数据分析与处理技术

探索性分析

南京审计大学商学院物流管理系

数据的计量类型

数据集可以看作**数据对象**的集合。

- 数据对象有时也称作：记录、点、向量、模式、事件、样本、观测或实体

数据对象(data object)用一组刻画对象基本特征的符号描述事物**属性(attribute)**

- **属性(attribute)**是事物的客观性质或特征；**属性值(attribute value)**是描述属性的符号
- 属性值被赋予属性来**度量(measure)**事物的特征
- 属性值可以是数字或非数字符号

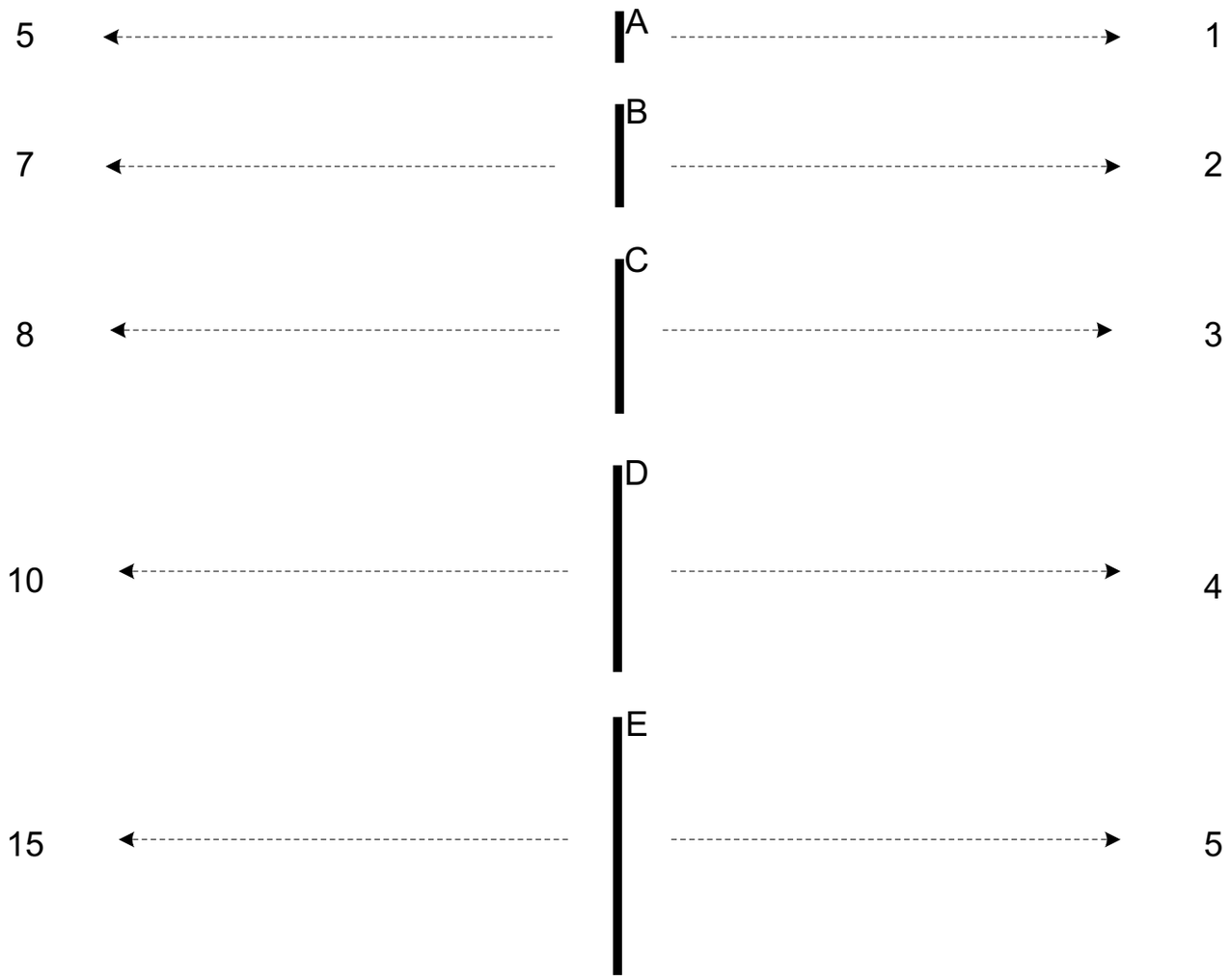
Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

长度属性的度量

同一属性可以有不同度量方法，但未必满足使用需求



不同度量方法影响可使用的运算：

- 分类运算
- 比较运算
- 加法运算
- 乘法运算

参考资料https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

属性值类型

1. 标称属性
   颜色、邮政编码
2. 有序属性
   年级、职称
3. 区间属性
   日期间隔、温度区间
4. 比率属性
   温度、长度

属性类型		描　述	例　子	操　作
分类的 (定性的)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 （=，≠）	邮政编码、雇员ID号、眼球颜色、性别	众数、熵、列联相关、χ^2检验
	序数	序数属性的值提供足够的信息确定对象的序 （<，>）	矿石硬度、{好，较好，最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定量的)	区间	对于区间属性，值之间的差是有意义的，即存在测量单位 （+，−）	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、*t*和*F*检验
	比率	对于比率变量，差和比率都是有意义的 （*，/）	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差

探索性数据分析(Exploratory Data Analysis)

探索性分析是通过分析数据集以决定哪些方法时候推断、建模的过程，也称为**描述性统计分析**

事物属性处于什么水平?
属性差异性有多大?
是否呈现正态分布特征?
分布状况是否存在偏态?
平均水平是否具有代表性?
属性之间是否存在影响以及分组特征?
……

数据的分布特征

频数：观察对象出现某属性值的次数

频率：频数占观察对象总数的百分比

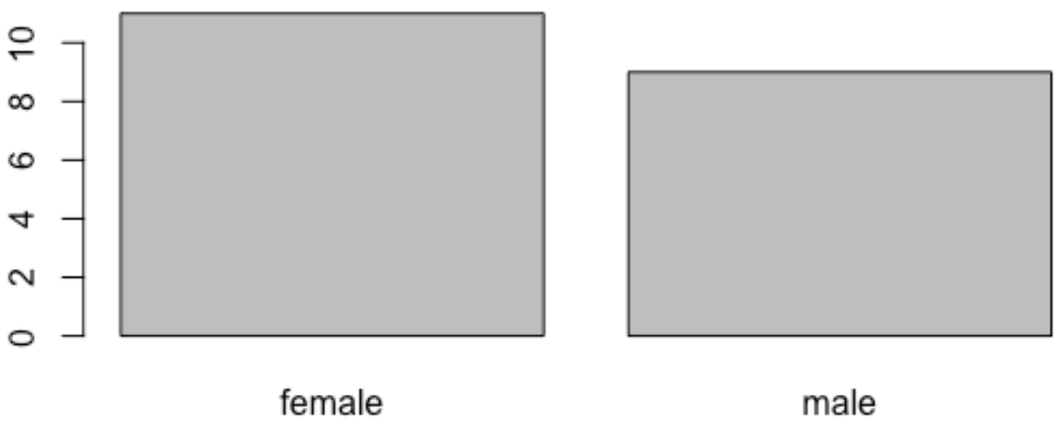
常见的频率：球队胜率、学生就业率

其中频率最高的属性值称之为**众数**

```
> table(titanic$Sex)

female   male
    11      9

> prop.table(table(titanic$Sex))

female   male
  0.55   0.45
```

年级	人数	频率
一年级	200	0.33
二年级	160	0.27
三年级	130	0.22
四年级	110	0.18

条形图(barplot)




```
> barplot(table(titanic$Sex))
```

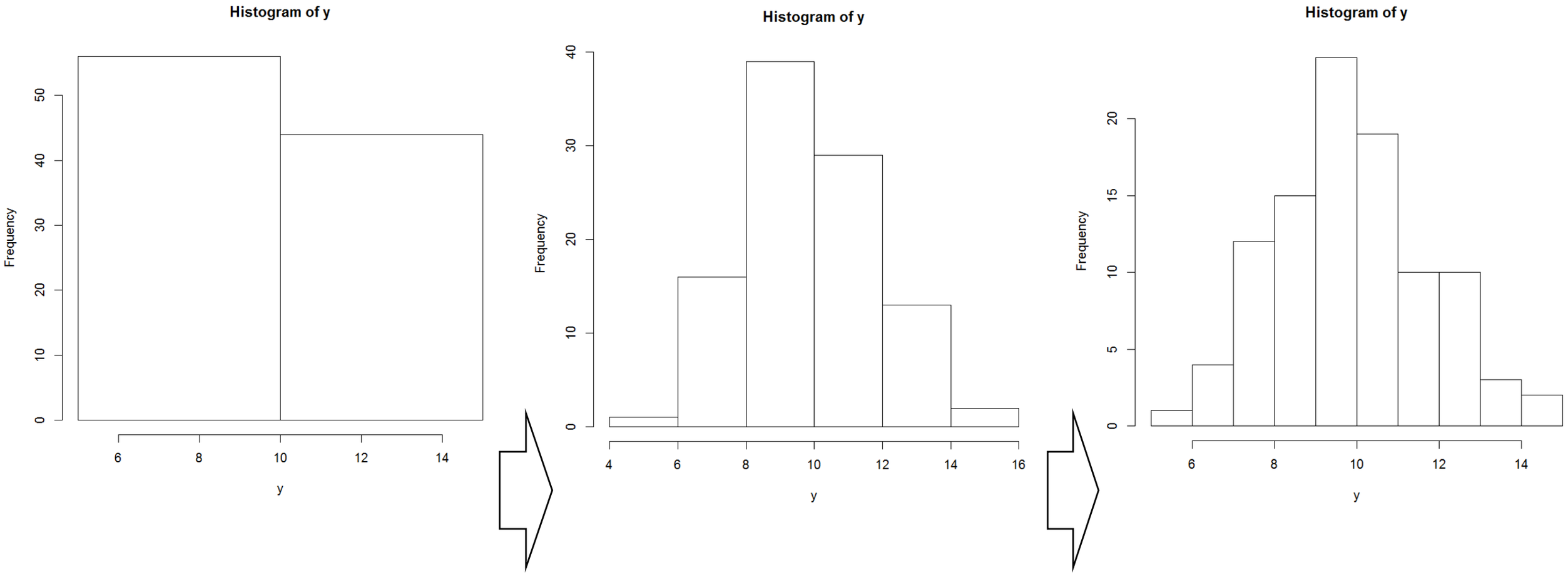
标称型和有序型属性可以方便的直接按属性值分组计算频数，而比率型则需要先对属性值进行分组统计

茎叶图(stem)

```
> stem(mtcars$mpg)

  The decimal point is at the |

  10 | 44
  12 | 3
  14 | 3702258
  16 | 438
  18 | 17227
  20 | 00445
  22 | 88
  24 | 4
  26 | 03
  28 |
  30 | 44
  32 | 49
```

直方图(hist)，对连续数值型的属性进行分组统计，观察没个区间上的频率

例如：利用直方图分析某连续数值型的变量y的分布情况

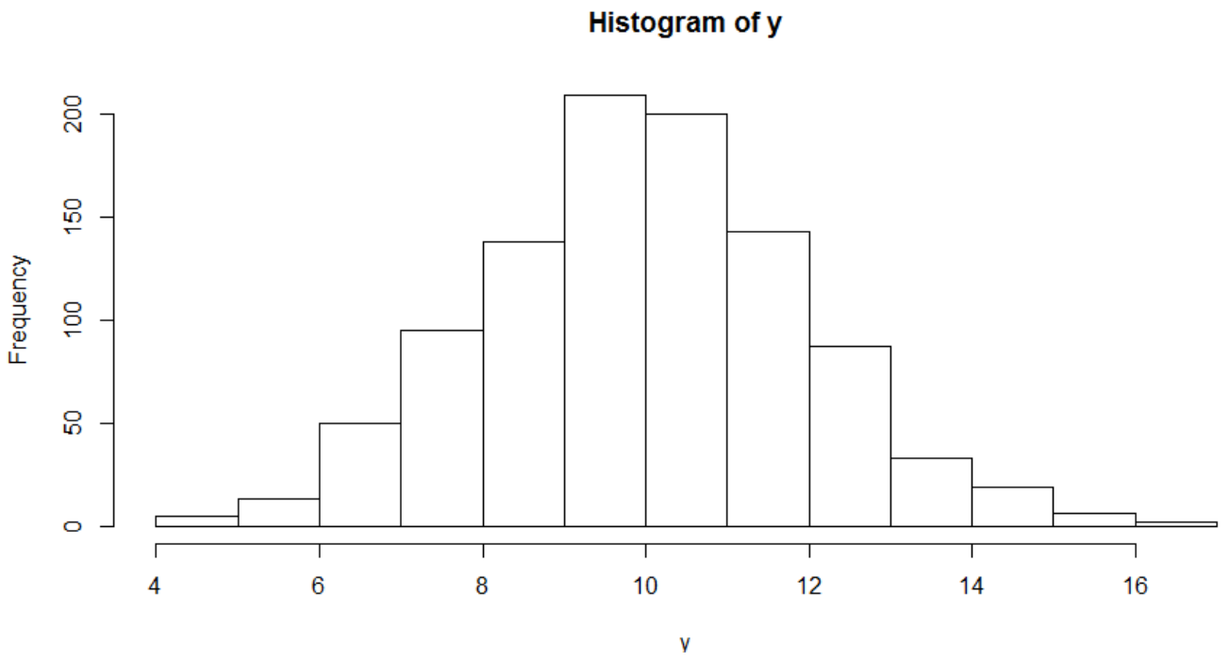


随着分组标志变量的区间缩小，直方图越来越密集，分布特征也越来越清晰

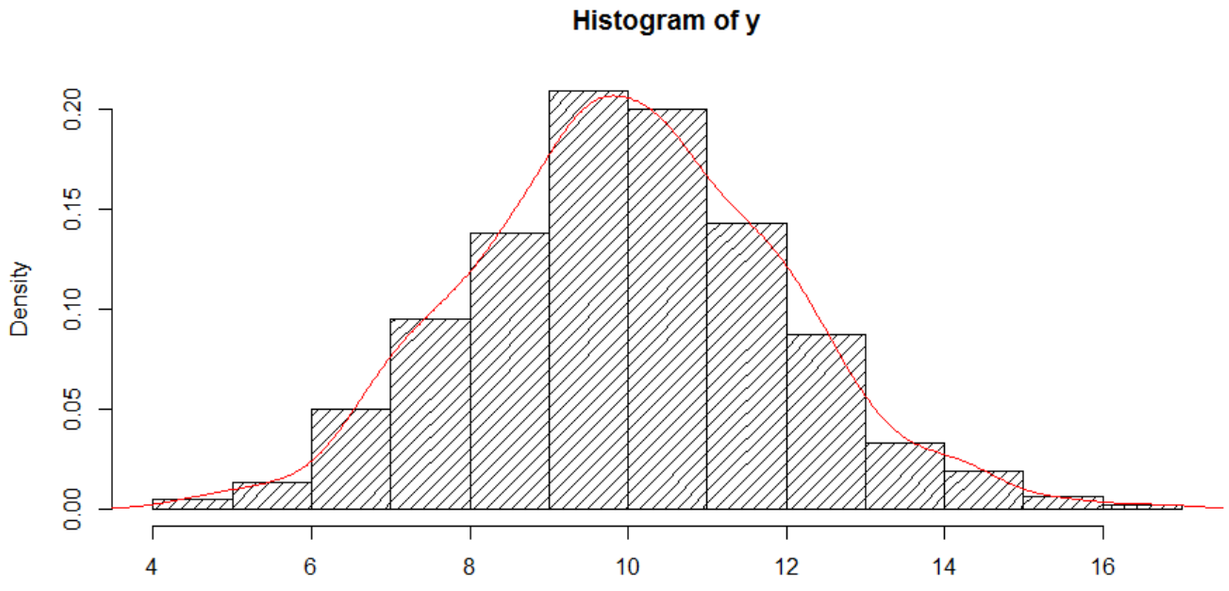
```
> hist(y,breaks = 2)
> hist(y,breaks = 5)
> hist(y,breaks = 9)
```

频数变成频率

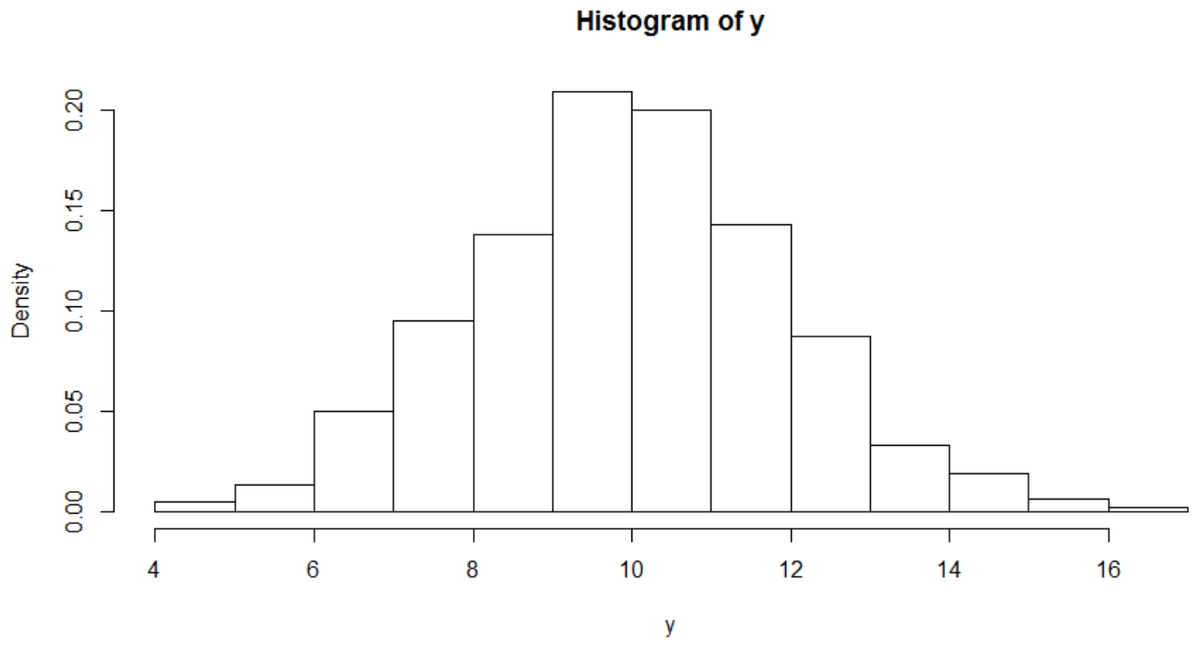
```
> hist(y)
```



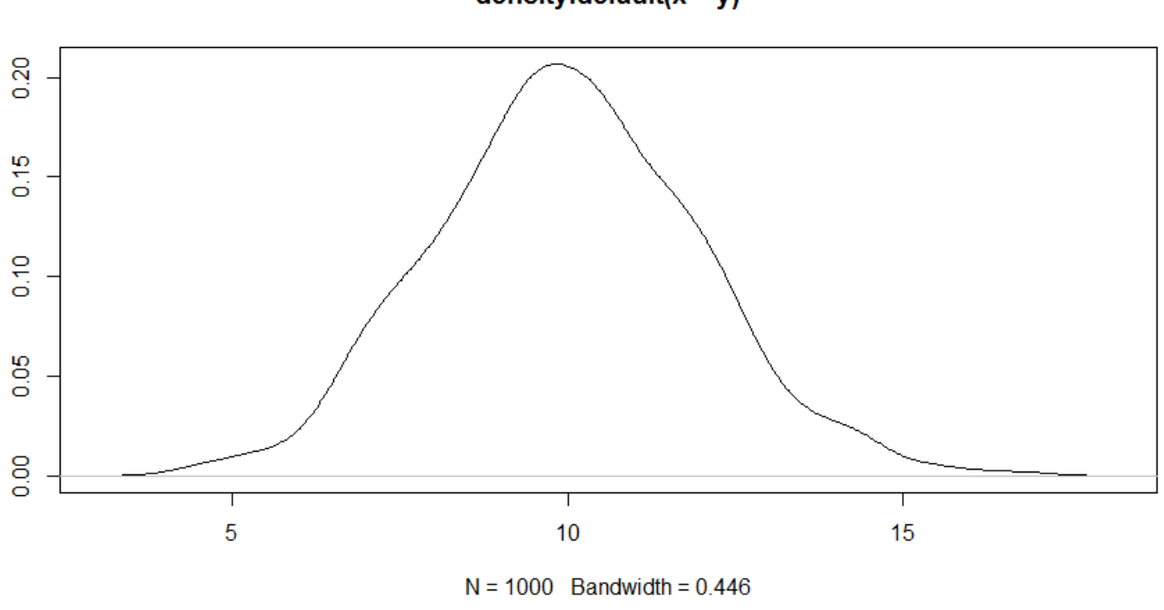

```
> hist(y,freq = F)
> lines(density(y),col='red')
```




```
> hist(y,freq = F)
```




```
> plot(density(y))
```



数据的集中程度

平均值衡量变量水平集中在哪个水平?

了解一个班级数学程度
进入一家饭店需要掌握消费水平在什么程度
……

算术平均值 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

```
> mean(iris$Sepal.Length)
[1] 5.843333
```

几何平均值 $\bar{x} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$

企业销售额3年增长速度5%、6%和7%，三年平均增速是多少

利息3.5%，5年平均利息是多少?

.....

psych工具包计算几何平均值

```
> library(psych)
> geometric.mean(c(0.05,0.06,0.07))
[1] 0.05943922
```

中位数

$$median(x) = x_{50\%} = \begin{cases} x_{(r+1)} & n\text{是奇数, n=2r+1} \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & n\text{是偶数, n=2r} \end{cases}$$

```
> median(mtcars$mpg)
[1] 19.2
```

p分位数

从1到10的整数的百分位数$x_{0\%}, x_{10\%},\ldots, x_{90\%}, x_{100\%}$依次为：1.0, 1.5, 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, 8.5, 9.5, 10.0。其中，$\min(x) = x_{0\%}$，而$\max(x) = x_{100\%}$。

```
> quantile(mtcars$mpg)
    0%    25%    50%    75%   100%
10.400 15.425 19.200 22.800 33.900
```

```
> fivenum(mtcars$mpg)
[1] 10.40 15.35 19.20 22.80 33.90
```

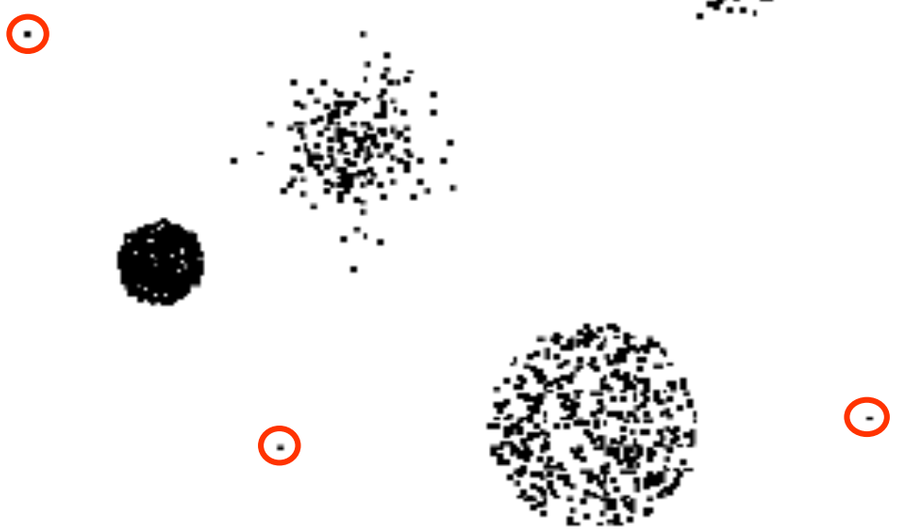
```
> quantile(mtcars$mpg,probs = c(0.3,0.7))
  30%   70%
15.98 21.47
```

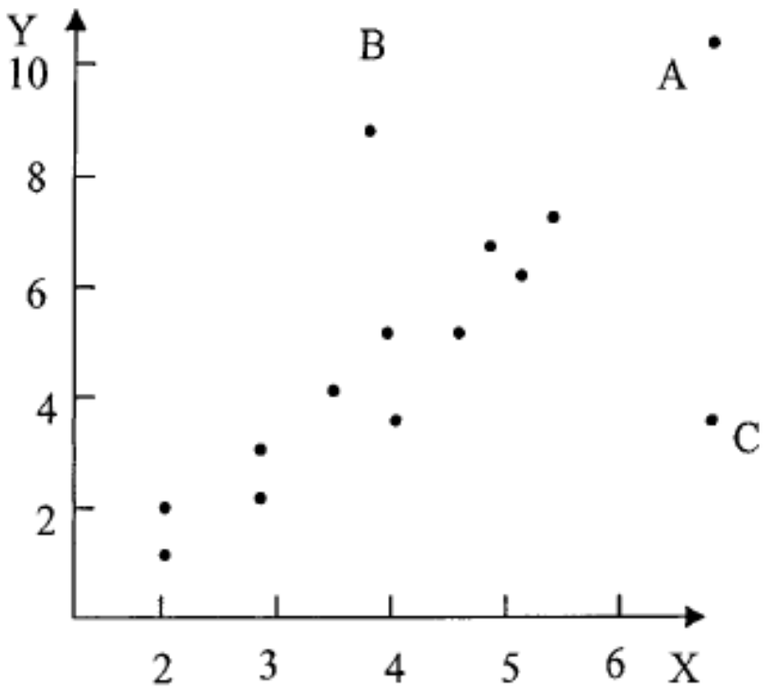
离群值问题

离群值是否等于错误值?

离群值对分析数据有什么影响?

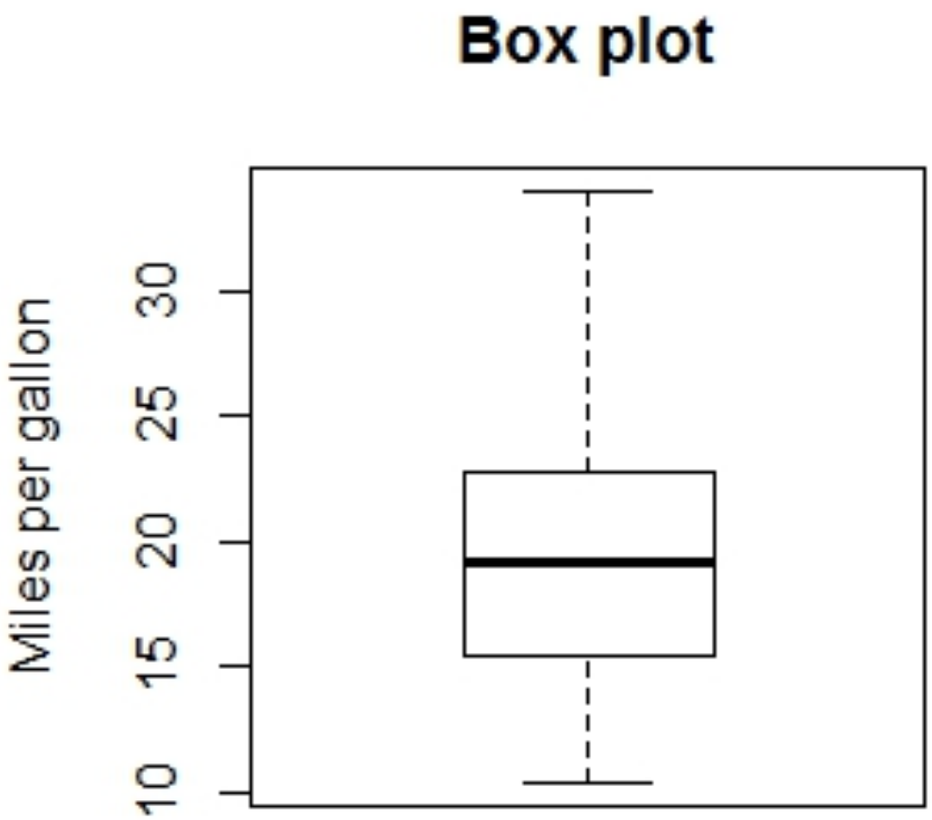
利用点图plot或points分析离群值

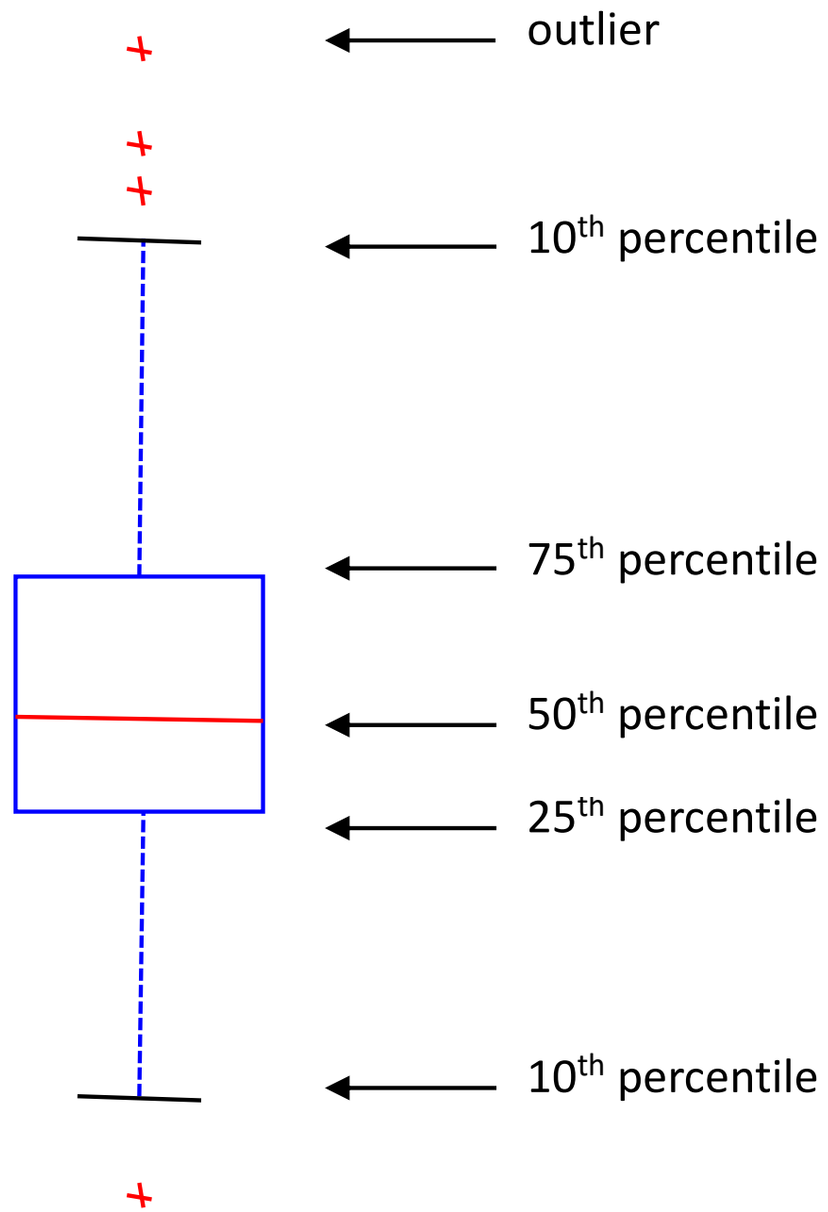




利用箱图分析数据分布结构

```
> boxplot(mtcars$mpg)
```





箱图含义

集中性与异常值

均值极易受异常值影响，尤其是算术均值。异常值的存在使得均值代表性出现明显下降，即出现“被平均”的感觉。

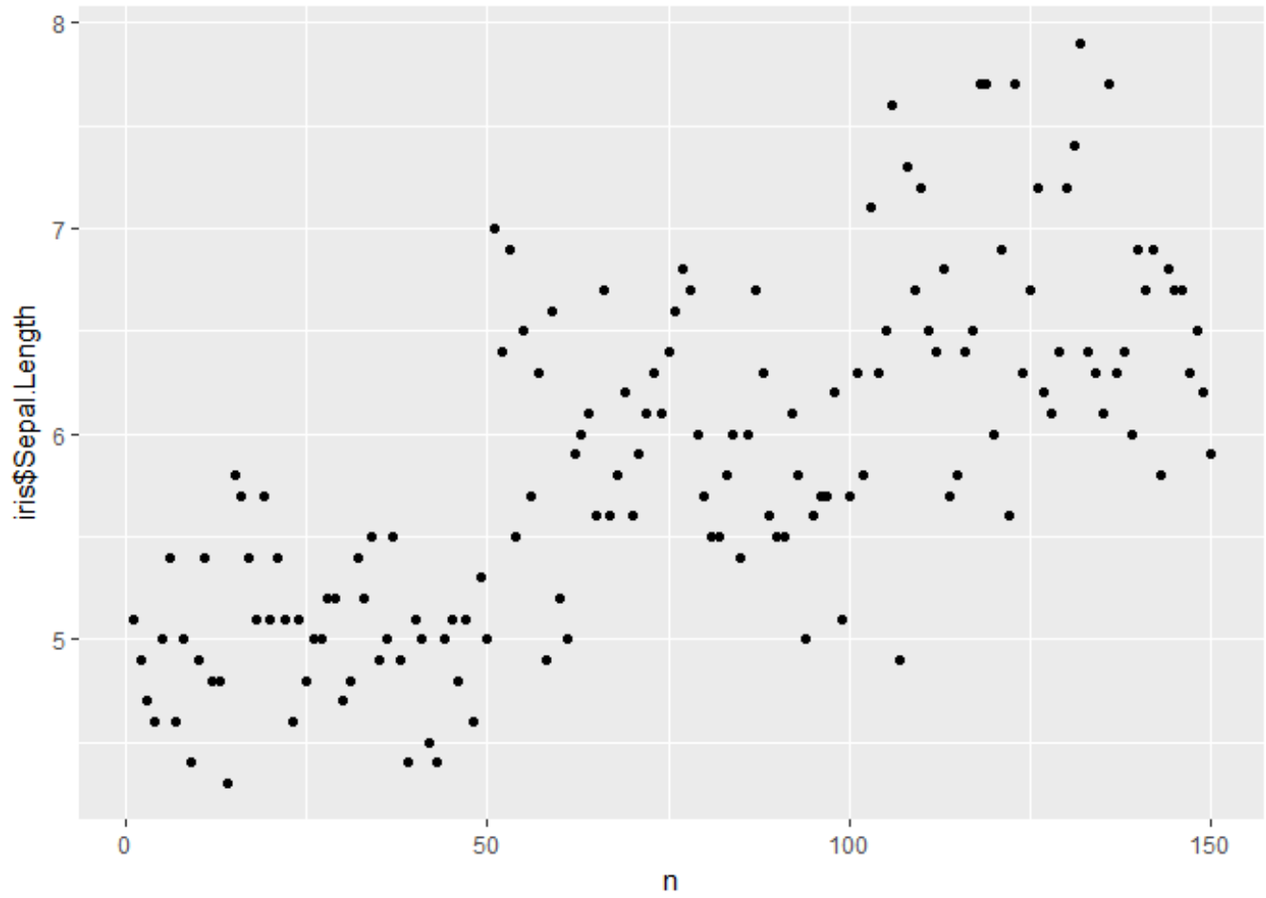
截断均值：为了排除异常值的干扰，使用截取上下p分位数以内的部分做均值

体育比赛打分时，去掉一个最高分和一个最低分，剩下得分求平均

```
> mean(mtcars$mpg,trim = 0.1)
[1] 19.69615
```

数据的离散程度

数据的分散程度是探索数据的另一个重要角度，但离散程度的描述通常需要一个参照，即平均值。



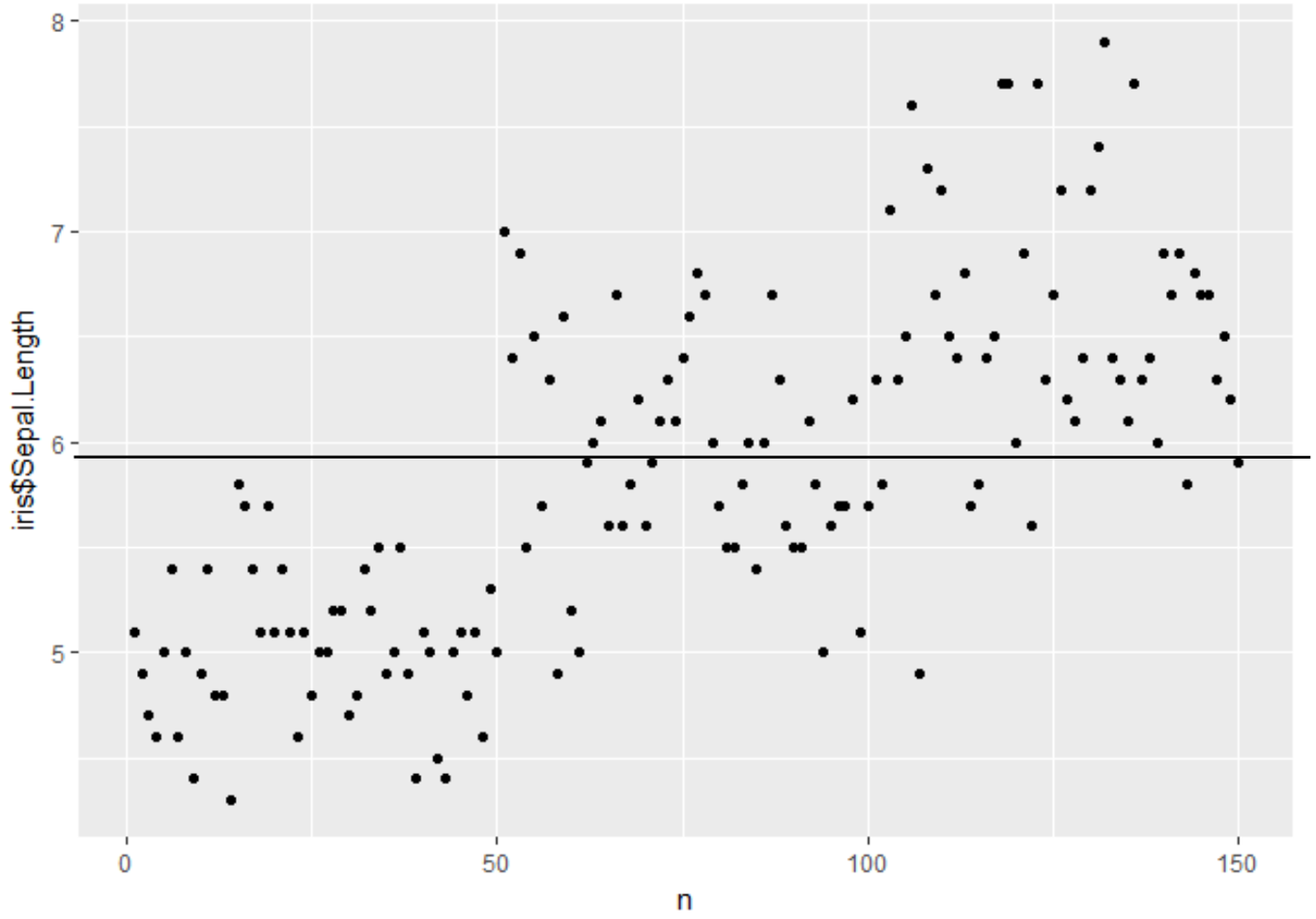
为什么使用方差和标准差

离差 $$d_i = x_i - \bar{x}$$

平均差 $$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

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

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



方差与标准差的计算

```
> var(mtcars$mpg)
[1] 36.3241
> sd(mtcars$mpg)
[1] 6.026948
```

差异性与异常值

在衡量差异性的方法中，极差、标准差等也是易受异常值影响。为了弱化异常值效果，可以采用**分位差**。

分位差

$$\mathrm{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

四分位差

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$