

# 基本统计分析

# 其他参考书

- 埃维森，统计学-基本概念和方法
  - 几乎没有公式，通俗易懂
- 贾俊平等，统计学
- 弗里德曼，统计学
- **Dawn Griffiths**，深入浅出（Head first）统计学

# 背景

- 在数据被组织成合适的形式后，开始使用图形探索数据，而下一步通常就是使用数值描述每个变量的分布，接下来则是两两探索所选择变量之间的关系。其目的是回答如下问题。
  - 各车型的油耗如何？特别是，在对车型的调查中，每加仑汽油行驶英里数的分布是什么样的？（均值、标准差、中位数、值域等。）
  - 在进行新药实验后，用药组和安慰剂组的治疗结果（无改善、一定程度的改善、显著的改善）相比如何？实验参与者的性别是否对结果有影响？
  - 收入和预期寿命的相关性如何？它是否明显不为零？
  - 美国的某些地区是否更有可能因为你犯罪而将你监禁？不同地区的差别是否在统计上显著？

# 本章内容

- 描述性统计分析
- 频数表（列联表）
- 独立性检验
- 相关性检验
- t检验
- 方差分析
- 组间差异的非参数检验

# 描述性统计分析

# 统计函数（1/2）

- `mean(x)` 平均数
  - `mean(c(1,2,3,4))`返回值为2.5
- **`median(x)`** 中位数
  - `median(c(1,2,3,4))`返回值为2.5
- `var(x)` 方差（ $\sigma^2$  (pronounced "sigma squared")）
$$\text{Var}(X) = \text{E}[(X - \mu)^2]$$
$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$
  - **`var(c(1,2,3,4))`**返回值为1.67
- `sd(x)` 标准差（ $\sigma$ ）
  - `sd(c(1,2,3,4))`返回值为1.29

# 统计函数（2/2）

- `range(x)` 求值域
  - `x <- c(1,2,3,4)` `range(x)`返回值为`c(1,4)`
- `sum(x)` 求和
  - `sum(c(1,2,3,4))`返回值为10
- `min(x)` 求最小值
  - `min(c(1,2,3,4))`返回值为1
- `max(x)` 求最大值
  - `max(c(1,2,3,4))`返回值为4

# 描述性统计分析

- 准备数据
  - mtcars数据集：Moto Trend杂志车辆路试
  - 子数据集：重点关注每加仑行驶英里数(mpg)、马力(hp)、车重(wt)。  
**mt** <- mtcars[c("mpg", "hp", "wt", "am")]
- 我们首先查看所有32种车型的描述性统计量，然后按照变速箱类型（am）和汽缸数（cyl）考察描述性统计量。
  - 变速箱类型是一个以0表示自动挡、1表示手动挡来编码的**二分变量**，而汽缸数可为4、5或6。



# 描述性统计量的计算函数（1/3）

## summary()

- 描述性统计量的计算函数极其多
- summary()函数
  - summary()函数提供了最小值、最大值、四分位数和数值型变量的均值，以及因子向量和逻辑型向量的频数统计。
  - 例：

**summary(mt)**

mpg	hp	wt	am
Min. :10	Min. : 52	Min. :1.5	Min. :0.00
1st Qu.:15	1st Qu.: 96	1st Qu.:2.6	1st Qu.:0.00
Median :19	Median :123	Median :3.3	Median :0.00
Mean :20	Mean :147	Mean :3.2	Mean :0.41
3rd Qu.:23	3rd Qu.:180	3rd Qu.:3.6	3rd Qu.:1.00
Max. :34	Max. :335	Max. :5.4	Max. :1.00

# 描述性统计量的计算函数（2/3）

## apply()

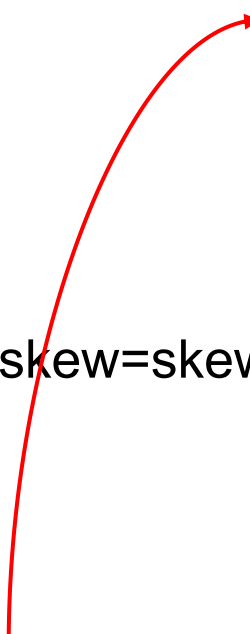
- `apply()`函数，使用格式为：  
**apply**(x, FUN, options)
- 其中的x是数据框（或矩阵），FUN为一个任意的函数。**apply对每列**应用函数FUN。如果指定了options，它们将被传递给FUN。在这里插入的典型函数有mean、sd、var、min、max、median、length、range和quantile。
- lapply returns a list of the same length as X, each element of which is the result of applying FUN to the corresponding element of X.
- apply is a user-friendly version and wrapper of lapply by default returning a vector, matrix or, if simplify = "array", is the same as lapply.

# 描述性统计量的计算函数 (3/3)

## sapply() 例

```
mystats <- function(x, na.omit=FALSE){  
  if (na.omit)  
    x <- x[!is.na(x)] #忽略缺失值  
  m <- mean(x)  
  n <- length(x)  
  s <- sd(x)  
  skew <- sum((x-m)^3/s^3)/n  
  kurt <- sum((x-m)^4/s^4)/n - 3  
  return(c(n=n, mean=m, stdev=s, skew=skew, kurtosis=kurt))  
}
```

supply对每列  
应用函数FUN。



n	32.00	32.00	32.000
mean	20.09	146.69	3.217
stdev	6.03	68.56	0.978
skew	0.61	0.73	0.423
kurtosis	-0.37	-0.14	-0.023

```
myvars <- c("mpg", "hp", "wt")
```

```
sapply(mtcars[myvars], mystats)
```

```
sapply(mtcars[myvars], mystats, na.omit=TRUE) #指定option
```

# 分组计算描述性统计量 (1/3)

## aggregate()

- 在比较多组个体或观测时，关注的焦点经常是各组的描述性统计信息，而不是样本整体的描述性统计信息。

- aggregate()函数 c("mpg", "hp", "wt") 单返回值函数  
**aggregate**(mtcars[myvars], **by=list**(am=mtcars\$am), **mean**)

	am	mpg	hp	wt
1	0	17	160	3.8
2	1	24	127	2.4

注意list(am=mtcars\$am)的使用。如果使用的是list(mtcars\$am)，则am列将被标注为Group.1而不是am。

aggregate(mtcars[myvars], by=list(am=mtcars\$am), **sd**)

#例子：求标准差

	am	mpg	hp	wt
1	0	3.8	54	0.78
2	1	6.2	84	0.62

# 分组计算描述性统计量 (2/3)

## by()

- **aggregate()**仅允许在每次调用中使用平均数、标准差这样的**单返回值函数**。它**无法一次返回若干个统计量**。要完成这项任务，可以使用**by()**函数。格式为：

**by(data, INDICES, FUN)**

- 其中**data**是一个数据框或矩阵，**INDICES**是一个因子或因子组成的列表，定义了分组，**FUN**是任意函数：单返回值函数和**多返回值函数**均可。

# 分组计算描述性统计量 (3/3)

## by() 例

#apply(x, mystats)见先前“sapply() 例”页。

**dstats** <- function(x)sapply(x, mystats)

myvars <- c("mpg", "hp", "wt")

**by**(mtcars[myvars], mtcars\$am, **dstats**)

任意函数

```
mystats <- function(x,
na.omit=FALSE){
  if (na.omit)
    x <- x[!is.na(x)] #忽略缺失值
  m <- mean(x)
  n <- length(x)
  s <- sd(x)
  skew <- sum((x-m)^3/s^3)/n
  kurt <- sum((x-m)^4/s^4)/n - 3
  return(c(n=n, mean=m, stdev=s,
skew=skew, kurtosis=kurt))
}
```

mtcars\$am: 0

	mpg	hp	wt
n	19.00000000	19.00000000	19.00000000
mean	17.14736842	160.26315789	3.7688947
stdev	3.83396639	53.90819573	0.7774001
skew	0.01395038	-0.01422519	0.9759294
kurtosis	-0.80317826	-1.20969733	0.1415676

mtcars\$am: 1

	mpg	hp	wt
n	13.00000000	13.00000000	13.00000000
mean	24.39230769	126.8461538	2.4110000
stdev	6.16650381	84.0623243	0.6169816
skew	0.05256118	1.3598859	0.2103128
kurtosis	-1.45535200	0.5634635	-1.1737358

# 频数表（列联表）

# 频数表和列联表

- **频数表**是将数据集按照某个特定列分类（分组）时观察每个类/组中数据出现次数的表。
- **列联表**（contingency tables）**也是频数表**，只不过它会分析的是将数据集按两个或两个以上类别变量联合分组时观察数据在每个分组中出现频数的表。



# 数据集

- **vcd包中的Arthritis:** 风湿性关节炎新疗法的双盲临床实验的结果。

```
library(vcd)
```

```
head(Arthritis)
```

	ID	Treatment	Sex	Age	Improved
1	57	Treated	Male	27	Some
2	46	Treated	Male	29	None
3	77	Treated	Male	30	None
4	17	Treated	Male	32	Marked
5	36	Treated	Male	46	Marked
6	23	Treated	Male	58	Marked

类别型因子：Treatment（Placebo安慰剂治疗、Treated）、Sex（Male、Female）、Improved（None、Some、Marked）。

# 一维频数表

- 可以用**table()**函数生成简单的频数统计表。
- 用**prop.table()**将这些频数转化为比例值。

```
mytable <- with(Arthritis, table(Improved))
```

```
mytable # frequencies
```

```
Improved
  None   Some Marked
   42    14    28
```

```
prop.table(mytable) # proportions
```

```
Improved
  None   Some Marked
0.500000 0.166667 0.333333
```

```
prop.table(mytable)*100 # percentages
```

```
Improved
  None   Some Marked
50.00000 16.66667 33.33333
```

# 二维列联表 (1/3)

- 对于二维列联表，**table()**函数的使用格式为：

`mytable <- table(A, B)`

其中的A是行变量，B是列变量。

- **xtabs()**函数还可使用公式风格的输入创建列联表，格式为：

`mytable <- xtabs(~A + B, data=mydata)`

其中的mydata是一个矩阵或数据框。要进行交叉分类的变量应出现在公式的右侧（即~符号的右方），以+作为分隔符。若某个变量写在公式的左侧，则其为一个频数向量。

## 二维列联表 (2/3)

```
mytable <- xtabs(~Treatment+Improved, data=Arthritis)
```

频数向量

	Improved		
Treatment	None	Some	Marked
Placebo	29	7	7
Treated	13	7	21

要进行交叉分类的变量

- 使用`margin.table()`函数生成边际和。  
`margin.table(mytable, 1) # row sums`  
`margin.table(mytable, 2) # column sums`
- 使用`prop.table()`函数生成比例（小数）。  
`prop.table(mytable) # cell proportions`  
`prop.table(mytable, 1) # row proportions`  
`prop.table(mytable, 2) # column proportions`

## 二维列联表 (3/3)

- 可以使用`addmargins()`函数表格添加边际和。  
`addmargins(mytable) # add row and column sums to table`
- 更加复杂的例子
  - `addmargins(prop.table(mytable))`
  - `addmargins(prop.table(mytable, 1), 2)`
  - `addmargins(prop.table(mytable, 2), 1)`

# 用CrossTable()创建二维列联表 (1/2)

- gmodels包中的CrossTable()函数仿照
  - **SAS**中PROC FREQ或
  - **SPSS**中CROSSTABS的形式  
生成二维列联表。
- 例：

```
library(gmodels)  
CrossTable(Arthritis$Treatment,  
Arthritis$Improved)
```

# 用CrossTable()创建二维列联表 (2/2)

cell contents	
	N
Chi-square contribution	
N / Row Total	
N / Col Total	
N / Table Total	

Total Observations in Table: 84

Arthritis\$Treatment	Arthritis\$Improved			Row Total
	None	Some	Marked	
Placebo	29	7	7	43
	2.616	0.004	3.752	
	0.674	0.163	0.163	0.512
	0.690	0.500	0.250	
	0.345	0.083	0.083	
Treated	13	7	21	41
	2.744	0.004	3.935	
	0.317	0.171	0.512	0.488
	0.310	0.500	0.750	
	0.155	0.083	0.250	
column Total	42	14	28	84
	0.500	0.167	0.333	

# 多维列联表（1/2）

- **table()**和**xtabs()**可以基于三个或更多的类别型变量生成多维列联表。
- **margin.table()**、**prop.table()**和**addmargins()**函数可以推广到高于二维的情况。
- **ftable()**函数可以以一种紧凑而吸引人的方式输出多维列联表。



# 多维列联表 (2/2)

#listing 7.11 - Three way table

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

mytable

**ftable**(mytable)

Treatment	Sex	Improved	None	Some	Marked
Placebo	Female		19	7	6
	Male		10	0	1
Treated	Female		6	5	16
	Male		7	2	5

margin.table(mytable, 1)

margin.table(mytable, 2)

margin.table(mytable, 2)

margin.table(mytable, c(1,3))

ftable(prop.table(mytable, c(1,2)))

ftable(addmargins(prop.table(mytable, c(1, 2)), 3))

, , Improved = None

Treatment	Sex	Female	Male
Placebo		19	10
Treated		6	7

, , Improved = Some

Treatment	Sex	Female	Male
Placebo		7	0
Treated		5	2

, , Improved = Marked

Treatment	Sex	Female	Male
Placebo		6	1
Treated		16	5

# 列联表的用处

- 列联表可以告诉我们组成表格的各种变量组合的频数或比例
- 独立性检验：列联表中的变量是否相关或独立？（下一节内容）

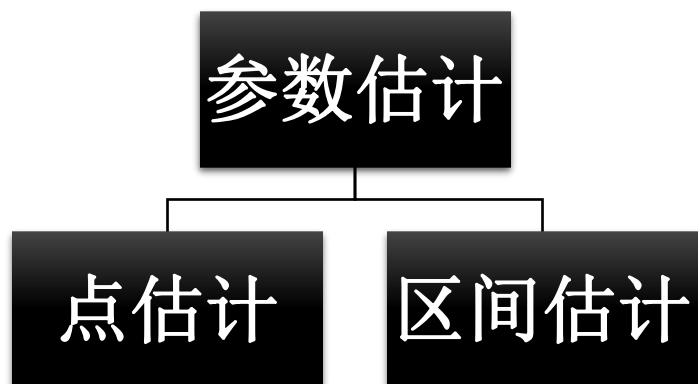
# 参数估计（理论复习）

……这么说吧，小伙子！  
她们都一个样儿，相一个  
就等于相全部！



# 参数估计

- 参数估计分类



# 点估计

- 基本思路
  - 从总体中抽取一个样本，根据该样本的统计量对总体的未知参数做出一个数值点的估计
- 例如
  - 用样本均值作为总体未知均值 $\mu$ 的估计值  
例：“X公司工资多少啊？”“我几个师兄平均**20K**吧。”
- 注意
  - 点估计没有给出估计值接近总体未知参数程度的信息

# 区间估计：自信地猜测

- 例

甲：X公司工资多少啊？

乙：20K。因为我师兄们平均都这些（点估计）。

甲：你师兄都厉害。我去能给多钱啊？

乙：如果你能被录用，我能99%确定，你的工资在5k-35k之间。（区间估计）

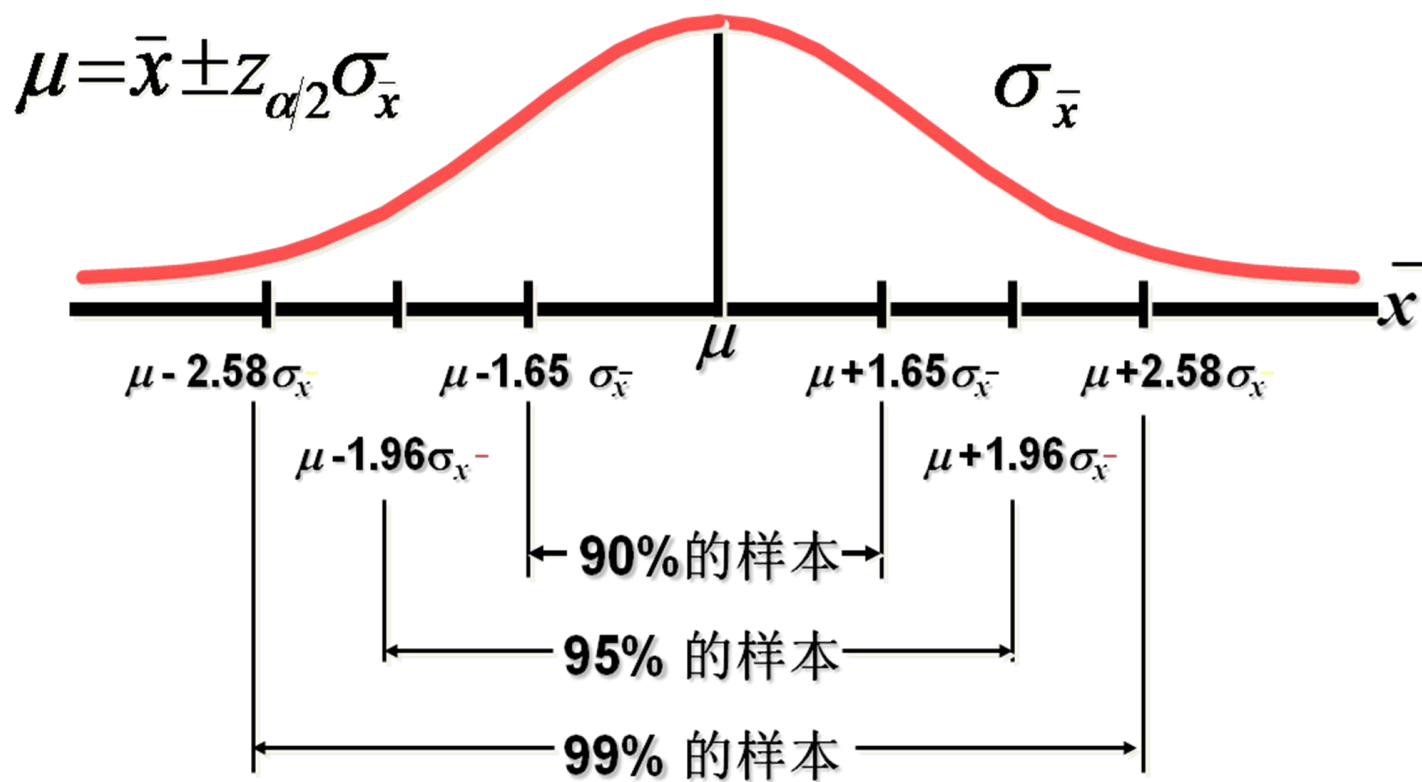
甲：@%&.....##&@

# 区间估计 (1)

- 区间估计
  - 在点估计的基础上，给出总体参数落在某一区间的概率。
  - 区间就是根据一个样本的观察值给出的总体参数的估计范围，可通过样本统计量加减抽样误差的方法计算。
  - 例如，总体均值落在**50~70**之间的概率为 **0.95**。



# 区间估计 (2)



总体的均值区间估计示意图

# 区间估计（3）

## 区间估计的指标

### – 置信区间（Confidence interval）

- 是指由样本统计量所构造的总体参数的估计区间，其中区间的最小值和最大值分别称为置信下限和置信上限。

### – 置信水平（Confidence level）

- **置信水平**是指总体未知参数落在区间内的概率，表示为  **$1 - \alpha$** ， **$\alpha$ 为显著性水平**，即总体参数**未**在区间内的概率。
- 在构造置信区间时，可以**任意**设置目标置信水平。常用的置信水平及正态分布曲线下**右侧面积为 $\alpha/2$ 时的 **$z$ 值****（ $z_{\alpha/2}$ ）。

# 区间估计 (4)

常用置信水平的 $z_{\alpha/2}$ 值

置信水平	$\alpha$	$\alpha/2$	$Z_{\alpha/2}$
90	0.1	0.05	1.645
95	0.05	0.025	1.96
99	0.01	0.005	2.58

# 区间估计 (5)

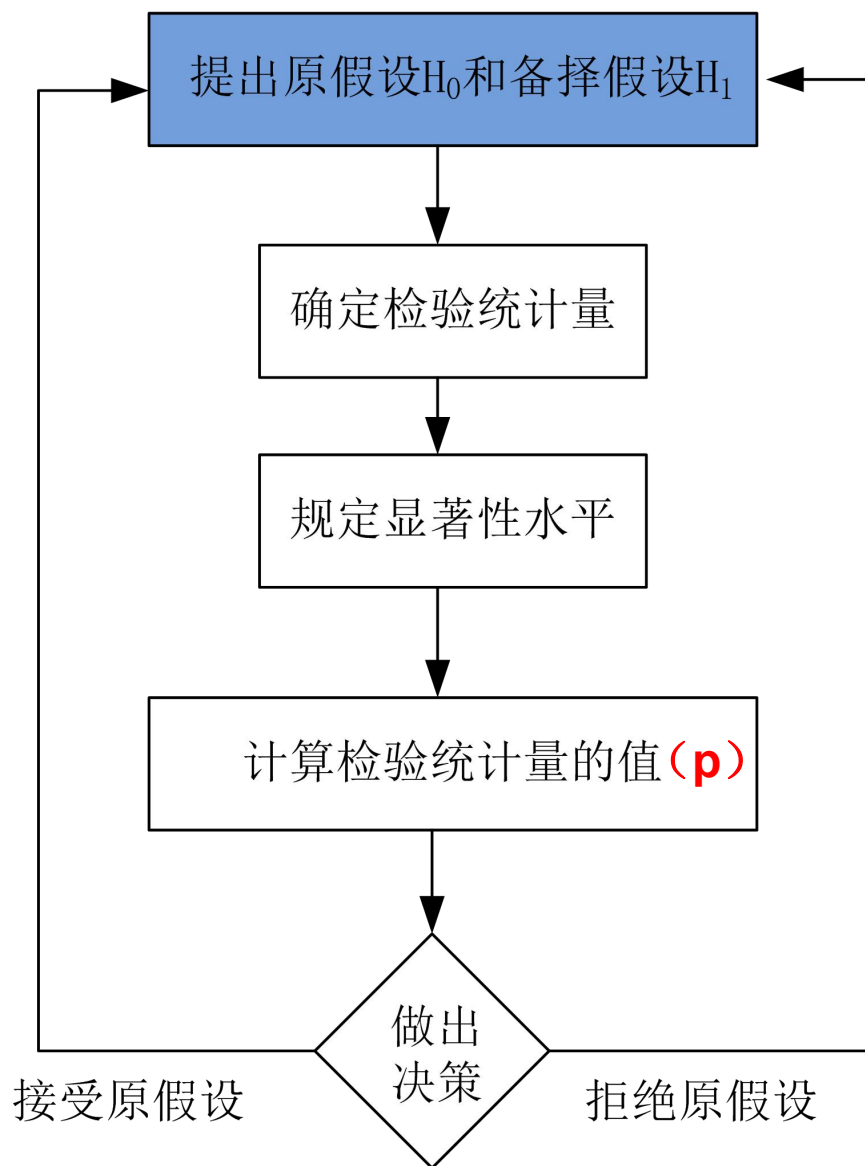
- 参数估计中，用于估计总体某一参数的随机变量称为**估计量**（estimator），如**样本均值就是总体均值 $\mu$ 的一个估计量**。
- 判断估计量的优良性准则：
  - 无偏性（Unbiasedness）
    - 估计量的数学期望等于被估计的总体参数；
  - 有效性(Efficiency)
    - 一个方差较小的无偏估计量称为一个更有效的估计量。如，与其他估计量相比，样本均值是一个更有效的估计量；
  - 一致性(Consistency)
    - 随着样本容量的增大，估计量越来越接近被估计的总体参数。

# 假设检验（理论复习）

# 假设检验

- 在数据科学中，经常采用“假设/演绎式分析方法”——先对总体参数或分布形式作出某种假设，然后利用数据（或样本信息）证明（或判断）原假设是否成立。
- 假设检验主要以小概率原理为基础，采用的是逻辑反证法。
  - 根据费希尔德（Ronald Fisher）的观点：小概率的标准是小于/等于**0.05**。小概率是指在一次试验中，一个几乎不可能发生的事件发生的概率。即，**如果在一次试验中出现了小概率事件，就有理由拒绝原假设。**

# 假设检验的基本步骤



# 假设检验6步骤

即我们要对其进行试验的断言。

→ ① 确定要进行检验的假设

② 选择检验统计量

← 我们需要选取能最有效地对断言进行检验的统计量。

我们需要使用某种确定性水平。

→ ③ 确定用于做决策的拒绝域

④ 求出检验统计量的p值

← 我们需要了解在假定断言为真的情况下，我们的试验结果的可信程度。

⑤ 查看样本结果是否位于拒绝域内

⑥ 作出决策

← 接着需要了解试验结果是否位于确定性限值范围中。



# 假设检验两种错误（1/2）

- 在进行假设检验时提出原假设和备择假设，原假设实际上是正确的，但我们做出的决定是拒绝原假设，此类错误称为第一类错误（ **$\alpha$ 错误**）。
- 原假设实际上是不正确的，但是我们却做出了接受原假设的决定，此类错误称为第二类错误（ **$\beta$ 错误**）。

# 假设检验两种错误（2/2）

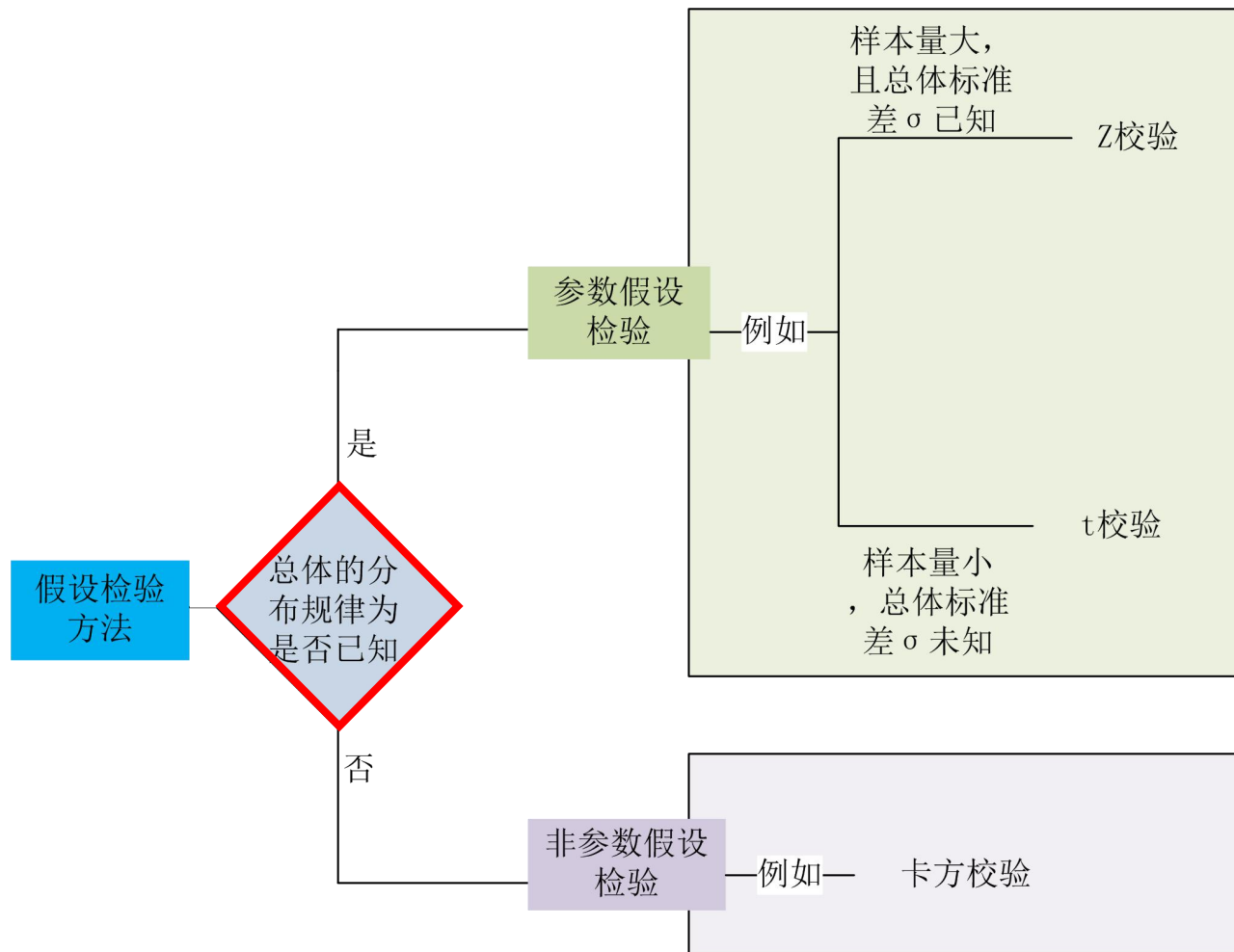
## 假设检验中各种可能的概率

项目	没有拒绝原假设 $H_0$	拒绝原假设 $H_0$
$H_0$ 为真	$1-\alpha$ （正确决策）	$\alpha$ （弃真错误）
$H_0$ 为假	$\beta$ （取伪错误）	$1-\beta$ （正确决策）

- 注意：

- 如果减少 $\alpha$ 错误，就会增加犯 $\beta$ 错误的可能
- 如果减少 $\beta$ 错误，就会增加犯 $\alpha$ 错误的可能

# 假设检验的方法



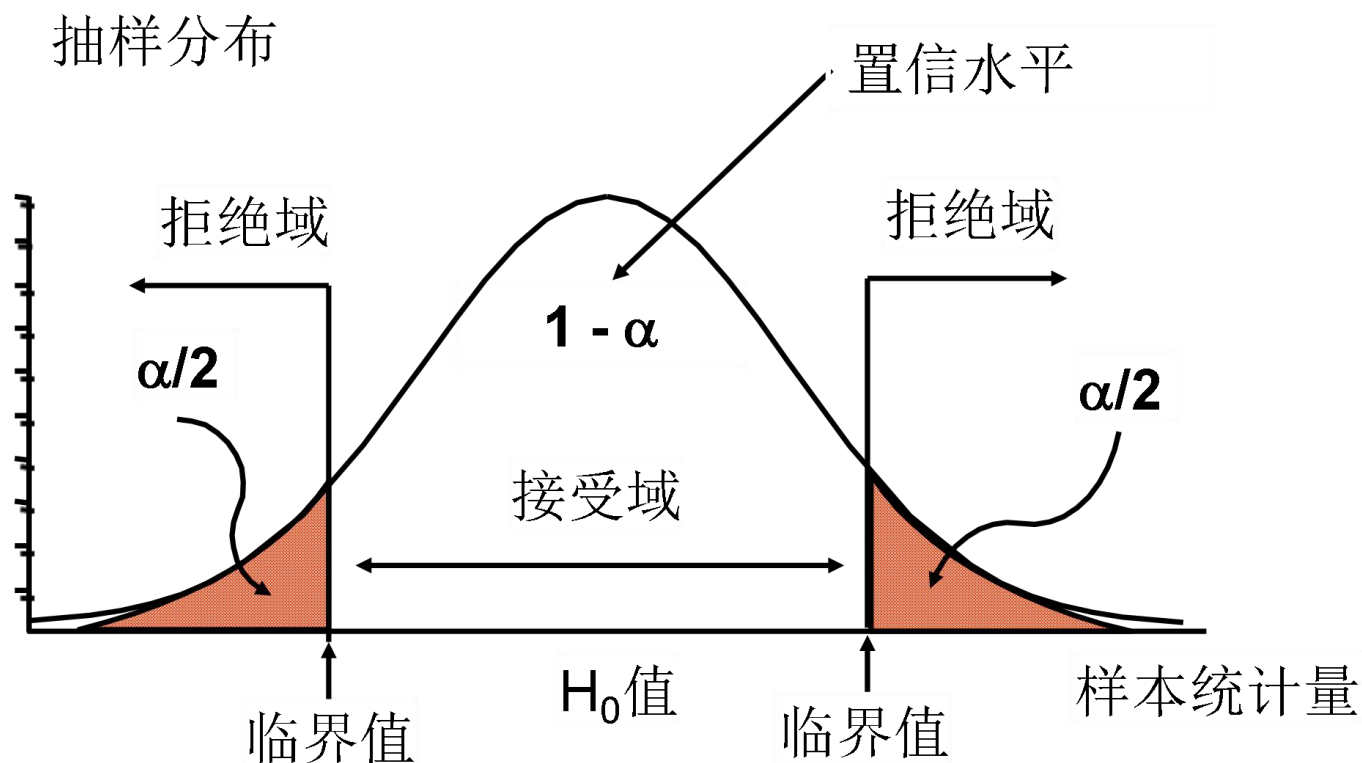
注：非参数检验是在总体方差未知或知道甚少的情況下，利用样本数据对总体分布形态等进行推断的方法。由于非参数检验方法在推断过程中不涉及有关总体分布的参数，因而得名为“非参数”检验。

# 参数检验（1/2）

左单侧检验、右单侧检验和双侧检验

检验	研究的问题		
	双侧检验	左单侧检验	右单侧检验
$H_0$	$=$	$\geq$	$\leq$
$H_1$	$\neq$	$<$	$>$

# 参数检验 (2/2)



双侧检验

# 独立性检验

# 独立性检验

- R提供了多种检验类别型变量独立性的方法。
- 三种检验：
  - 卡方独立性检验
  - Fisher精确检验
  - Cochran-Mantel-Haenszel检验

# 卡方 ( $\chi^2$ ) 假设检验步骤 (1/4)

- 1 确定要进行检验的假设及其备择假设
- 2 求出期望频数和自由度
- 3 确定用于做决策的拒绝域
- 4 计算检验统计量 $\chi^2$
- 5 查看检验统计量是否位于拒绝域以内
- 6 作出决策



# 卡方 ( $X^2$ ) 独立性检验 (1/3)

## Chi-square test of independence

- 卡方独立性检验的**原假设** $H_0$ 是假设**各属性之间相互独立**。
- **$X^2$ 统计量**可以用来作相关性的度量。 **$X^2$ 越小说明变量之间越独立**， **$X^2$ 越大说明变量之间越相关**。
- 例：患者接受的治疗方式 (Treatment) 和改善水平 (improved) 是否独立？
  - 原假设 $H_0$ ：Treatment和improved独立。
  - 备择假设 $H_1$ ：Treatment和improved不独立，即存在某种关系。

# 卡方 ( $X^2$ ) 独立性检验 (2/3)

- 使用 `chisq.test()` 函数对二维列联表的行变量和列变量进行卡方独立性检验。代码：

```
library(vcd)
mytable <- xtabs(~Treatment+Improved,
data=Arthritis)
chisq.test(mytable)
```

生成列联表

Pearson's Chi-squared test

```
data: mytable
X-squared = 13.055, df = 2, p-value = 0.001463
```

$X^2=13.055$

自由度=2

p远小于0.05，否定 $H_0$ ，即Treatment和Improved不独立。

自由度 (**df**) 指的是计算某一统计量时，取值不受限制的变量个数。通常 $df=n-k$ 。其中n为样本数量，k为被限制的条件数或变量个数或计算某一统计量时用到其它独立统计量的个数。

# 卡方 ( $X^2$ ) 独立性检验 (3/3)

- 另一个例子：性别 (**Sex**) 和和改善水平 (**improved**) 是否独立？

- 代码：

```
mytable <- xtabs(~Improved+Sex, data=Arthritis)
chisq.test(mytable)
```

Pearson's Chi-squared test

data: mytable

X-squared = 4.8407, df = 2, p-value = 0.08889

Warning message:

In chisq.test(mytable) : Chi-squared近似算法有可能不准

← p大于0.05，故没有足够理由否定 $H_0$ ，即没有足够理由说明Treatment和Sex是不独立的。

↑ 卡方检验有四条假设（参考Wikipedia: Pearson's chi-squared test），其中一条假设是每个单元频数都要大于5，而mytable中有一个小于5的值。

# Fisher精确检验 (1/2)

## (Fisher's exact test)

- Fisher精确检验的原假设是：边界固定的列联表中行和列是相互独立的。
- 可以使用`fisher.test()`函数进行Fisher精确检验。其调用格式为`fisher.test(mytable)`，其中的`mytable`是一个二维列联表。
- 注意：`fisher.test()`函数可以在任意行列数大于等于2的二维列联表上使用，但不能用于 $2 \times 2$ 的列联表。

# Fisher精确检验 (2/2)

- 例:

```
mytable <- xtabs(~Treatment+Improved,  
data=Arthritis)
```

```
fisher.test(mytable)
```

Fisher's Exact Test for Count Data

```
data: mytable  
p-value = 0.001393  
alternative hypothesis: two.sided
```

p远小于0.05，否定H0，即Treatment和Improved不独立。

# Cochran-Mantel-Haenszel检验

## (1/2)

- `mantelhaen.test()`函数可用来进行Cochran-Mantel-Haenszel卡方检验。
- 其原假设是：两个名义变量在第三个变量的每一层中都是条件独立的。
- 例：分性别来看，患者接受的治疗方式（**Treatment**）和改善水平（**improved**）是否独立？

# Cochran-Mantel-Haenszel检验

## (2/2)

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

```
mantelhaen.test(mytable)
```

Cochran-Mantel-Haenszel test

data: mytable

Cochran-Mantel-Haenszel  $M^2 = 14.632$ ,  $df = 2$ ,  $p\text{-value} = 0.0006647$

$p$ 远小于0.05, 否定 $H_0$

- 结果表明, 患者接受的治疗与得到的改善在性别的每一水平下并不独立 (即, 分性别来看, 用药治疗的患者较接受安慰剂的患者有了更多的改善)。

# 相关性的度量（1/2）

## Measures of association for a two-way table

- 上一节中的显著性检验评估了是否存在充分的证据以拒绝变量间相互独立的原假设。
- 如果可以拒绝原假设，那么我们的兴趣就会自然而然地转向用以衡量相关性强弱的相关性度量。
- **vcd**包中的**assocstats()**函数可以用来计算二维列联表的
  - phi系数
  - 列联系数
  - Cramer's V系数



# 相关性的度量 (2/2)

```
library(vcd)
mytable <- xtabs(~Treatment+Improved,
data=Arthritis)
assocstats(mytable)
```

	X <sup>2</sup>	df	P(> X <sup>2</sup> )
Likelihood Ratio	13.530	2	0.0011536
Pearson	13.055	2	0.0014626

Phi-Coefficient	: NA
Contingency Coeff.	: 0.367
Cramer's V	: 0.394

- 总体来说，较大的值意味着较强的相关性。

# 相关性检验

# 相关

- 相关系数可以用来描述定量变量之间的关系。相关系数的符号（ $\pm$ ）表明关系的方向（正相关或负相关），其值的大小表示关系的强弱程度（完全不相关时为0，完全相关时为1）。

# 相关的类型

- R可以计算多种相关系数，包括
  - Pearson相关系数
  - Spearman相关系数
  - Kendall相关系数
  - 偏相关系数
  - 多分格（polychoric）相关系数
  - 多系列（polyserial）相关系数

# Pearson、Spearman和Kendall 相关

- **Pearson**积差相关系数衡量了两个定量变量之间的线性相关程度。
- **Spearman**等级相关系数则衡量分级定序变量之间的相关程度。
- **Kendall's Tau**相关系数是一种非参数的等级相关度量。

# 协方差 (Covariance)

- 在概率论和统计学中，协方差用于衡量两个变量的总体误差。方差是协方差的一种特殊情况，即当两个变量是相同的情况。

$$\text{cov}(X, Y) = \mathbf{E} [(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$$

$$\text{即 } \text{cov}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- 通俗解释：你变大，同时我也变大，说明两个变量是同向变化的，这时协方差就是正的。你变大，同时我变小，说明两个变量是反向变化的，这时协方差就是负的。从数值来看，协方差的数值越大，两个变量同向程度也就越大。反之亦然。

# R的计算相关系数的函数

- **cor()**函数可以计算Pearson、Spearman和Kendall相关系数。
- **cov()**函数可以用来计算协方差。

# 例1：协方差

`states<- state.x77[,1:6]` #准备数据集：收入、文盲率、寿命、高中毕业率等

**cov(states)** #计算协方差

			文盲率			高中毕业率
	Population	Income	Illiteracy	Life Exp	Murder	HS Grad
Population	19931683.7588	571229.7796	292.8679592	-407.8424612	5663.523714	-3551.509551
Income	571229.7796	377573.3061	-163.7020408	280.6631837	-521.894286	3076.768980
Illiteracy	292.8680	-163.7020	0.3715306	-0.4815122	1.581776	-3.235469
Life Exp	-407.8425	280.6632	-0.4815122	1.8020204	-3.869480	6.312685
Murder	5663.5237	-521.8943	1.5817755	-3.8694804	13.627465	-14.549616
HS Grad	-3551.5096	3076.7690	-3.2354694	6.3126849	-14.549616	65.237894

e.g., 收入和高中毕业率之间存在很强的正相关。



## 例2: Pearson积差相关系数

**cor**(states)

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad
Population	1.00000000	0.2082276	0.1076224	-0.06805195	0.3436428	-0.09848975
Income	0.20822756	1.00000000	-0.4370752	0.34025534	-0.2300776	0.61993232
Illiteracy	0.10762237	-0.4370752	1.00000000	-0.58847793	0.7029752	-0.65718861
Life Exp	-0.06805195	0.3402553	-0.5884779	1.00000000	-0.7808458	0.58221620
Murder	0.34364275	-0.2300776	0.7029752	-0.78084575	1.00000000	-0.48797102
HS Grad	-0.09848975	0.6199323	-0.6571886	0.58221620	-0.4879710	1.00000000

e.g.,

收入和高中毕业率之间存在很强的正相关。

文盲率和预期寿命之间存在很强的负相关。

# 例3: Spearman相关系数

**cor**(states, method="spearman")

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad
Population	1.0000000	0.1246098	0.3130496	-0.1040171	0.3457401	-0.3833649
Income	0.1246098	1.0000000	-0.3145948	0.3241050	-0.2174623	0.5104809
Illiteracy	0.3130496	-0.3145948	1.0000000	-0.5553735	0.6723592	-0.6545396
Life Exp	-0.1040171	0.3241050	-0.5553735	1.0000000	-0.7802406	0.5239410
Murder	0.3457401	-0.2174623	0.6723592	-0.7802406	1.0000000	-0.4367330
HS Grad	-0.3833649	0.5104809	-0.6545396	0.5239410	-0.4367330	1.0000000

e.g.,

收入和高中毕业率之间存在很强的正相关。

文盲率和预期寿命之间存在很强的负相关。

# 非方形的相关矩阵

- 在默认情况下得到的结果是一个方阵（所有变量之间两两计算相关）。
- `cor()`可以计算非方形的相关矩阵。例：  
x <- states[,c("Population", "Income", "Illiteracy", "HS Grad")]  
y <- states[,c("Life Exp", "Murder")]  
`cor(x,y)`

	Life Exp	Murder
Population	-0.06805195	0.3436428
Income	0.34025534	-0.2300776
Illiteracy	-0.58847793	0.7029752
HS Grad	0.58221620	-0.4879710

# 偏相关（1/2）

- 偏相关是指在控制一个或多个定量变量（要排除影响的定量变量）时，另外两个定量变量之间的相互关系。
- 可以使用ggm包中的pcor()函数计算偏相关系数。

pcor()函数调用格式为：**pcor**(**u**, S)

其中的**u**是一个数值向量，前两个数值表示要计算相关系数的变量下标，其余的数值为条件变量（即要排除影响的变量）的下标。**S**为变量的协方差阵。

## 偏相关（2/2）

- 例：求在控制了收入、文盲率和高中毕业率的影响时，人口和谋杀率之间的相关系数：

```
library(ggm)
```

```
pcor(c(1,5,2,3,6), cov(states))
```



```
[1] 0.3462724
```

- 偏相关系数常用于社会科学的研究中。

# 相关性的显著性检验

# 相关性的显著性检验

- 在计算好相关系数以后，可以对它们进行统计显著性检验。
- 常用的**原假设**为：**变量间不相关**（即总体的相关系数为0）。
- 可以使用`cor.test()`函数对单个的**Pearson**、**Spearman**和**Kendall**相关系数进行检验。

# cor.test() (1/2)

- cor.test()函数的简化的使用格式为：  
**cor.test**(x, y, **alternative** = , **method** = )
  - 其中的x和y为要检验相关性的变量，
  - **alternative**用来指定进行双侧检验或单侧检验（取值为"two.side"、"less"或"greater"），
  - **method**用以指定要计算的相关类型（"pearson"、"kendall"或"spearman"）。
  - 当研究的假设为总体的相关系数小于0时，请使用 **alternative="less"**。在研究的假设为总体的相关系数大于0时，应使用 **alternative="greater"**。在默认情况下，假设为 **alternative="two.side"**（总体相关系数不等于0）。



# cor.test() (2/2)

- 例：以下代码检验了文盲率和谋杀率的Pearson相关系数为0的原假设。

文盲率                  谋杀率  
**cor.test**(states[,3], states[,5])

Pearson's product-moment correlation

```
data: states[, 3] and states[, 5]
t = 6.8479, df = 48, p-value = 1.258e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5279280 0.8207295
sample estimates:
      cor
0.7029752
```

远小于0.05。拒绝原假设。即文盲率和谋杀率之间的总体相关度不为零。

# corr.test() ( 1/2 )

- `cor.test`每次只能检验一种相关关系。而`psych`包中提供的`corr.test()`函数可以一次做更多事情。`corr.test()`函数可以为Pearson、Spearman或Kendall相关计算相关矩阵和显著性水平。
- **`corr.test(x, y, use= , method = )`**
  - 参数`use=`的取值可为"pairwise"或"complete"  
(分别表示对缺失值执行成对删除或行删除)。
  - 参数`method=`的取值可为"pearson" (默认值)、"spearman"或"kendall"。

# corr.test() (2/2)

```
library(psych)
```

```
corr.test(states, use="complete")
```

相关系数  
矩阵

```
Call: corr.test(x = states, use = "complete")
```

```
Correlation matrix
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad
Population	1.00	0.21	0.11	-0.07	0.34	-0.10
Income	0.21	1.00	-0.44	0.34	-0.23	0.62
Illiteracy	0.11	-0.44	1.00	-0.59	0.70	-0.66
Life Exp	-0.07	0.34	-0.59	1.00	-0.78	0.58
Murder	0.34	-0.23	0.70	-0.78	1.00	-0.49
HS Grad	-0.10	0.62	-0.66	0.58	-0.49	1.00

```
Sample Size
```

```
[1] 50
```

p值矩阵上三角的数值使用多重检验进行了调整

```
Probability values (Entries above the diagonal are adjusted for multiple tests.)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad
Population	0.00	0.59	1.00	1.0	0.10	1
Income	0.15	0.00	0.01	0.1	0.54	0
Illiteracy	0.46	0.00	0.00	0.0	0.00	0
Life Exp	0.64	0.02	0.00	0.0	0.00	0
Murder	0.01	0.11	0.00	0.0	0.00	0
HS Grad	0.50	0.00	0.00	0.0	0.00	0

下三角是  
原始  
p值

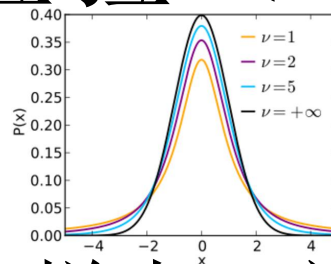
```
To see confidence intervals of the correlations, print with the short=FALSE option
```

t检验

# t检验

- **William Sealy Gosset**（牛津大学毕业），1908年，在吉尼斯啤酒厂（使用“**student**”笔名）提出，以便降低啤酒质量监控成本。
- 在研究中最常见的行为就是对两个组进行比较。
  - 接受某种新药治疗的患者是否较使用某种现有药物的患者表现出了更大程度的改善？
  - 某种制造工艺是否较另外一种工艺制造出的不合格品更少？
  - 两种教学方法中哪一种更有效？

# 独立样本的t检验（1/2）



- 原始测量值服从t分布。
- 原假设：针对两组独立样本（并且是从正态总体中抽得），两个总体的均值相等。
- t检验的第一种调用格式为：  
`t.test(y ~ x, data)`  
其中的y是一个数值型变量，x是一个二分变量。
- t检验的第二种调用格式为：  
`t.test(y1, y2)`  
其中的y1和y2为数值型向量。

# 独立样本的t检验 (2/2)

- 例：如果你在美国的南方犯罪，是否更有可能被判监禁？
- 原假设：南方和非南方拥有相同监禁概率。
- `library(MASS)` 监禁概率 “南方”标记（二分变量）
- `t.test(Prob ~ So, data=UScrime)`

Welch Two Sample t-test

```
data: Prob by So
t = -3.8954, df = 24.925, p-value = 0.0006506
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03852569 -0.01187439
sample estimates:
mean in group 0 mean in group 1
 0.03851265      0.06371269
```

拒绝原假设

# 非独立样本的t检验

- 例：14~24岁男性的失业率和35~39岁男性的失业率是否相同？（对于失业率来说，两个人群失业率是相关的，因此是非独立样本。）
- 原假设：两组失业率相同。

with(UScrime, **t.test**(U1, U2, **paired=TRUE**))

Paired t-test

```
data: U1 and U2
t = 32.407, df = 46, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 57.67003 65.30870
sample estimates:
mean of the differences
 61.48936
```

拒绝原假设。此外，通过计算两组的均值（分别为95.5和33.98）可知年轻男性失业率更高。



# 多于两组的情况

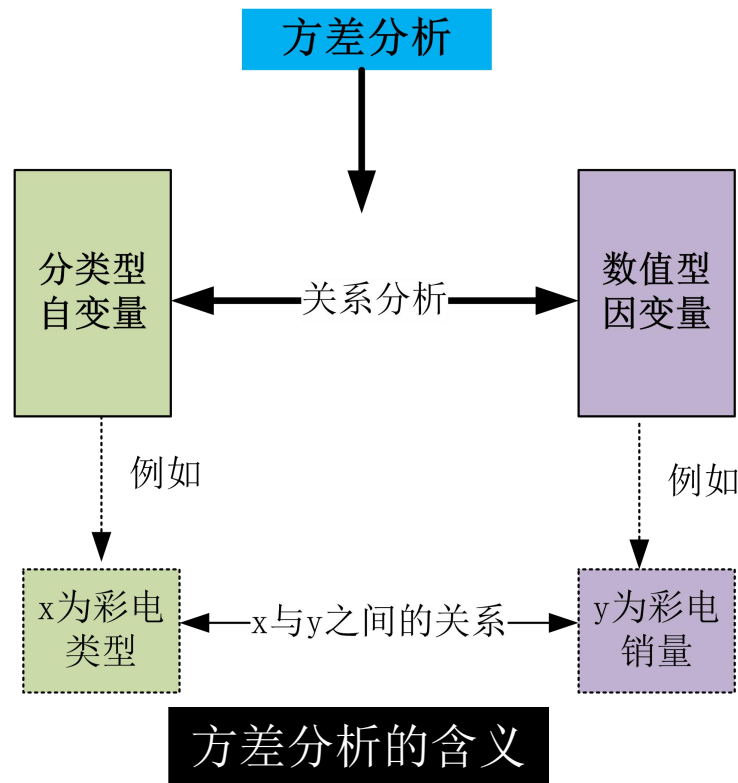
- 如果能够假设数据是从正态总体中独立抽样而得的，那么可以使用方差分析。
- 如果数据无法满足t检验或ANOVA的参数假设，可以转而使用非参数方法。

# 方差分析

(Analysis of variance,  
ANOVA)

# 方差分析概述（1/4）

方差分析主要用于分析**分类型自变量**和**数值型因变量**之间的关系，例如分析彩电的品牌对其销售量的影响；犯罪率是否与地区有关？



# 方差分析概述（2/4）

## 方差分析的基本假定

- 每个总体都应服从正态分布；
- 各个总体的方差必须相同；
- 观察值 是独立的

# 方差分析概述（3/4）

## 方差分析

- 是指通过检验各总体的均值是否相等来判断分类型自变量对数值型因变量是否有显著影响的方法。
- 其基本思想是采用方差比对比随机误差与系统误差的方法检验均值是否相等。
  - 如果系统误差显著地不同于随机误差，则均值就是不相等的；
  - 如果系统误差并不显著地不同于随机误差，均值就是相等的。

# 方差分析概述（4/4）

## 方差分析的类型

- 单因素方差分析
  - 只涉及一个分类型自变量，如分析品牌对彩电销售量的影响。
- 双因素方差分析
  - 涉及两个分类型自变量，如分析彩电的品牌和销售地区对销售量的影响。

# 例（服务质量差异）（1/2）

- 消费者协会在零售业、旅游业、航空公司、家电制造业分别抽取了7家、6家、5家、5家企业，分别统计它们的投诉次数。消费者协会想知道这几个行业之间的服务质量是否有显著差异。

零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

## 例（服务质量差异）（2/2）

- 要分析四个行业间的服务质量是否有显著差异，实际上是要判断“行业”对“投诉次数”是否有显著影响。作出这种判断最终被归结为**检验**这四个行业被投诉次数的**均值是否相等**。



# 方差分析及其术语（1/2）

- 表面上看，方差分析是检验多个总体均值是否相等的统计方法；但**本质上它研究的是分类型自变量对数值型因变量的影响**。例如，变量之间有没有关系，关系强度如何。

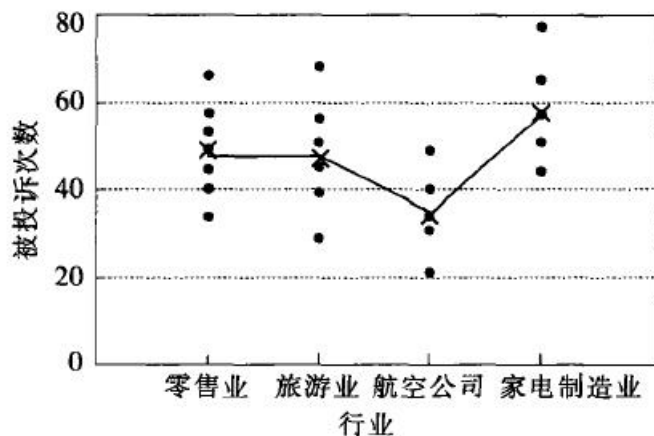
# 方差分析及其术语（2/2）

- 在方差分析中，所要检验的对象称为**因子**（**factor**），因子的不同表现称为**水平**或处理。每个因子水平下得到的样本数据称为**观测值**。

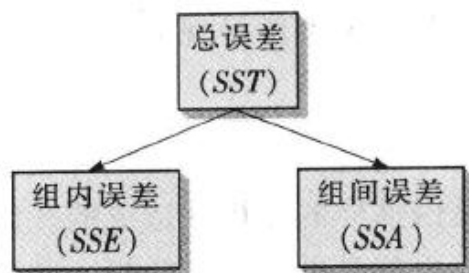
零售业	旅游业	航空公司	家电制造业
57	68	31	44
66	39	49	51
49	29	21	65
40	45	34	77
34	56	40	58
53	51		
44			

- 因为只涉及到“行业”一个因素，因此称为单因素4水平的实验。

# 方差分析基本原理 (1/3)



不同行业被投诉次数的散点图



误差分解图

- 图形描述
  - 折线是由均值连接而成的。可以看出有不同行业与被投诉次数一定差异。但这种差异也可能是由抽样的随机性造成的。
- 误差分解
  - **SST**: 反映全部数据误差大小的平方和称为总平方和。例如全部**23**家企业被投诉数之间的的误差平方和就是总平方和，它反映了全部观测值的离散状况。
  - **SSE**: 反映组内误差大小的平方和，也称误差平方和。每个样本内部数据平方和加在一起。反映了样本内各观测值的总离散状况。
  - **SSA**: 反映组间误差大小的平方和。例如四个行业被投诉次数之间的误差平方和就是**SSA**。

# 方差分析基本原理（2/3）

- 误差分析

- 如果不同行业对投诉次数没影响，则组间误差中只包含随机误差，而没有系统误差。这时组间误差和组内误差经平均后的数值就应该很接近（比值会接近1）；
- 反之，若不同行业对投诉次数有影响，则组间误差中除了随机误差还会包含系统误差，这时组间误差会大于组内误差平均后的数值（比值会大于1）。
- 当比值大到某种程度时，就认为因素的不同水平之间存在显著差异。

# 方差分析基本原理（3/3）

- 方差分析主要通过**F检验**来进行效果评测，如果**F检验**显著，则说明不同水平之间存在显著差异。
- **F检验**又叫方差齐性检验。从两研究总体中随机抽取样本，要对这两个样本进行比较的时候，首先要判断两总体方差是否相同，即方差齐性。即要判断两总体方差是否相同，就可以用**F检验**。
- **F检验**原假设：两总体方差相同。

# ANOVA 模型拟合: aov()函数

- aov()函数的语法为  
**aov(formula, data = dataframe)**
- 常见研究设计的表达式

设 计	表 达 式
单因素ANOVA	$y \sim A$
含单个协变量的单因素ANCOVA	$y \sim x + A$
双因素ANOVA	$y \sim A * B$
含两个协变量的双因素ANCOVA	$y \sim x1 + x2 + A*B$
随机化区组	$y \sim B + A$ (B是区组因子)
单因素组内ANOVA	$y \sim A + \text{Error}(\text{Subject}/A)$
含单个组内因子(W)和单个组间因子(B)的重复测量ANOVA	$y \sim B * W + \text{Error}(\text{Subject}/W)$

y是因变量, 字母A、B、C代表因子

# 单因素方差分析（1/3）

- 单因素方差分析中，比较分类因子定义的两个或多个组别中的因变量均值。

- 例：multcomp包中的cholesterol数据集：

	trt	response
1	1time	3.8612
2	1time	10.3868
3	1time	5.9059
4	1time	3.0609
5	1time	7.7204
6	1time	7.7130

- 50个患者均接受降低胆固醇药物治疗（trt）五种疗法中的一种疗法。
- 其中三种治疗条件使用药物相同，分别是20 mg一天一次（1time）、10 mg一天两次（2times）和5 mg一天四次（4times）。
- 剩下的两种方式（drugD和drugE）代表候选药物。
- 哪种药物疗法降低胆固醇（response）最多呢？

# 单因素方差分析 (2/3)

```
library(multcomp)
```

```
attach(cholesterol)
```

```
table(trt) #计算各组频数 (即各组样本大小)
```

```
aggregate(response, by=list(trt), FUN=mean) #各组均值
```

```
aggregate(response, by=list(trt), FUN=sd) #各组方差
```

```
fit <- aov(response ~ trt)
```

```
summary(fit) #summary函数得到方差分析表
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
trt	4	1351.4	337.8	32.43	9.82e-13 ***
Residuals	45	468.8	10.4		

ANOVA对治疗方式 (trt) 的  
F检验非常显著 ( $p < 0.0001$ )。  
说明五种疗法的效果不同。

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

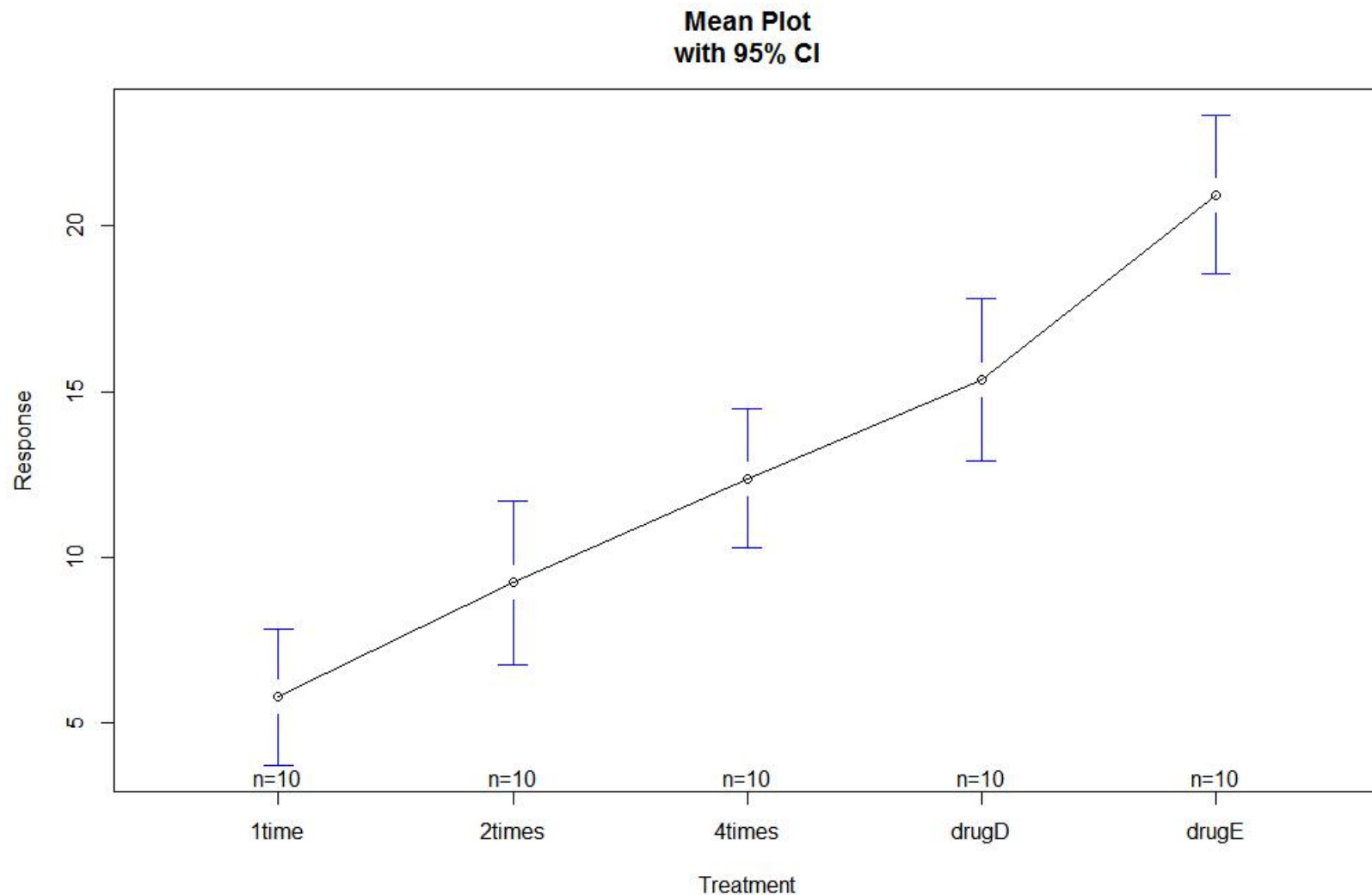
```
library(gplots) #下面代码用于绘制各组均值 (图在下页)
```

```
plotmeans(response ~ trt, xlab="Treatment", ylab="Response", main="Mean  
Plot\nwith 95% CI") #plotmeans()用来绘制带有置信区间的组均值图形
```

```
detach(cholesterol)
```



# 单因素方差分析（3/3）



五种降低胆固醇药物疗法的均值，含95%的置信区间

# 多重比较（1/3）

- 虽然ANOVA对各疗法的F检验表明五种药物疗法效果不同，但是并没有告诉我们哪种疗法与其他疗法不同。**多重比较**可以解决这个问题。
- TukeyHSD()函数提供了对各组均值差异的成对检验。

# 多重比较 (2/3)

## TukeyHSD(fit)

Tukey multiple comparisons of means  
95% family-wise confidence level

Fit: aov(formula = response ~ trt)

\$trt		diff	lwr	upr	p_adj
2times-1time	3.44300	-0.6582817	7.544282	0.1380949	
4times-1time	6.59281	2.4915283	10.694092	0.0003542	
drugD-1time	9.57920	5.4779183	13.680482	0.0000003	
drugE-1time	15.16555	11.0642683	19.266832	0.0000000	
4times-2times	3.14981	-0.9514717	7.251092	0.2050382	
drugD-2times	6.13620	2.0349183	10.237482	0.0009611	
drugE-2times	11.72255	7.6212683	15.823832	0.0000000	
drugD-4times	2.98639	-1.1148917	7.087672	0.2512446	
drugE-4times	8.57274	4.4714583	12.674022	0.0000037	
drugE-drugD	5.58635	1.4850683	9.687632	0.0030633	

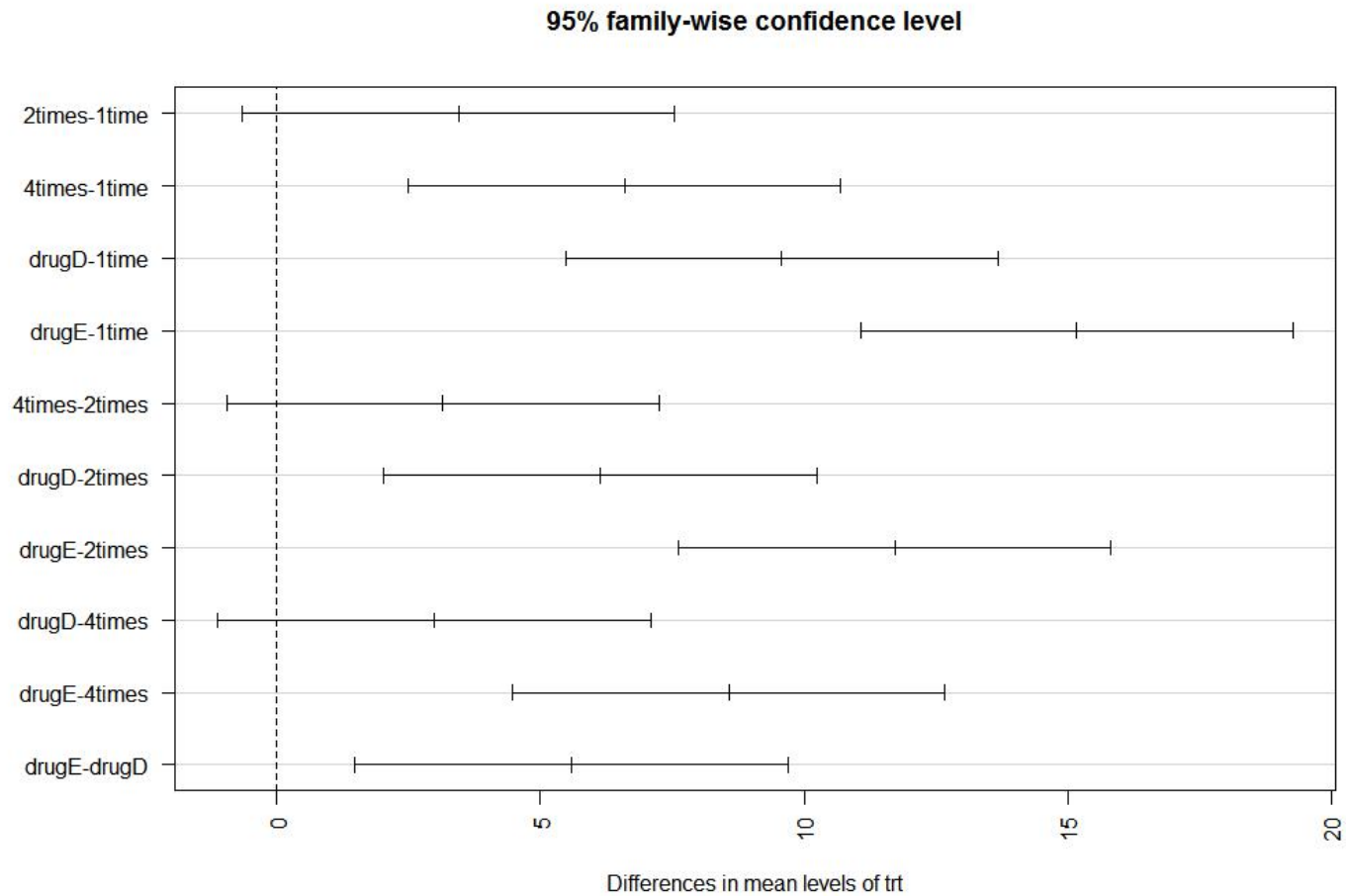
可以看到，1time和2times的均值差异不显著（ $p=0.138$ ），而1time和4times间的差异非常显著（ $p<0.001$ ）。

`par(las=2)` #用来旋转轴标签

`par(mar=c(5,8,4,2))` #用来增大左边界的面积

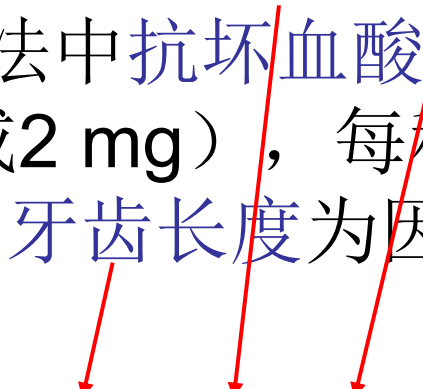
`plot(TukeyHSD(fit))`

# 多重比较 (3/3)



# 双因素方差分析（1/7）

- 例：基础安装中的ToothGrowth数据集：随机分配60只豚鼠，分别采用两种喂食方法（橙汁或维生素C），各喂食方法中抗坏血酸含量有三种水平（0.5 mg、1 mg或2 mg），每种处理方式组合都被分配10只豚鼠。牙齿长度为因变量。



	len	supp	dose
1	4.2	VC	0.5
2	11.5	VC	0.5
3	7.3	VC	0.5
4	5.8	VC	0.5
5	6.4	VC	0.5
6	10.0	VC	0.5
7	11.2	VC	0.5

# 双因素方差分析 (2/7)

```
attach(ToothGrowth)
```

```
table(supp,dose)
```

```
      dose
supp 0.5  1  2
OJ   10 10 10
VC   10 10 10
```

列联表表明该设计是均衡设计（各设计单元中样本大小都相同（10））。

获得各单元的均值和标准差

```
aggregate(len, by=list(supp,dose), FUN=mean)
```

```
aggregate(len, by=list(supp,dose), FUN=sd)
```

```
fit <- aov(len ~ supp*dose)
```

```
summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
supp	1	205.4	205.4	12.317	0.000894	***
dose	1	2224.3	2224.3	133.415	< 2e-16	***
supp:dose	1	88.9	88.9	5.333	0.024631	*
Residuals	56	933.6	16.7			

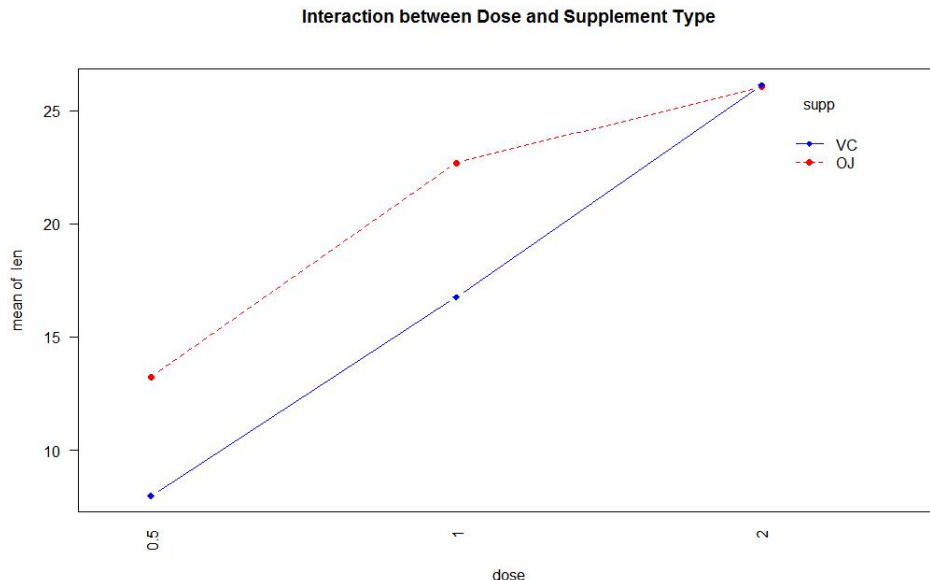
用**summary()**函数得到方差分析表，可以看到主效应（**supp**和**dose**）和交互效应都非常显著。

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 双因素方差分析（3/7）

- 可用`interaction.plot()`函数来展示双因素方差分析的交互效应。

```
interaction.plot(dose, supp, len, type="b",  
col=c("red", "blue"), pch=c(16, 18), main =  
"Interaction between Dose and Supplement Type")
```



- 图形展示了（两种喂食方法）各种剂量喂食下豚鼠牙齿长度的均值。

# 双因素方差分析（4/7）

- 可以用gplots包中的**plotmeans()**函数来展示交互效应。

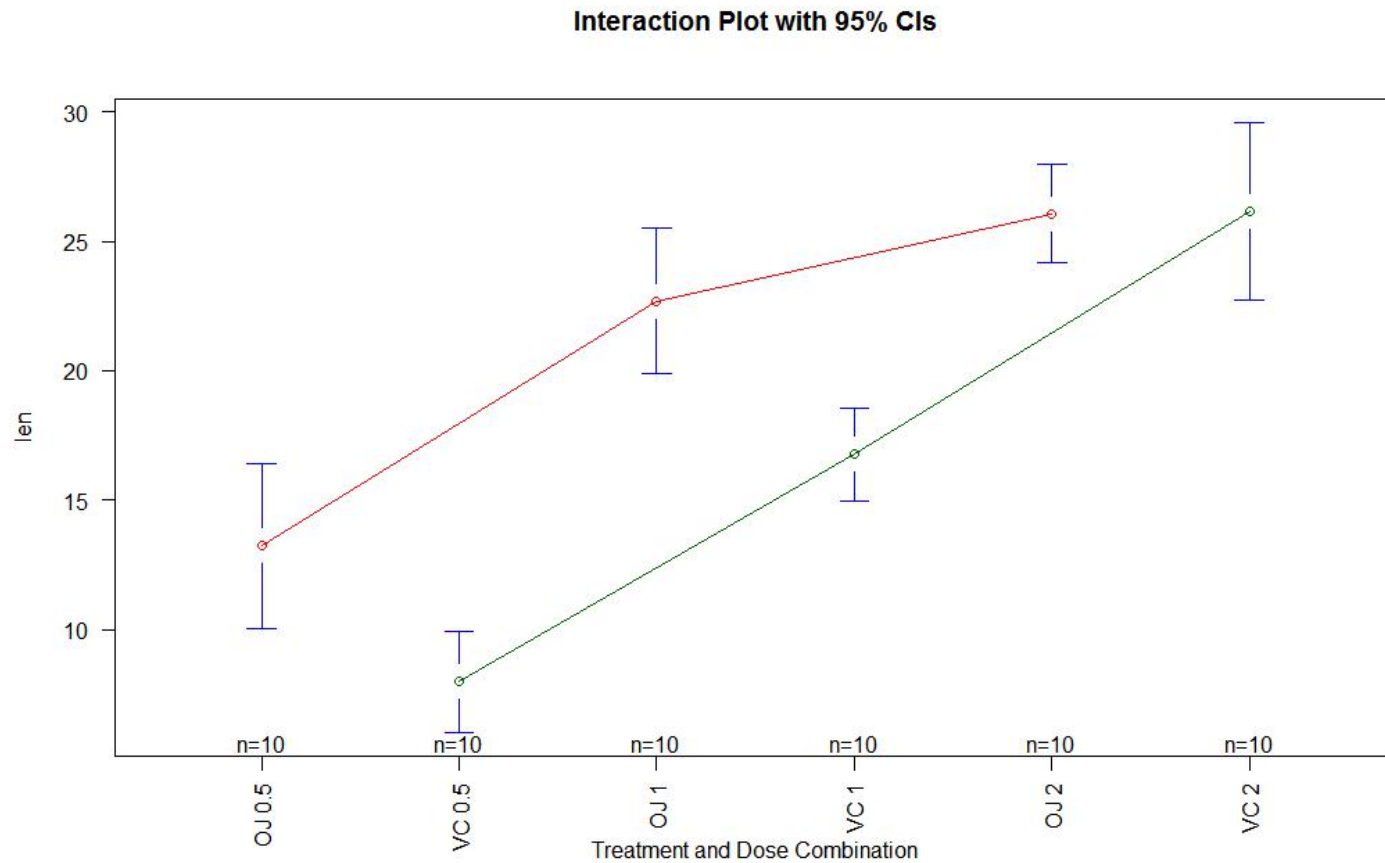
```
library(gplots)
```

```
plotmeans(len ~ interaction(supp, dose, sep=" "),  
            connect=list(c(1, 3, 5),c(2, 4, 6)),  
            col=c("red","darkgreen"),  
            main = "Interaction Plot with 95% CIs",  
            xlab="Treatment and Dose Combination")
```

- 图在下页



# 双因素方差分析（5/7）



喂食方法和剂量对牙齿生长的交互作用。用`plotmeans()`函数绘制的95%的置信区间的牙齿长度均值。

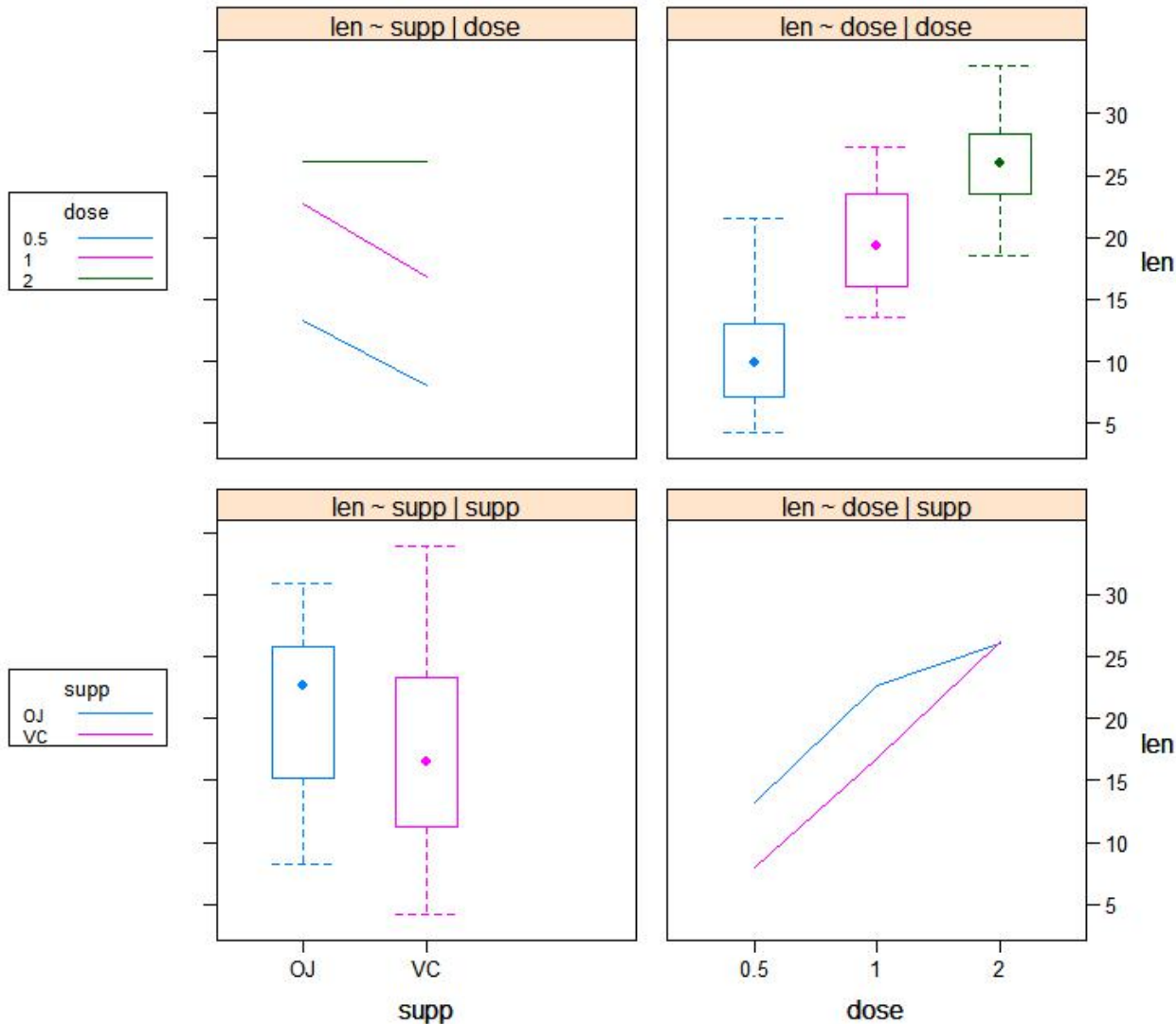
# 双因素方差分析（6/7）

- 可以HH包中的interaction2wt()函数来**可视化**结果，图形对任意顺序的因子设计的主效应和交互效应都会进行展示。
- **推荐使用**。因为它能展示任意复杂度设计（双因素方差分析、三因素方差分析等）的主效应（箱线图）和交互效应。
- 代码：

```
library(HH)  
interaction2wt(len~supp*dose)
```
- 图在下页。

# 双因素方差分析 (7/7)

len: main effects and 2-way interactions



- 之前的三幅图形都表明随着橙汁和维生素C中的抗坏血酸剂量的增加，牙齿长度变长。
- 对于0.5 mg和1 mg剂量，橙汁比维生素C更能促进牙齿生长；对于2 mg剂量的抗坏血酸，两种喂食方法下牙齿长度增长相同。

# 组间差异的非参数检验

# 组间差异的非参数检验

- 如果数据无法满足t检验或ANOVA的参数假设，可以转而使用非参数方法。比如，若结果变量在本质上就严重偏倚或呈现有序关系。
- 注：非参数检验是在总体方差未知或知道甚少的情況下，利用样本数据对总体分布形态等进行推断的方法。由于非参数检验方法在推断过程中不涉及有关总体分布的参数，因而得名为“非参数”检验。

# 两组的比较（1/2）

- 若两组数据独立，可以使用Wilcoxon秩和检验（更广为人知的名字是Mann–Whitney U检验）来评估观测是否是从相同的概率分布中抽得的。
- 第一种调用格式为：  
**wilcox.test**(y ~ x, data)  
其中的y是数值型变量，而x是一个二分变量。
- 第二种调用格式为：  
**wilcox.test**(y1, y2)  
其中的y1和y2为各组的结果变量。

## 两组的比较 (2/2)

- 例：原假设：南方和非南方拥有相同监禁概率。

```
library(MASS)
wilcox.test(Prob ~ So, data=UScrime)
```

监禁概率 “南方”标记

```
wilcoxon rank sum test

data: Prob by So
W = 81, p-value = 8.488e-05
alternative hypothesis: true location shift is not equal to 0
```

拒绝原假设

# 多于两组的比较（1/3）

- 如果无法满足**ANOVA**设计的假设，那么可以使用非参数方法来评估组间的差异。
- 如果各组独立，可以使用**Kruskal—Wallis**检验。

**kruskal.test(y ~ A, data)**

- 如果各组不独立（如重复测量设计或随机区组设计），可以使用**Friedman**检验。

**friedman.test(y ~ A | B, data)**



## 多于两组的比较（2/3）

- 考虑7.4节中的**state.x77**数据集。它包含了美国各州的人口、收入、文盲率、预期寿命、谋杀率和高中毕业率数据。
- 如果想比较美国四个地区（东北部、南部、中北部和西部）的文盲率，应该怎么做呢？

## 多于两组的比较 (3/3)

```
states <- data.frame(state.region, state.x77)  
#将地区名称添加到数据集中。
```

```
kruskal.test(illiteracy ~ state.region,  
data=states)
```

```
Kruskal-wallis rank sum test
```

```
data: illiteracy by state.region
```

```
Kruskal-wallis chi-squared = 22.672, df = 3, p-value = 4.726e-05
```

- 显著性检验的结果意味着美国四个地区的文盲率各不相同 ( $p < 0.001$ )。