	3.833333	56.083333	郊区环境评价点	116.136	39.742
3776	丰台花园	20160101	212.291667	186.12500	4.4708333	120.166667	3.541667	39.166667	城区环境评价点	116.279	39.863
				165.75000	3.4416667	102.416667	6.916667	40.541667	城区环境评价点	116.184	39.914
				151.54167	3.3458333	121.583333	5.250000	49.083333	城区环境评价点	116.339	39.929
				116.16667	1.9500000	57.826087	4.750000	14.583333	郊区环境评价点	116.628	40.328
				277.79167	4.4666667	109.708333	2.000000	17.333333	对照点及区域点	116.000	39.580
				113.75000	2.3916667	90.608696	8.625000	21.041667	郊区环境评价点	116.106	39.937
				134.58333	2.7291667	67.958333	10.208333	28.166667	郊区环境评价点	116.832	40.370
				94.37500	1.5375000	39.458333	13.791667	15.083333	对照点及区域点	116.911	40.499
				181.50000	3.5958333	112.916667	8.458333	35.916667	交通污染监控点	116.368	39.856
				180.20833	3.8333333	119.875000	30.750000	35.375000	城区环境评价点	116.461	39.937

```
> length(unique(exData$date))    #爆表天数
[1] 21
> sort(table(exData[,"SiteName"]),decreasing = TRUE)   #各站点的爆表情况

    榆垡   琉璃河   永乐店   东高村     大兴     房山   南三环     通州 永定门内   东四环 农展馆 万寿西宫     亦庄
      17       12       12       11       10        9        9        9        9        8        7        7
    东四     平谷     前门     顺义 奥体中心 北部新区 丰台花园     古城     官园 西直门北     天坛     万柳     云岗
       6        6        6        6        5        5        5        5        5        5        4        4
    昌平     怀柔   门头沟     定陵   植物园     密云 密云水库     延庆
       3        3        3        2        2        1        1        1
> unique(exData[,"SiteName"])
 [1] "奥体中心" "北部新区" "昌平"     "大兴"     "定陵"     "东高村"   "东四"     "东四环"   "房山"     "丰台花园"
[11] "古城"     "官园"     "怀柔"     "琉璃河"   "门头沟"   "密云"     "密云水库" "南三环"   "农展馆"   "平谷"
[21] "前门"     "顺义"     "天坛"     "通州"     "万柳"     "万寿西宫" "西直门北" "延庆"     "亦庄"     "永定门内"
[31] "永乐店"   "榆垡"     "云岗"     "植物园"
```

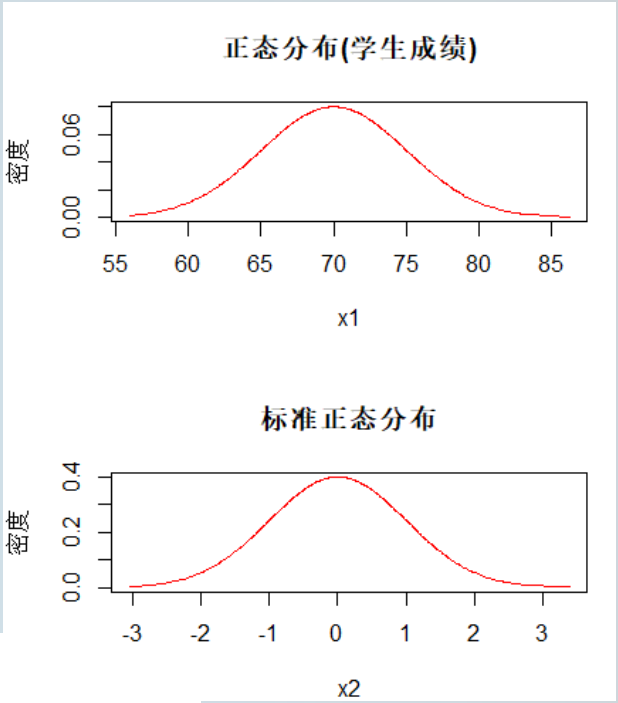
第4讲 R的基本分析和统计图形

- 有关单个数值变量的其他R内容：随机数和分布
  - 例如： 正态分布:rnorm(),dnorm(),pnorm(),qnorm()

```
#################################学生成绩：正态分布
set.seed(123)
x1<-rnorm(1000,70,5)
summary(x1)
mean(x1)
var(x1)
sd(x1)

par(mfrow=c(2,1))
hist(x1,main="学生成绩(正态分布)",ylab="密度",freq=FALSE)
x1<-x1[order(x1)]
lines(x1,dnorm(x1,70,5),type="l",col=2,main="正态分布(学生成绩)",ylab="密度")
print(pnorm(60,70,5))
print(qnorm(0.25,70,5))
print(pnorm(80,70,5))
print(1-pnorm(80,70,5))
print(pnorm(80,70,5)-pnorm(60,70,5))
```

```
> print(pnorm(60,70,5))
[1] 0.02275013
> print(qnorm(0.25,70,5))
[1] 66.62755
> print(pnorm(80,70,5))
[1] 0.9772499
> print(pnorm(80,70,5)-pnorm(60,70,5))
[1] 0.9544997
```



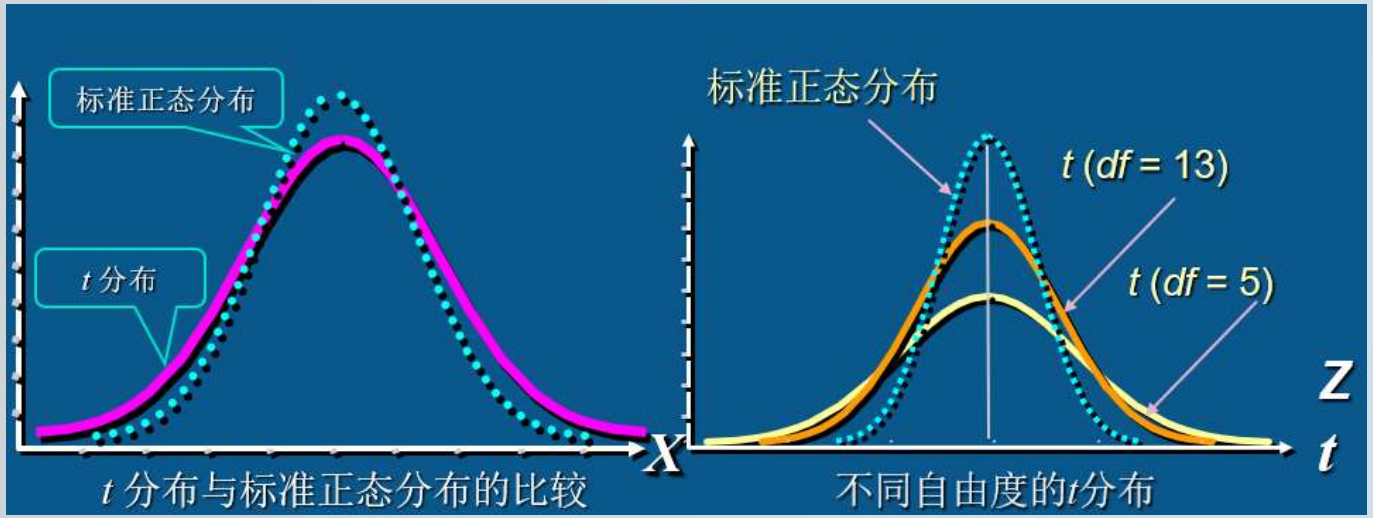

```
x2<-rnorm(1000,0,1)
summary(x2)
x2<-x2[order(x2)]
plot(x2,dnorm(x2,0,1),type="l",col=2,main="标准正态分布",ylab="密度")
print(qnorm(0.05,0,1))
print(qnorm(0.025,0,1))
print(qnorm(0.95,0,1))
print(qnorm(0.975,0,1))
```

```
> print(qnorm(0.05,0,1))
[1] -1.644854
> print(qnorm(0.025,0,1))
[1] -1.959964
> print(qnorm(0.95,0,1))
[1] 1.644854
> print(qnorm(0.975,0,1))
[1] 1.959964
```

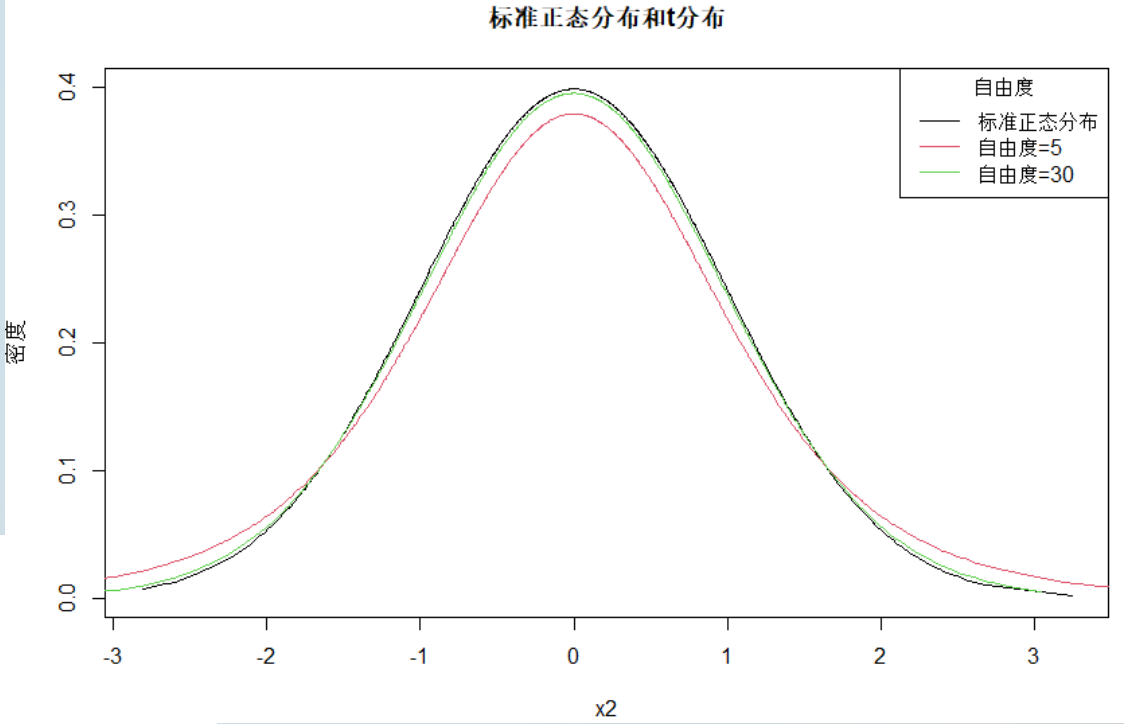
“3σ准则:“68-95-99.7规则”

第4讲 R的基本分析和统计图形

- 有关单个数值变量的其他R内容：随机数和分布
  - 例如：t分布:对称分布，比正态分布平坦。依赖于自由度，随自由度增大，分布也逐渐趋于正态分布



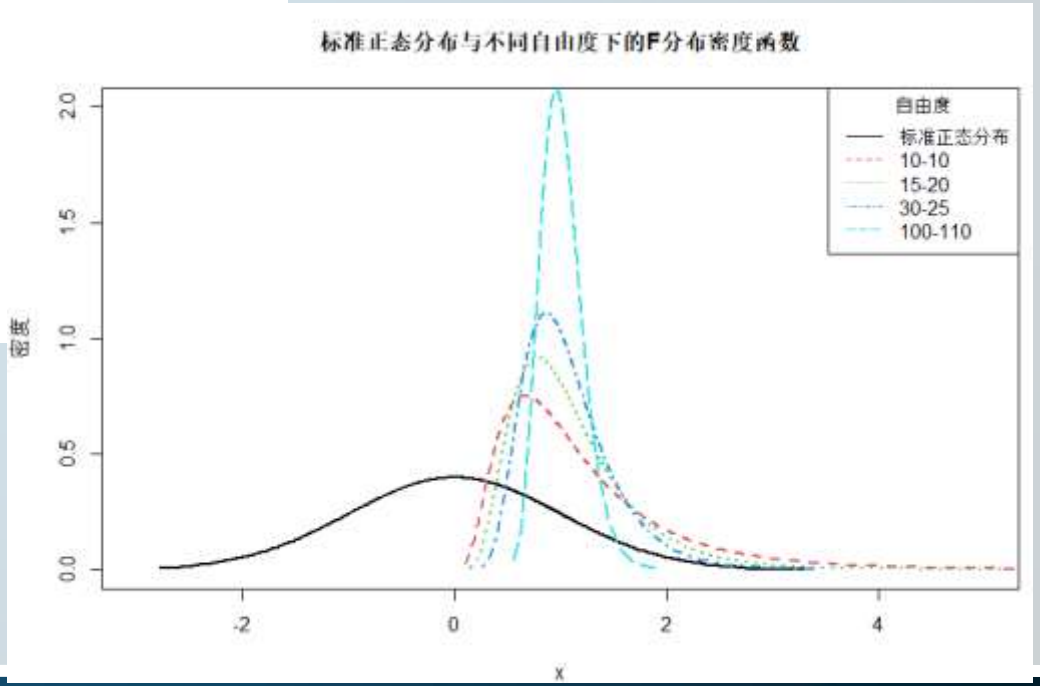

```
###################################正态分布和不同自由度下的t分布
set.seed(123)
x2<-rnorm(1000,0,1)
x2<-x2[order(x2)]
plot(x2,dnorm(x2,0,1),type="l",col=1,main="标准正态分布和t分布",ylab="密度")
x3<-rt(1000,5)
x3<-x3[order(x3)]
lines(x3,dt(x3,5),col=2)
x3<-rt(1000,30)
x3<-x3[order(x3)]
lines(x3,dt(x3,30),col=3)
legend("topright",title="自由度",c("标准正态分布","自由度=5","自由度=30"),lty=c(1,1,1),col=c(1,2,3))
```



第4讲 R的基本分析和统计图形

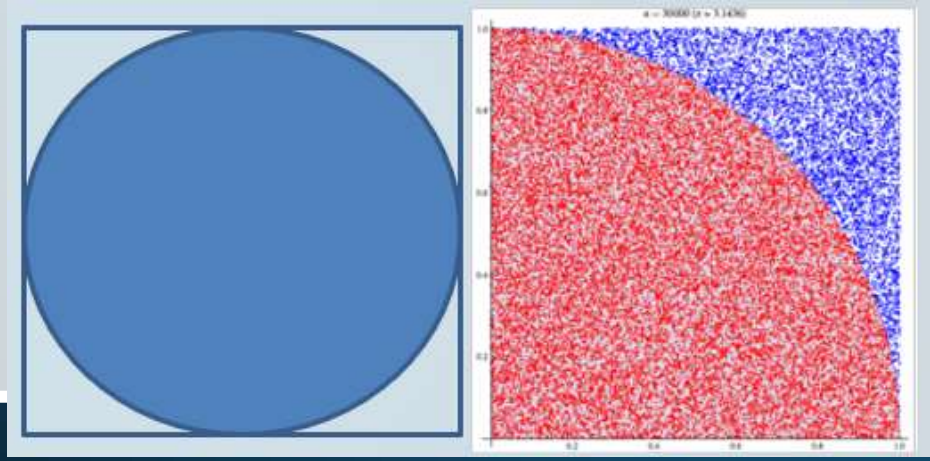
- 有关单个数值变量的其他R内容：随机数和分布
  - 例如：F分布:非对称分布，两个自由度参数

```
###################################正态分布和不同自由度下的F分布
set.seed(12345)
x<-rnorm(1000,0,1)
ord<-order(x,decreasing=FALSE)
x<-x[ord]
y<-dnorm(x,0,1)
plot(x,y,xlim=c(-3,5),ylim=c(0,2),type="l",ylab="密度",main="标准正态分布与不同自由度下的F分布密度函数",lwd=2)
#######不同自由度的F分布
df1<-c(10,15,30,100)
df2<-c(10,20,25,110)
for(i in 1:4){
  x<-rf(1000,df1[i],df2[i])
  ord<-order(x,decreasing=FALSE)
  x<-x[ord]
  y<-df(x,df1[i],df2[i])
  lines(x,y,col=i+1,lty=i+1,lwd=2)
}
legend("topright",title="自由度",c("标准正态分布",paste(df1,df2,sep="-")),lty=1:5,col=1:5)
```



R编程延展

- 示例：利用随机模拟估计π的近似值
  - 蒙特卡洛随机模拟:利用随机数进行计算机随机模拟的方法。对研究对象进行随机抽样，通过对样本值的观测统计，求得所研究系统的某些参数
- 利用随机投点法实现
  - 内接圆的面积是正方形(边长为1)面积的π/4：s1/s2=π/4
  - 随机生成在[0,1]区间内的30000个点。以点的个数作为面积的测度
  - 统计落入曲线(圆)内的点数
  - 点数的比率乘以4得到π的近似值



```
##############求解pi
Pi<-vector(length=500)
set.seed(123456)
for(i in 1:500){
  X<-runif(30000,0,1)
  Y<-runif(30000,0,1)
  Z<-X*X+Y*Y  #利用圆的标准方程(x-a)^2+(y-b)^2=r^2,a=b=0
  Data<-data.frame(x=X,y=Y,z=Z)
  if(i==1){  #只在第一次显示图
    plot(Data$x,Data$y,cex=0.2,col=ifelse(Data$z<=1,1,2))
  }
  SubSet<-subset(Data,Data$z<=1)
  Pi[i]<-4*length(SubSet$x)/30000
}
print(mean(Pi))
```

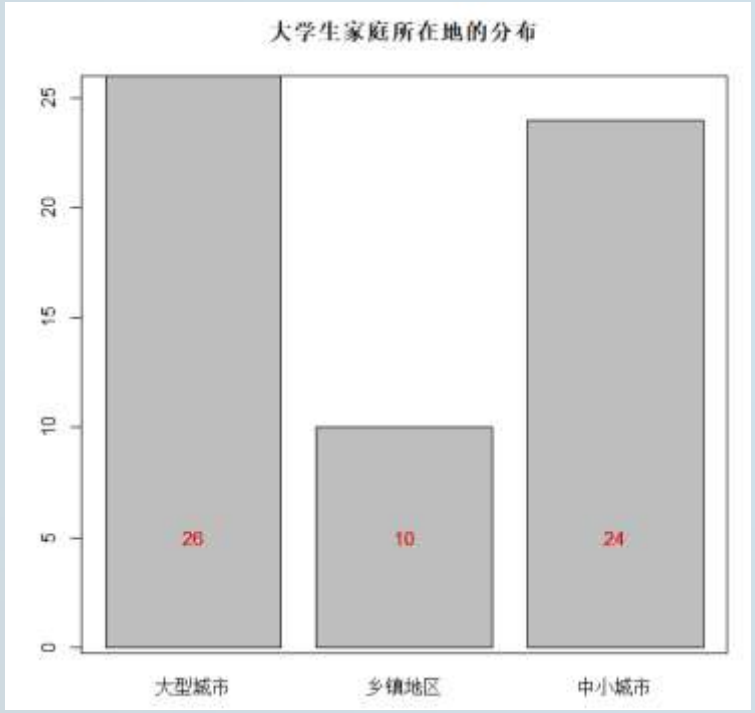
课后练习

- 对于学生成绩数据(ReportCard.txt)进行如下计算：
  - 分类变量的描述统计
    - 分析平均分等级的分布
    - 绘制平均分等级的各种统计图形
  - 数值型变量的描述统计
    - 计算政治课程的描述统计量
    - 利用sapply函数计算每门课程的平均分
    - 编写用户自定义函数(Des.Fun)，计算每门课程的描述统计量
    - 分性别计算各门课程的描述统计量
    - 绘制数学成绩的统计图形(直方图并添加核密度估计曲线、箱线图)

xh	sex	poli	chi	math	fore	phy	che	geo	his	sumScore	avScore
92101	F	96	96	87.5	72	93	65	76	92	677.5	B
92102	M	94	97	86.5	61	93	64	79.5	95	670	B
92103	F	NA	NA	NA	66	98	79	89	81	NA	NA
92104	F	89	97	69.5	86	83	62	83	94	663.5	B
92105	M	82	85	79.5	60	88	66	72.5	98	631	C
92106	F	88	88	78	60	90	70	81.5	77	632.5	C
92108	F	84	90	69.5	50	80	60	86.5	94	614	C
92110	M	92	94	71	65	78	62	83	87	632	C
92111	M	61	86	74	51	74	61	76	91	574	C
92112	F	81	75.5	76.5	43	78	83	78	91	606	C
92113	M	70	85	66	63	86	65	64	84	583	C
92115	M	85	91	72.5	63	70	50	65	82	578.5	C
92116	F	84	87	67.5	52	82	60	79	71	582.5	C
92117	M	83	91	80.5	44	89	56	55	62	560.5	C
92120	M	90	84	55	50	82	60	67.5	81	569.5	C
92122	M	80	88.5	63.5	44	75	70	61.5	66	548.5	D
92124	M	82	76	61	42	80	50	60	87	538	D
92125	F	88	80	53.5	60	78	51	69	94	573.5	C
92126	M	70	92	56	40	66	56	70.5	91	541.5	D
92127	F	87	97	52	38	76	47	79	84	560	C

课后练习

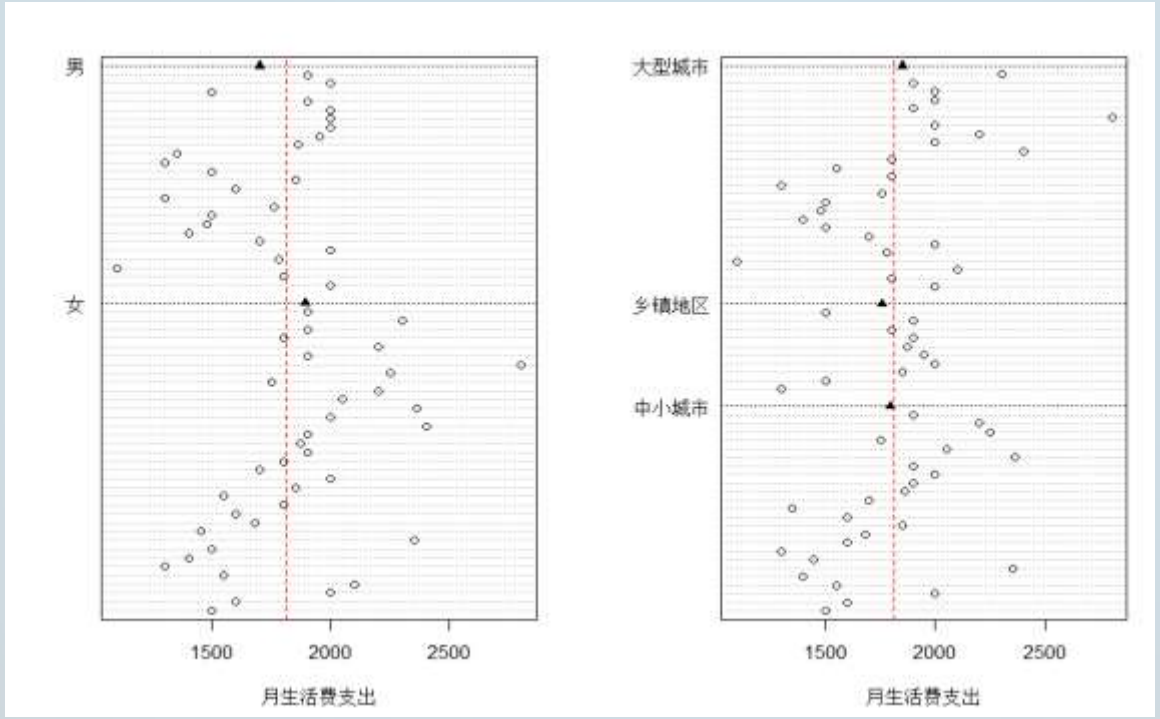
- 对于大学生综合调查数据(大学生综合调查.txt)做以下基本分析：
  - 对大学生的家庭所在地进行分析(统计量，图形)

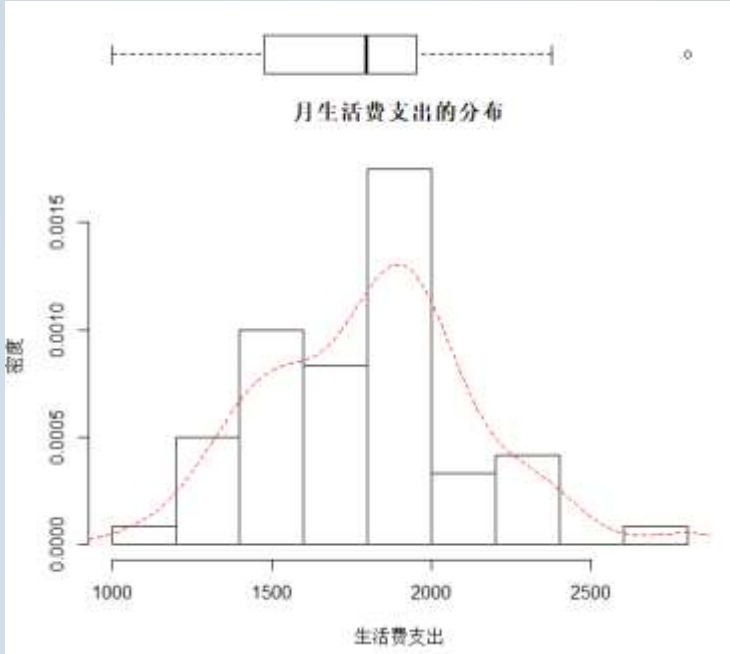



```
"性别" "家庭所在地" "月生活费支出"
"女" "中小城市" 1500
"男" "大型城市" 2000
"男" "大型城市" 1800
"女" "中小城市" 1600
"女" "中小城市" 2000
"女" "大型城市" 2100
"男" "大型城市" 1100
"男" "大型城市" 1780
"女" "中小城市" 1550
"女" "乡镇地区" 1300
"男" "大型城市" 2000
"男" "大型城市" 1700
"女" "中小城市" 1400
"女" "大型城市" 1500
"男" "大型城市" 1400
"男" "大型城市" 1480
"女" "中小城市" 2350
"女" "中小城市" 1450
"男" "大型城市" 1500
"男" "大型城市" 1760
"男" "中小城市" 1300
"男" "中小城市" 1600
```

课后练习

- 对大学生的月生活费支出进行分析(统计量，图形)
  - par(fig=c(0,1,0,0.8))；par(fig=c(0,1,0.6,1),new=TRUE)
  - 判断是否存在离群点
- 分男生和女生、家庭所在地分别对生活费支出进行分析(统计量，图形)





课后练习

- 针对标准正态分布，你可通过模拟验证3σ准则吗？