数据分析与处理技术

随机抽样与样本检验

南京审计大学商学院物流管理系

随机数生成

目前任何计算机和软件均无法生成真正的随机数，只能利用一些算法做出较为逼真的伪随机数。
R中生成随机数依靠随机种子(seed)，可以人为设定随机种子以达到计算可以复现的目的。

设置随机种子 `> set.seed(10086)`

正态分布随机数

```
> rnorm(100,mean = 15,sd=6)
```

均匀分布随机数

```
> runif(100,max = 20,min = 9)
```

卡方分布随机数

```
> rchisq(100,df=5)
```

其他分布随机数函数

rexp	指数分布
rf	F分布
rgamma	Gamma分布
rgeom	几何分布
rhyper	超几何分布
rlnorm	对数正态分布
rlogis	Logistic分布
rmultinom	多项分布
rnbinom	负二项分布
rpois	泊松分布
rt	t分布
rchisq	卡方分布

所有分布有关函数都有四种形态，d-密度函数,p-分布函数,q-分位数函数和r-随机数

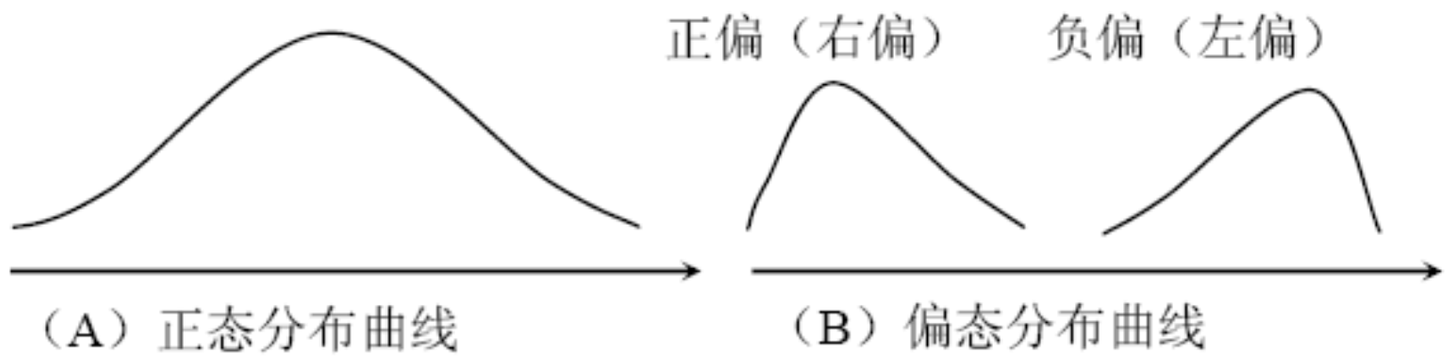
正态总体检验

正态分布描述了大多数现象的常见规律。但数据是否真的来自正态总体这一事实直接影响了后续的数据研究方法。

问题：如何检验一个变量的数据是来自正态总体的?

偏度与峰度检验

数据的**偏度**



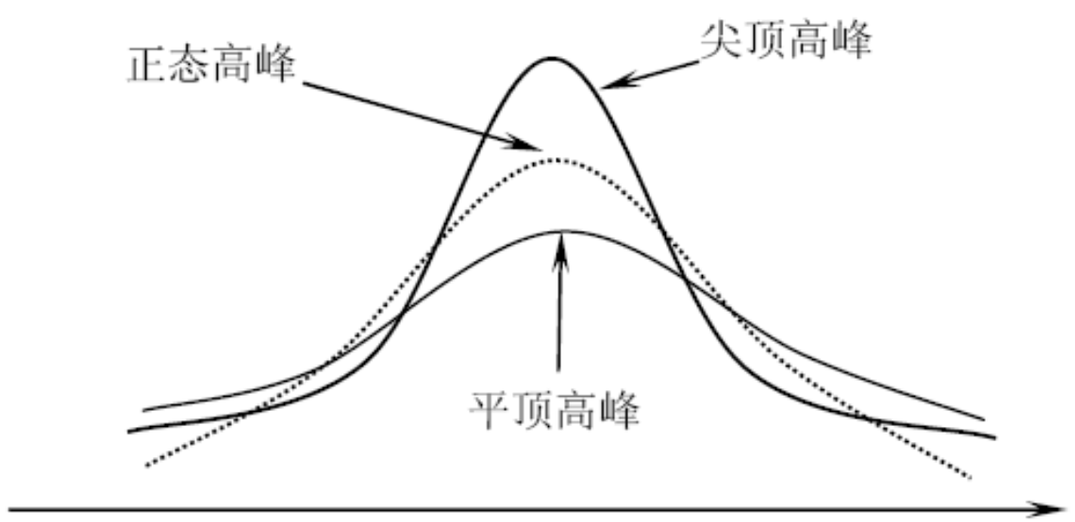

```
> library(moments)
> skewness(mtcars$mpg)
[1] 0.6404399
```

$$\alpha=\frac{\sum(x_i-\bar{x})^3/n}{[\sqrt{\sum(x_i-\bar{x})^2/n}]^3}\begin{cases}>0, & \text{右偏}\\ =0, & \text{正态或对称}\\ <0, & \text{左偏}\end{cases}$$

数据的**峰度**

```
> kurtosis(mtcars$mpg)
[1] 2.799467
```

moments包计算峰度没有减3，因此它的峰度以3为参照值。



$$\beta=\frac{\sum(x_i-\bar{x})^4/n}{[\sum(x_i-\bar{x})^2/n]^2}-3 \quad \begin{cases} >0, & \text{尖顶高峰} \\ =0, & \text{正态高峰} \\ <0, & \text{平顶高峰} \end{cases}$$

Jarque-Bera检验：偏度与峰度的联合检验

```
> library(tseries)

> jarque.bera.test(mtcars$mpg)

        Jarque Bera Test

data:  mtcars$mpg
X-squared = 2.2412, df = 2, p-value = 0.3261
```

对于随机数y

将其排序 $y_{(1)}, y_{(2)}, \cdots y_{(i)} \cdots y_{(n)}$

若y来自正态总体 $y \sim N(\mu, \sigma^2)$

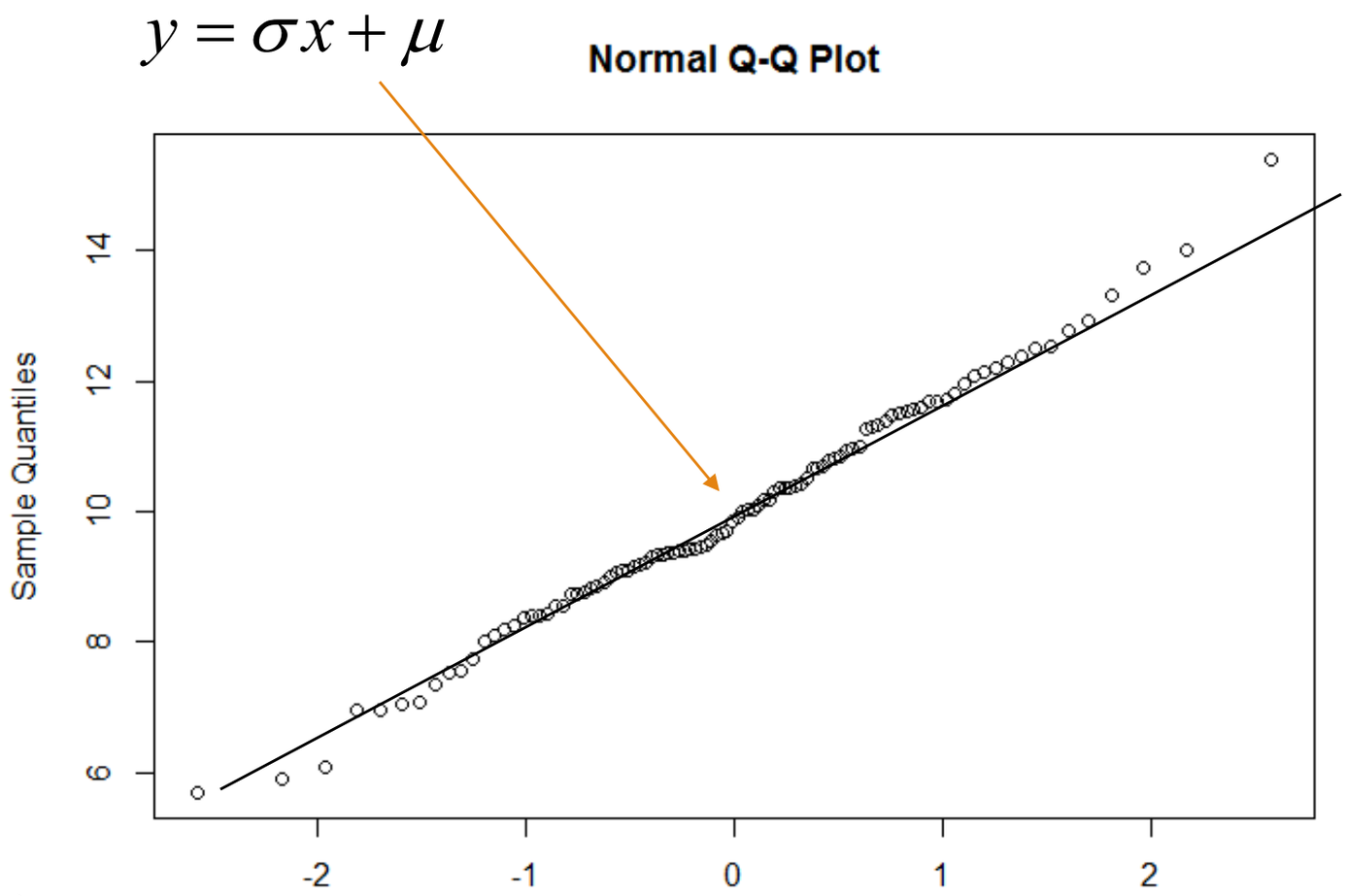
那么应该存在理论值

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2} \right)$$

反解之后组合成点 $\left(f^{-1}(\frac{i-0.375}{n+0.25}), y_{(i)} \right)$

标准正态分布应当在Q-Q图中形成严格直线。

$y = \sigma x + \mu$




```
> qqnorm(mtcars$mpg)
> qqline(mtcars$mpg)
```

什么是卡方分布：

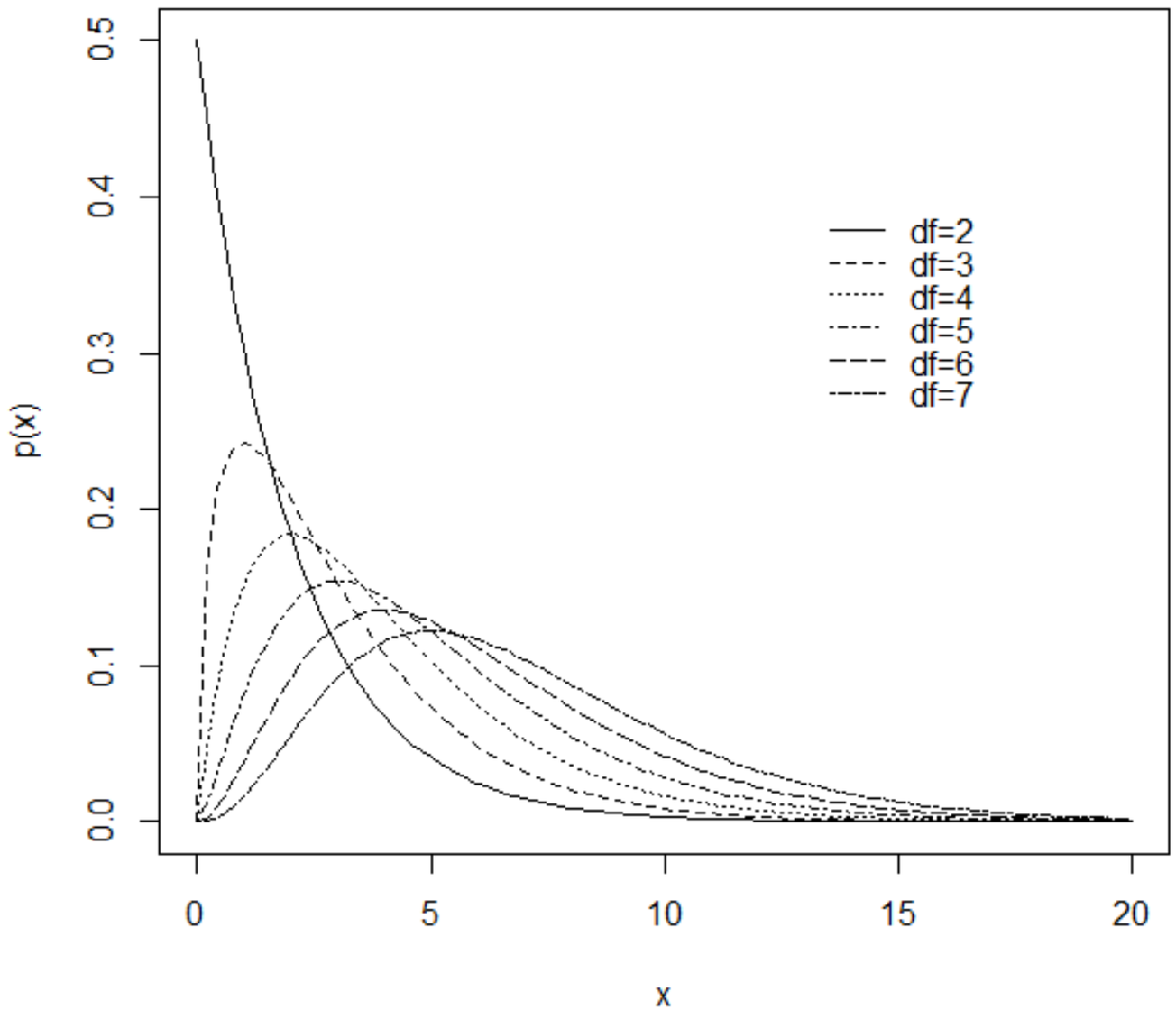
$$\chi^2=\sum_{i=1}^{n} y_i^2, y_i \sim N(0,1)$$

卡方分布是由n个标准正态分布加和构成，它的关键参数是自由度，即几个正态变量构成。

自由度参数　$df = n$

则其期望　$E(\chi^2)=n$

而方差则　$D(\chi^2)=2n$



正态样本的方差服从卡方分布，因此质量管理中常用卡方分布检查方差是否符合要求

当自由度n->∞时，卡方分布趋近于正态分布

拟合优度检验 课本p146案例

```
> freq=c(22,21,22,27,22,36)
> probs=c(1,1,1,1,1,1)/6
> chisq.test(freq,p=probs)

        Chi-squared test for given probabilities

data:  freq
X-squared = 6.72, df = 5, p-value = 0.2423
```

卡方可以用来检验事物是否与它应有的理论概率相符合，即数据是否来自正态总体。

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

正态分布：正态分布代表了属性正常时水平，虽然存在一定随机性，但随着观察对象的增长最终属性平均水平还是会趋近其理论水平。

卡方检验：事物属性出现哪个属性值是随机的，但随着观察对象增多，属性出现的频率依然会呈现正态特性趋近于其理论值。

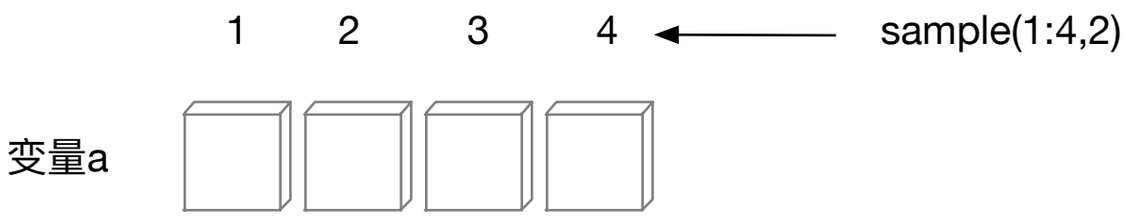
大数定律：一切随机现象重复次数越多，其平均水平越呈现出向某个具体数值(理论值)趋近的特性。

中心极限定理：大量独立随机变量之和具有近似正态的分布，其平均数是以正态分布为极限的。

随机抽样

```
> sample(1:10,size = 5)
[1]  4  9  2 10  3
```

sample随机抽取的是对象的元素编号



例如数据集mtcars，从这个32个对象11个变量的数据集中随机抽取10个样本

```
> mtcars[sample(1:nrow(mtcars),10),]
```

可重复的随机抽样

```
> sample(1:10,5,replace = T)
[1]  4  6  6  1 10
```

单变量分析-假设检验

参考课本第七章

当样本达到一定数量时，属性观测值的平均水平应该在其理论水平附近，如果n样本的平均值落入了几乎不可能出现的范围内，则有足够的可信度拒绝原假设。

问题：生产直径5.00mm的生产线是否合格?

原假设 $H_0:\mu=5$

对立的备择假设 $H_1:\mu\neq5$

默认参数为检验 $\mu=0$

```
> t.test(product)

        One Sample t-test

data:  product
t = 21.564, df = 19, p-value = 8.072e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 4.628905 5.624095
sample estimates:
mean of x
   5.1265
```

设置参数检验 $\mu=5$

```
> t.test(product,mu=5)

        One Sample t-test

data:  product
t = 0.53209, df = 19, p-value = 0.6008
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.628905 5.624095
sample estimates:
mean of x
   5.1265
```

t.test( )函数参数

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

单边检验

```
> t.test(x,alternative = "greater",mu=10)
```

分组对照检验

sleep数据集：两组服用不同药物的病人睡眠状况的记录

问题：观察值是否说明两组数据平均水平不同?

```
> t.test(sleep$extra[1:10],sleep$extra[11:20])

        Welch Two Sample t-test

data:  sleep$extra[1:10] and sleep$extra[11:20]
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean of x mean of y
     0.75      2.33

> t.test(sleep$extra~sleep$group)

        Welch Two Sample t-test

data:  sleep$extra by sleep$group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

多变量分析

前边已经介绍了许多针对单变量的分析方法和工具，这些工具可以逐个研究每个变量，却无法揭示变量之间的关系。

多变量分析重点是探索变量之间的关系，通过数据展现出的现象推断属性之间的相互影响。

标称型属性案例

观察一组数据，安装并加载工具包vcd，其中一组数据集Arthritis记录了一组治疗关节炎的记录数据。加载包后可以用data( )函数将其从包环境调取至全局环境便于观察

```
> library(vcd)
载入需要的程辑包: grid
> data("Arthritis")
> Arthritis
```

交叉表分析标称型变量

xtabs( )函数与table效果相似，但可以使用公式，使用方法更为灵活

```
> a<-xtabs(~Treatment+Sex+Improved,data=Arthritis)
> a
, , Improved = None

         Sex
Treatment Female Male
  Placebo     19   10
  Treated      6    7

, , Improved = Some

         Sex
Treatment Female Male
  Placebo      7    0
  Treated      5    2

, , Improved = Marked

         Sex
Treatment Female Male
  Placebo      6    1
  Treated     16    5
```

样本分布按Treatment和Improved进行了展开，但同时还可以加入取子集的条件

```
> xtabs( ~Treatment+Improved, data = Arthritis,subset=Sex=="Female")
         Improved
Treatment None Some Marked
  Placebo   19    7      6
  Treated    6    5     16
```

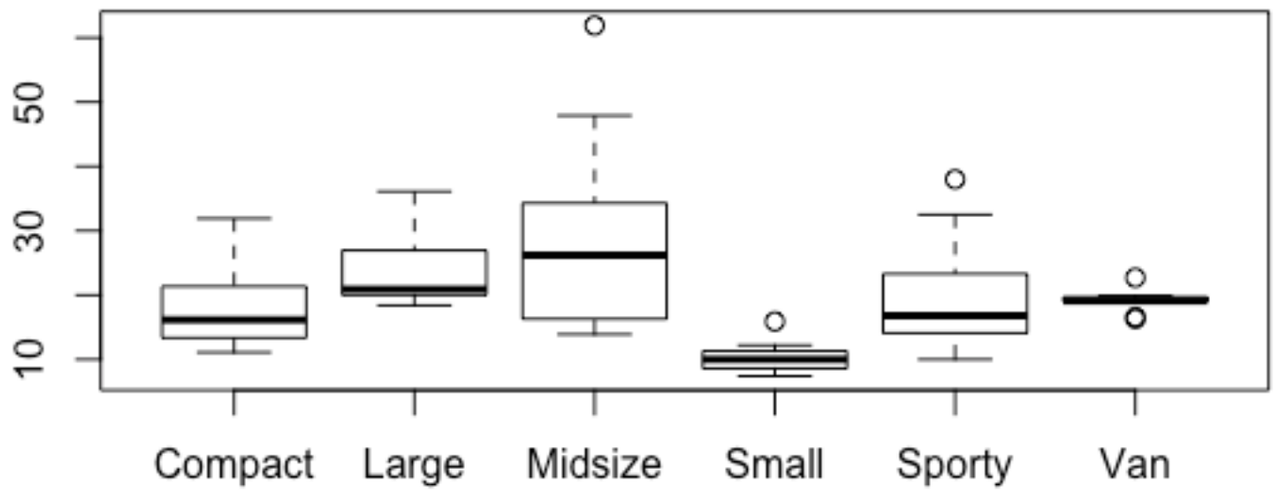
```
> ftable(a)
                 Improved None Some Marked
Treatment Sex
Placebo   Female            19    7      6
          Male              10    0      1
Treated   Female             6    5     16
          Male               7    2      5
>
```

并列箱图

当数据中既有标称型又有数值型(比率型)，则使用标称型作为分组标志，分组统计观察数值型变量有何差异。

```
> boxplot(Cars93$Price~Cars93$Type)
```

MASS包中数据



思考如何对该数按Type变量分组计算平均值，并做出条形图

公式符号 ~

R中另一种常见的运算符号其实也是函数：+ - * ^ 甚至包括许多想不到的符号，如括号(，都是函数，只是这些函数形式并非我们通常理解的左侧函数名右侧参数的形式，而是函数名在数据中间，将两边数据作为其参数。

该函数作用为创建公式，一般右侧为自变量、或称为解释变量（依据变量），而左侧则为响应变量或被解释变量，在~的函数体系当中，左右侧都被视为创建公式的参数，左侧参数根据环境有时可以空缺，而右侧参数多数不可省。

```
> a<-xtabs(~Treatment+Sex+Improved,data=Arthritis)

> boxplot(Cars93$Price~Cars93$Type)
```

相关分析

两个变量之间是否存在互动关系，通过观察两个变量增长关系推测他们之间的相关性。

例如：刚工作的大学生每月消费支出与他们的收入水平呈现相关关系；一个国家的财政收入与它的税收之间呈现出较为明显相关性，而财政收入与国家经济规模也呈现出相关性。

函数关系：有明确的映射关系，也称为确定性关系

相关关系：非确定型关系，存在随机性

相关系数代表了变量之间相关性的强弱，取值范围[-1,1]，小于0为负相关，大于0为正相关

$$\rho = \frac{Cov(X,Y)}{Var(X)Var(Y)}$$

相关系数计算方法上等于协方差除以两个变量标准差的乘积。

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

```
> cor(iris$Sepal.Length,iris$Sepal.Width)
[1] -0.1175698
```

检验相关系数的显著性，可以用cor.test

```
> cor.test(x=iris[,1],y=iris[,2])
```

iris数据集记录的鸢尾花

三种花色
- Setosa
- Virginica
- Versicolour

四个属性
- 萼片的宽度、长度
- 花瓣的宽度、长度

cor()函数同时可以计算多个变量相互间相关性，并构成相关性矩阵，而corrplot则提供了对**相关系数矩阵**作图的专用语句。

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

```
> library(corrplot)
> corrplot(cor(iris[,1:4]))
```

```
> corrplot(cor(iris[,1:4]),method='number')
```

method可以的方法还有： "circle", "square", "ellipse", "number", "shade", "color", “pie"

相关系数的特点：

只能用于定性分析，比较、排序等

没有加减运算，也不能比较差距

相关性分析以客观事物的定性分析为前提

方差分析

在单变量分析中，我们关注较多的是检验变量的平均水平。当检验的分组多于三个时就无法在使用t检验，则需要使用方差分析，而其中分组标志则归于一个因素变量当中。方差分析本质上也是一种检验均值是否一致的方法。

观察变量 ~ 因素变量

问题：检验因素变量是否对观察变量有明显作用

参见课本第九章或15级课件或任何一本讲解方差分析的教材