

概率统计与R软件

李东风

2010 年 12 月 2 日

目录

第一章 R软件基础与概率论	1
1.1 R软件介绍	2
1.1.1 R的历史和特点	2
1.1.2 R的使用样例	3
1.1.3 R向量	6
1.1.4 R矩阵	12
1.1.5 R输入输出	13
1.1.6 R函数	14
1.2 概率论	16
1.2.1 事件和概率	16
1.2.2 随机变量和数学期望	19
1.2.3 随机变量联合分布	21
1.2.4 常用分布	21
第二章 置信区间和假设检验	27
2.1 总体与样本	28
2.1.1 统计基本概念	28
2.1.2 探索性数据分析(EDA)	29
2.2 参数估计	32
2.2.1 点估计方法	32
2.2.2 置信区间	35
2.3 假设检验	38
2.3.1 假设检验的概念	38
2.3.2 单正态总体参数的检验	39
2.3.3 两正态总体参数的检验	43
2.3.4 成对t检验	45
2.3.5 正态性检验	46
2.3.6 比例的检验	47
2.4 练习	49
第三章 方差分析	53
3.1 两组比较进阶	54
3.1.1 位置检验的非参数方法	54
3.1.2 卡方检验	57
3.2 单因素方差分析	60
3.2.1 单因素方差分析	60

3.2.2	多重比较	62
3.2.3	方差齐性检验	63
3.2.4	非参数Kruskal-Wallis检验	64
3.3	两因素方差分析	65
3.4	协方差分析	69
3.5	正交设计	71
第四章	统计模型	77
4.1	相关与回归	78
4.1.1	相关分析	78
4.1.2	一元回归分析	80
4.1.3	曲线拟合	83
4.1.4	多元线性回归	84
4.1.5	Logistic回归	85
4.2	多元分析	87
4.2.1	主成份分析	87
4.2.2	因子分析	88
4.2.3	判别分析	89
4.2.4	聚类分析	89
4.2.5	典型相关分析	91
4.3	时间序列分析	93
4.3.1	时间序列	93
4.3.2	时间序列的检验	95
4.3.3	时间序列分析模型	97
第五章	质量管理	99
5.1	质量概述	100
5.2	质量理念与框架	103
5.3	过程管理	109
5.4	绩效测量与信息管理	111
5.5	六西格玛的原理	116
5.5.1	六西格玛的统计基础	116
5.5.2	六西格玛项目的选择	118
5.5.3	六西格玛的问题解决	119
5.6	统计思考和应用	122
5.7	六西格玛设计	124
5.7.1	概念开发中的工具	124
5.7.2	设计开发中的工具	125
5.7.3	可靠性预测	126

5.7.4	过程优化中的工具	127
5.7.5	设计验证中的工具	129
5.8	过程改进工具	131
5.8.1	过程改进的方法论	131
5.8.2	过程改进的基本工具	131
5.9	统计过程控制	145
5.9.1	质量控制测量指标	145
5.9.2	R的qcc软件包	145
5.9.3	能力与受控	149
5.9.4	计量值数据控制图	149
5.9.5	计数值的数据控制图	157

第一章 R软件基础与概率论

第一讲内容概要

- R软件基础：向量，矩阵，输入输出，函数。
- 概率，随机变量
- 分布

§1.1 R软件介绍

R软件介绍

- R介绍。
- 向量，下标。
- 矩阵。
- 输入输出。
- 函数。

§1.1.1 R的历史和特点

R的历史

- S语言：Rick Becker, John Chambers等人在贝尔实验室开发¹，著名的C语言、Unix系统也是贝尔实验室开发的。

¹ 第一个版本开发于1976-1980，基于Fortran；于1980年移植到Unix，并对外发布源代码。1984年出版的“棕皮书”

Becker, Richard A. and John M. Chambers (1984), “S: An Interactive Environment for Data Analysis and Graphics”, Wadsworth Advanced Books Program, Belmont CA

总结了1984年为止的版本，并开始发布授权的源代码。这个版本叫做旧S。与我们现在用的S语言有较大差别。

1989-1988对S进行了较大更新，变成了我们现在使用的S语言，称为第二版。1988年出版的“蓝皮书”做了总结：

Becker, Richard A., John M. Chambers and Allan R. Wilks (1988), “The New S Language”, Chapman and Hall, New York.

1992年出版的“白皮书”描述了在S语言中实现的统计建模功能，增强了面向对象的特性。软件称为第三版，这是我们现在用的多数版本。

Chambers, John M. and Trevor Hastie, eds. (1992), “Statistical Models in S”, Chapman and Hall, New York.

1998年出版的“绿皮书”描述了第四版S语言，主要是编程功能的深层次改进。现行的S系统并没有都采用第四版，S-PLUS的第5版才采用了S语言第四版。

John M. Chambers (1998), “Programming with Data,” New York: Springer

- 商业版本为S-PLUS, 1988年发布, 现在为Tibco Software拥有。
- R是一个自由软件, GPL授权, 最初由新西兰Auckland 大学的Ross Ihaka 和Robert Gentleman 于1997年发布, 实现了与S语言基本相同的功能和统计功能。现在由R核心团队开发, 但全世界的用户都可以贡献软件包。见R的网站: <http://www.r-project.org/>。

R的特点

- 自由软件, 免费;
- 完整的程序设计语言, 基于函数和对象, 可以自定义函数, 调入C、C++、Fortran编译的代码;
- 具有完善的数据类型, 如向量、矩阵、因子、数据集、一般对象等, 代码像伪代码一样简洁、可读。
- 强调交互式数据分析, 支持复杂算法描述, 图形功能强。
- 实现了经典的、现代的统计方法, 如参数和非参数假设检验、线性回归、广义线性回归、非线性回归、可加模型、树回归、混合模型、方差分析、判别、聚类、时间序列分析等。
- 统计科研工作者广泛使用R进行计算和发表算法。R有数以千计的软件包。

§1.1.2 R的使用样例

R安装、启动、退出

- 从R的网站<http://www.r-project.org/>获得软件并安装。
- 在MS Windows系统中, 把R的程序快捷方式复制到工作目录如C:\work中, 从右键菜单选择“属性”, 把“起始位置”栏清空。
- 双击R图标可以启动R, 为命令行界面: 输入一行命令, 就在后面显示计算结果。
- 可以用向上和向下箭头访问历史命令; 可以从上面的命令中用Ctrl+C复制或Ctrl+X复制, Ctrl+X会自动粘贴到当前行。
- 关闭窗口退出, 退出时询问是否保存工作空间, 如果保存, 下次进入时可以访问前面各次已定义的变量。

R向量例子

- R语言以向量为最小单位。用<-赋值。

```
x1 <- 0:100
```

- 可以对向量进行通常的数乘，或与一个常数相加、相减等：

```
x2 <- x1 * 2 * pi / 100
```

- 一元函数可以对向量的每个元素取值：

```
y1 <- sin(x2)
```

方便的绘图功能

- 曲线图：

```
plot(x2, y1, type="l")
```

- 添加参考线：

```
abline(h=0, lwd=2)  
abline(v=(0:4)/2*pi, lty=3, col="gray")
```

- 添加一条余弦曲线：

```
y2 <- cos(x2)  
lines(x2, y2, lty=2, col="green")
```

- 其它图形、图像演示：

```
demo("graphics")  
demo("image")
```

R基本统计功能演示

- 读入CSV格式的数据集:

```
cl <- read.csv("class.csv", header=TRUE)
```

- 概括统计:

```
summary(cl)
```

- 统计函数:

```
mean(cl$height)
var(cl$height)
```

包括sum, mean, var, sd, min, max, range等。

回归分析

- 回归结果返回为一个“对象”，可以进一步利用回归结果。

```
lm1 <- lm(weight ~ height + age + sex,
           data=cl)
print(summary(lm1))
```

- 逐步回归:

```
lm2 <- step(lm1,
            weight ~ height + age + sex, data=cl)
```

记录输出结果

- 例:

```
sink("myres.txt", split=TRUE)
print(A)
print(Ai)
sink()
```

- `sink`函数指定一个输出记录文件，结果在屏幕显示的同时输出到指定文件中。
- 用空`sink()`关闭输出记录。

在线帮助

- 用帮助菜单中“html帮助”查看帮助文档。“Search engine and keywords”项下面有分类的帮助。
- 在命令行，用问号后面跟随函数名查询某函数的帮助。

§1.1.3 R向量

R数据类型

- 数据类型有向量、矩阵、列表、对象等。
- 常量和变量。
- 变量名规则。句点可以作为变量名的一部分，但有部分隐含作用。

常量

- 数值型：包括整型、单精度、双精度等，一般不需要区分。
- 字符型。
- 逻辑性：TRUE和FALSE。
- 缺失值：用NA表示。
- 复数：如 $2.2 + 3.5i$, $1i$ 等。

向量与赋值

- 用`<-`赋值。
- 用`c()`函数组合向量元素。

```
marks <- c(10, 6, 4, 7, 8)
x <- c(1:3, 10:13)
x1 <- c(1, 2)
x2 <- c(3, 4)
x <- c(x1, x2)
x
```


- 显示向量时在左边方括号中提示显示行的第一个元素的下标:

```
> 1234501:1234520
[1] 1234501 1234502 1234503 1234504 1234505 1234506
[7] 1234507 1234508 1234509 1234510 1234511 1234512
[13] 1234513 1234514 1234515 1234516 1234517 1234518
[19] 1234519 1234520
```

比如, 第13号元素1234513, 前面有一个 “[13]”。

- `length(x)`可以求x的长度。

向量与标量运算

- 向量与标量的运算为每个元素与标量的运算。
- 四则运算: `+` `-` `*` `/` `^`。
- 如:

```
> x <- c(1, 10)
> x + 2
[1] 3 12
> x - 2
[1] -1 8
> x * 2
[1] 2 20
> x / 2
[1] 0.5 5.0
> x ^ 2
[1] 1 100
> 2 / x
[1] 2.0 0.2
> 2 ^ x
[1] 2 1024
```

等长向量运算

- 为对应元素两两运算。
- 如:

```
> x1 <- c(1, 10)
> x2 <- c(4, 2)
> x1 + x2
[1] 5 12
> x1 - x2
[1] -3 8
> x1 * x2
[1] 4 20
> x1 / x2
[1] 0.25 5.00
```

不等长向量的运算

- 如果长度为倍数关系，每次从头重复利用短的一个。如

```
> x1 <- c(1, 10)
> x2 <- c(1, 3, 5, 7)
> x1 + x2
[1] 2 13 6 17
> x1 * x2
[1] 1 30 5 70
```

- 如果长度不是倍数关系，会给出警告信息。如

```
> c(1,2) + c(1,2,3)
[1] 2 4 4
警告信息:
In c(1, 2) + c(1, 2, 3) :
  长的对象长度不是短的对象长度的整倍数
```

向量函数

- 一元函数以向量为自变量，对每个元素计算。如sqrt, log, exp, sin, cos, tan等。
- 统计函数: sum, mean, var, sd, min, max, range等。

- `cumsum`和`cumprod`计算累加和累乘积。
- `sort(x)`返回排序结果; `order(x)`返回排序用的下标。如

```
> sort(c(3, 5, 1))  
[1] 1 3 5  
> order(c(3, 5, 1))  
[1] 3 1 2
```

- 有缺失值的运算: 缺失元素参加的运算相应结果元素仍缺失。

seq函数

- `seq`函数是冒号运算符的推广。
- 如: `seq(5)`等同于`1:5`。
- `seq(2,5)`等同于`2:5`。
- `seq(11, 15, by=2)`产生11,13,15。
- `seq(0, 2*pi, length=100)`产生从0到 2π 的等间隔序列, 序列长度指定为100。
- S函数可以带自变量名调用。
- `seq(to=5, from=2)`等同于`2:5`。

rep函数

- 产生一个初值为零的长度为n的向量: `x <- rep(0, n)`。
- `rep(c(1,3), 2)`相当于`c(1,3,1,3)`。
- `rep(c(1,3), c(2,4))`相当于`c(1,1,3,3,3,3)`。
- `rep(c(1,3), each=2)`相当于`c(1,1,3,3)`。

逻辑向量

- 逻辑值: TRUE, FALSE, 缺失时为NA。
- 向量比较结果为逻辑型向量。
- 比较运算符为<, <=, >, >=, ==, !=。

- 逻辑运算符为`&`, `|`, `!`, 另外`&&`, `||`是短路的标量逻辑与和逻辑或。
- `all(cond)`测试`cond`的所有元素为真; `any(cond)`测试`cond`至少一个元素为真。

字符型向量

- 字符型向量是元素为字符串的向量。
- `paste`函数: 连接两个字符型向量, 元素一一对应连接。缺省用空格连接。如`paste(c("ab", "cd"), c("ef", "gh"))`相当于`c("ab ef", "cd gh")`。
- 可以一对多连接, 可以在连接时把数值型转换为字符型, 如`paste("x", 1:3)`相当于`c("x 1", "x 2", "x 3")`。
- 用`sep`=指定分隔符, 如`paste("x", 1:3, sep="")`相当于`c("x1", "x2", "x3")`。
- 使用`collapse`=参数连接字符型向量的各个元素为一个字符串。如`paste(c("a", "b"), collapse="")`相等于`"ab"`。

字符型数据获取

- 利用字符串处理的方法可以从网页文件或不规则的Excel文件读取数据。
- `grep`, `grep1`函数从字符串中查询某个模式, 模式用perl格式的**正则表达式**(regular expression)定义。
- `sub`, `gsub`替换某模式。
- 正则表达式功能强大但也不容易掌握。
- `substr`和`substring`函数抽取子字符串, 也可以修改子字符串。

复数向量

- 复数常数表示如`3.5+2.4i`, `1i`。
- 用函数`complex`生成复数向量, 指定实部和虚部。如`complex(c(1,0,-1,0), c(0,1,0,-1))`相当于`c(1+0i, 1i, -1+0i, -1i)`。
- 可以用`mod`和`arg`指定模和辐角, 如`complex(mod=1, arg=(0:3)/2*pi)`结果同上。
- 用`Re(z)`求`z`的实部, 用`Im(z)`求`z`的虚部, 用`Mod(z)`或`abs(z)`求`z`的模, 用`Arg(z)`求`z`的辐角, 用`Conj(z)`求`z`的共轭。
- `log`, `exp`, `sin`等函数对复数也有定义。

向量下标

- 设 `x <- c(1, 4, 6.25)`。
- `x[2]` 取出第二个元素; `x[2] <- 99` 修改第二个元素。
- `x[c(1,3)]` 取出第1、3号元素; `x[c(1,3)] <- c(11, 13)` 修改第1、3号元素。
- 下标可重复, 如 `x[c(1,3,1)]` 为 1, 6.25, 1。
- 负下标表示“扣除”, 如 `x[-c(1,3)]` 为第二个元素 4。

逻辑下标

- 下标可以是一个条件, 如 `x[x>3]` 取出 4, 6.25。
- 例: 示性函数。设输入向量 `x`, 求每个元素的示性函数值(元素非负时取1, 否则取0)。

```
y <- numeric(length(x))
y[x >= 0] <- 1
y[x < 0] <- 0 # 此语句多余
```

元素名

- 向量可以为每个元素命名。如

```
ages <- c("李明"=30, "张聪"=25, "刘颖"=28)
```

或

```
ages <- c(30, 25, 28)
names(ages) <- c("李明", "张聪", "刘颖")
```

- 这时可以用 `ages["张聪"]`, `ages[c("李明", "刘颖")]` 这样的方法访问。
- 这样建立了字符串到数值的映射表。
- 在矩阵和数据框中尤其有用。

整个数组

- 用 `x[] <- 0` 把 `x` 的所有元素赋0，但长度不变。
- 这与 `x <- 0` 不同，后者把 `x` 变成标量0。

§1.1.4 R矩阵

R矩阵

- 矩阵用 `matrix` 函数定义。缺省为按列存储。

```
A <- matrix(1:6, nrow=3, ncol=2)
B <- matrix(1,-1, 1,1,
  nrow=2, ncol=2, byrow=TRUE)
C1 <- A %*% B
C2 <- A + 2
C3 <- A / 2
```

- 可以与标量四则运算；相同形状矩阵可以四则运算，表示对应元素的运算。
- 可以用 `%*%` 表示矩阵乘法。

矩阵下标

- `A[1,]` 取出 `A` 的第一行。
- `A[,1]` 取出 `A` 的第一列。
- `A[1:3,1:2]` 取出子矩阵。
- 用列名访问矩阵：

```
colnames(A) <- c("x", "y")
A[, "y"]
```

矩阵求逆

- `solve(B)` 返回 `B` 的逆矩阵。
- `solve(B, c(1,2))` 解线性方程组

$$\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} x = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

§1.1.5 R输入输出

R输入输出

- 显示某个变量: `print(x)`。
- `cat()`函数: 用如

```
cat("x =", x, "\n")
```

- 写入指定文件中并且在文件尾部添加:

```
cat("x =", x, "\n",  
    file="res.txt", append=TRUE)
```

- 把某些变量保存起来, 以及从文件中恢复某些变量:

```
save(x, y, file="x-y.RData")  
load("x-y.RData")
```

- 开始把输出分流到指定的文件中:

```
sink("allres.txt", split=TRUE)
```

- 终结前面的`sink()`的作用:

```
sink()
```

- 用`scan()`函数输入文本文件中的数值向量, 数值之间以空格分开。如

```
cat(1:12, file="c:/data/x.txt")  
x <- scan("c:/data/x.txt")  
A1 <- matrix(x, nrow=4, ncol=3, byrow=TRUE)
```

- 用`read.csv()`把一个第一行为变量名的CSV格式的数据表读入为R的数据框(data frame)。如

```
cl <- read.csv("class.csv", header=TRUE)
```

- 数据框是类似于矩阵的数据结构，但各列的数据类型可以不同，比如可以有字符串列、因子列（因子是对分类变量用数值编码后的结果）。数据框每列必须有变量名。访问数据框的子集，可以采用和访问矩阵子集相同的语法。
- 可以用如`as.matrix(cl[,c("height", "weight")])`的办法把数据框的数值类型的子集转换为矩阵。

§1.1.6 R函数

R函数

- 编程语言中函数的优点：代码复用、模块化设计。
- 把编程任务分解成小的模块，每个模块用一个函数实现，可以降低复杂性，防止变量混杂。
- 函数的自变量是只读的，函数中定义的局部变量只在函数运行时起作用，不会与外部或其它函数中同名变量混杂。
- 函数返回一个对象作为输出，如果需要返回多个变量，可以用列表进行包装。
- 例如

```
fsub <- function(x, y=0){  
  cat("x=", x, " y=", y, "\n")  
  x - y  
}
```

函数体内最后一个表达式为返回值。也可以用`return(x)`返回值。

- 定义时可以规定可选参数如上面的`y`。调用时可选参数可以省略，可以按次序对准，也可以用名字对准，如


```
fsub(5,3)
fsub(5)
fsub(x=5, y=3)
fsub(y=3, x=5)
```

- `browser()`函数可以令程序进入跟踪运行状态，在跟踪运行状态可以查看变量值，用`n`命令逐句运行，用`c`命令恢复正常运行。

练习

- 若`x <- 1:100`，如何访问`x`的第11—20和41, 45号元素？
- 若`x <- sin((1:100)/100*2*pi)`，访问`x`中大于0.75的元素集合。
- 若文件“d:\work\work.txt”中保存了如下数值：

```
2 3 5 7 11
13 17 19
```

把各数值读入到向量`x`中。

- 写一个函数`f(x)`，计算返回

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^3}{S^3}$$

其中 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ， n 为向量`x`的长度。

§1.2 概率论

§1.2.1 事件和概率

事件

- 随机试验或观测的可能结果称为**事件**，如{掷硬币得到正面}，{产品的疵点超过5个}，{明天有雨}。用集合 A, B, C 等表示。
- A 或 B 至少有一个发生记为 $A \cup B$ 。
- A 和 B 同时发生记为 AB 或 $A \cap B$ 。

概率

- 事件 A 的**概率** $P(A)$ 是一个在 $[0, 1]$ 取值的数，表示可能性的大小。
- $P(A)$ 可以是一个试验多次重复后事件 A 发生的比例的极限。
- 概率也可以是主观想象中的可能性大小。

概率的性质

- $0 \leq P(A) \leq 1$.
- 若事件 A 与 B 不能同时发生，称 A 和 B 互斥，记为 $AB = \phi$ ，则有概率的可加性：

$$P(A \cup B) = P(A) + P(B).$$

•

$$P(A \text{不发生}) = 1 - P(A).$$

- 比如，掷两枚硬币，记

$$A = \{\text{同为正面}\}$$

$$B = \{\text{同为反面}\}$$

$$C = \{\text{朝向相同}\}$$

$$D = \{\text{朝向相反}\}$$

则

$$\begin{aligned}P(A) &= \frac{1}{4} \\P(B) &= \frac{1}{4} \\P(C) &= P(A) + P(B) = \frac{1}{2} \\P(D) &= 1 - P(C) = \frac{1}{2}\end{aligned}$$

条件概率

- 同时掷两枚硬币，在已知朝向相同的条件下，发生“都是正面”的概率为 $\frac{1}{2}$ 。
- 已知 A 发生的条件下 B 发生的**条件概率**定义为

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

- 于是

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

这个公式是一般公式，不要求 $P(A) \neq 0$ 。

- $P(B|A)$ 不要求 B 是 A 的一部分。比如

$$A = \{\text{朝向相同}\} = \{(\text{正}, \text{正}), (\text{反}, \text{反})\}$$

$$B = \{\text{至少一个正面}\}$$

$$= \{(\text{正}, \text{正}), (\text{正}, \text{反}), (\text{反}, \text{正})\}$$

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P((\text{正}, \text{正}))}{P(A)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

事件的独立性

- 称事件 A 和 B 独立，若

$$P(AB) = P(A)P(B)$$

- 解释：若 $P(A) \neq 0$, $P(B) \neq 0$, A 与 B 独立当且仅当 $P(B|A) = P(B)$, $P(A|B) = P(A)$, 即 A 和 B 发生与否不影响对方发生的概率。
- 独立不是互斥，因为 A 和 B 互斥是指 A 和 B 不能同时发生，这样是不独立的。“凡是敌人反对的，我们就要拥护”？

全概公式

- 设试验的所有可能结果分解为互斥的事件 $A_1, A_2, \dots, A_k$, 即 $A_1, A_2, \dots, A_k$ 互斥且 $P(A_1 \cup A_2 \cup \dots \cup A_k) = 1$ 。
- 任一事件 B 就可以分解为

$$B = (BA_1) \cup (BA_2) \cup \dots \cup (BA_k)$$

分解是互斥的。

- 由概率可加性得如下全概公式

$$\begin{aligned} P(B) &= P(BA_1) + P(BA_2) + \dots + P(BA_k) \\ &= \sum_{j=1}^k P(BA_j) = \sum_{j=1}^k P(A_j)P(B|A_j) \end{aligned}$$

- 例如, 同时掷两枚硬币, $A_1 = \{\text{朝向相同}\}$, $A_2 = \{\text{朝向相反}\}$, $B = \{\text{至少有一个正面}\}$ 。
- A_1 与 A_2 互斥且构成所有结果。 $P(A_1) = P(A_2) = \frac{1}{2}$ 。

•

$$\begin{aligned} P(B) &= P(A_1)P(B|A_1) + P(A_2)P(B|A_2) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 1 = \frac{3}{4} \end{aligned}$$

贝叶斯公式

- 如果事件 A_1 和 A_2 互斥, 事件 B 发生蕴含 $A_1 \cup A_2$ 发生, 则

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2)$$

- 反向的条件概率 $P(A_1|B)$ 为

$$P(A_1|B) = \frac{P(BA_1)}{P(B)} = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)}$$

- 例如, 同时掷两枚硬币, $A_1 = \{\text{朝向相同}\}$, $A_2 = \{\text{朝向相反}\}$, $B = \{\text{至少有一个正面}\}$ 。
- A_1 与 A_2 互斥且构成所有结果。 $P(A_1) = P(A_2) = \frac{1}{2}$ 。

•

$$\begin{aligned} P(A_1|B) &= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot 1} = \frac{1}{3} \end{aligned}$$

§1.2.2 随机变量和数学期望

随机变量

- 如果随机试验的结果可以用数值来表示, 我们称这样的变量为**随机变量**。
- 例如, 掷一枚硬币, $X = 1$ 表示正面, $X = 0$ 表示反面。
- 同时掷两枚硬币, 可以用两个随机变量 (X, Y) 表示, 取值为 $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$ 。
- 测量某地中午12时的气温, 用随机变量 X 表示, 可以取连续值。

分布函数

- 随机变量是关于随机试验的结果概率的模型。我们关心随机变量取值的概率, 如硬币试验中 $P(X = 1)$, 气温测量中 $P(20 \leq X \leq 28)$ 。
- 分布函数

$$F(x) = P(X \leq x) \quad x \in (-\infty, \infty)$$

包含了随机变量的取值概率的规律。

- 分布指随机变量取值概率的规律。
- 若 $F(x)$ 可逆, 其逆函数 $q(y) = F^{-1}(y), y \in (0, 1)$ 称为该分布的**分位数函数**。一般地定义 $q(y) = \inf\{x : F(x) \geq y\}$ 。
- 随机变量分为离散型和连续型。
- 离散型随机变量 X 取有限个值 $\{x_1, x_2, \dots, x_k\}$ 或可列个值 $\{x_1, x_2, \dots\}$ 。可以用概率函数(PMF)来刻画其取值概率的规律:

$$P(X = x_j) = p_j, \quad j = 1, 2, \dots$$

- 连续型随机变量在一个区间上取值。其分布可以用分布密度(PDF) $f(x)$ 刻画:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

$$P(a < X \leq b) = \int_a^b f(t) dt$$

$$P(X = a) = 0$$

数学期望

- 随机变量 X 取值具有不确定性，如何求其平均值？
- 数学期望

$$EX = \sum_j x_j p_j \quad (\text{离散型})$$

$$EX = \int_{-\infty}^{\infty} x f(x) dx \quad (\text{连续型})$$

- EX 代表了随机变量 X 取值的平均。
- 性质:

$$E(\alpha X + \beta Y) = \alpha EX + \beta EY$$

- 性质:

$$E(g(X)) = \begin{cases} \sum_j g(x_j) p_j & (\text{离散型}) \\ \int g(x) f(x) dx & (\text{连续型}) \end{cases}$$

例子: 彩票获利期望

- 彩票获利的数学期望。设彩票面值为1元，共发行一百万张，其中有5张各奖10万元，其余没有奖。
- 令 $X = 100000 - 1$ 表示中奖时的获利， $X = -1$ 表示不中时的获利。
-

$$EX = (100000 - 1) * \frac{5}{1000000} + (-1) * \frac{1000000 - 5}{1000000}$$

$$= -0.5$$

- 即买一注平均亏损0.5元。

随机变量的数字特征

- 分布函数、PMF、PDF是随机变量的取值概率的完整刻画。
- 希望用少数的数值来概括分布主要特征。
- 位置特征: EX 为分布的平均值。
- 方差:

$$\text{Var}(X) = E(X - EX)^2 = E(X^2) - (EX)^2$$

- 标准差: $\sigma = \sqrt{\text{Var}(X)}$ 。方差和标准差概括了随机变量的分散程度特征。

§1.2.3 随机变量联合分布

随机变量联合分布

- 设有两个随机变量 X 和 Y ，它们的取值可能是有关系的。比如，同时掷两只骰子，令 X 表示点数和， Y 表示第一个骰子的点数，两者显然有关系。
- 两个随机变量 X 和 Y 的分布情况用联合分布表示，联合分布函数为

$$F(x, y) = P(X \leq x, Y \leq y), \quad x \in (-\infty, \infty), y \in (-\infty, \infty)$$

- 如果 X 与 Y 都是连续取值的，其分布常可以用联合分布密度 $f(x, y)$ 表示。

独立随机变量

- 称随机变量 X 与 Y 独立，若

$$F(x, y) = F(x)F(y), \quad \forall x, y$$

- 若连续型的 X 与 Y 独立，则联合密度

$$f(x, y) = f_X(x)f_Y(y)$$

其中 $f_X(x)$ 为 X 的密度， $f_Y(y)$ 为 Y 的密度。

- 定义 X 和 Y 的协方差为 $\text{Cov}(X, Y) = E[(X - EX)(Y - EY)]$ ，相关系数为

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

- X 与 Y 独立时有

$$E(XY) = E(X)E(Y)$$

$$\text{Cov}(X, Y) = 0$$

$$\rho(X, Y) = 0$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

- 随机变量联合分布和独立性可以推广到 n 个随机变量 $X_1, X_2, \dots, X_n$ 。如果 $X_1, X_2, \dots, X_n$ 相互独立且服从相同的分布，记为iid。

§1.2.4 常用分布

两点分布(Bernoulli分布) $B(1, p)$

- $X = 0, 1$ 。 $P(X = 1) = p$, $P(X = 0) = 1 - p$ 。 $p \in [0, 1]$ 。
- 表示结果的成功与失败。
- $EX = p$ 。

二项分布 $B(n, p)$

- n 为正整数, $p \in [0, 1]$ 。
- $X = 0, 1, \dots, n$ 。表示 n 次独立重复试验中成功的次数, 成功概率为 p 。
- $X = X_1 + X_2 + \dots + X_n$, 其中 $X_1, X_2, \dots, X_n$ 相互独立, 分布服从 $B(1, p)$ 。
- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \dots, n$ 。
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ 是从 n 个中选取 k 个的不同选法。
- $EX = np$, $\text{Var}(X) = np(1-p)$ 。

R中与分布有关的函数

- `pbinom(x, n, p)` 计算 $B(n, p)$ 在 x 处的分布函数值。
- `dbinom(x, n, p)` 计算 $B(n, p)$ 在 x 处的概率函数值。
- `qbinom(y, n, p)` 计算 $B(n, p)$ 在 y 处的分位数函数值, 定义为使得 $F(x) \geq y$ 的最小的 x 。
- `rbinom(K, n, p)` 产生 K 个服从 $B(n, p)$ 的随机数。

泊松分布

- $X = 0, 1, 2, \dots$ 。
- $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$, $\lambda > 0$ 。
- 通常表示一段时间内某种事件到来次数, 如一个小时内超市顾客数, 热线电话十分钟内接到的电话个数, 产品上的疵点个数, 等等。 λ 是平均数:

$$EX = \lambda$$

- R中有关函数为`ppois(x, lambda)`, `dpois(x, lambda)`, `qpois(p, lambda)`, `rpois(n, lambda)`。

均匀分布

- 设 X 在 (a, b) 取值可能性均等, 称 $X \sim U(a, b)$ 。
- 均匀分布是连续型分布, 密度函数为

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & \text{其它} \end{cases}$$

- R中`runif(n, a, b)`产生 n 个服从 $U(a, b)$ 分布的随机数。

正态分布

- $X \sim N(\mu, \sigma^2)$ 的密度为

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

$$x \in (-\infty, \infty), \mu \in (-\infty, \infty), \sigma \in (0, \infty)$$

- $EX = \mu, \text{Var}(X) = \sigma^2$ 。
- $N(0, 1)$ 称为标准正态分布，密度为

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

分布函数记为 $\Phi(x)$ 。

- $N(\mu, \sigma^2)$ 的分布密度为 $\Phi(\frac{x-\mu}{\sigma})$ 。
- 正态分布常表示误差、测量值等连续型量的分布。
- 正态分布密度是单峰对称的，中心点为 μ ， σ 代表宽度。
- 正态分布的经验规则：

$$P(X \in (\mu - 2\sigma, \mu + 2\sigma)) \approx 95\%$$

$$P(X \in (\mu - 3\sigma, \mu + 3\sigma)) \approx 99.73\%$$

- 见Flash演示。
- `pnorm(x, mean, sd)`、`dnorm`、`qnorm`、`rnorm` 分别为正态分布函数、分布密度、分位数函数、随机数函数，期望为 `mean`，标准差为 `sd`。

指数分布

- 参数为 λ 的指数分布密度为

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- 分布函数为 $F(x; \lambda) = 1 - e^{-\lambda x}$ 。
- $EX = \frac{1}{\lambda}$ 。
- 常用于表示电子产品的寿命，下一事件的到来间隔等，取正值。
- 具有“无记忆性”：

$$P(X > t + s | X > s) = P(X > t)$$

- `pexp`、`dexp`、`qexp`、`rexp` 分别为指数分布函数、分布密度、分位数函数、随机数函数。

威布尔分布

- 威布尔(Weibull)是指数分布的推广,用于各种寿命,不再保持无记忆性。
- 密度为

$$f(x; m, \eta) = \begin{cases} \frac{m}{\eta^m} x^{m-1} \exp\left\{-\left(\frac{x}{\eta}\right)^m\right\} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- 分布函数为 $F(x; m, \eta) = 1 - \exp\left\{-\left(\frac{x}{\eta}\right)^m\right\}$ 。

伽玛分布

- 伽玛分布(Gamma)也是正值随机变量,常用来表示气象中年降水量、最大风速等。
- 密度为

$$f(x; \alpha, \lambda) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- $\alpha = 1$ 时退化为指数分布。

卡方分布

- n 个自由度的卡方(χ^2)密度是 $\Gamma(\frac{n}{2}, \frac{1}{2})$ 分布。
- 密度为

$$f(x; n) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2}, \quad x > 0$$

- 若 $X_1, X_2, \dots, X_n$ iid $N(0,1)$, 则 $X_1^2 + \dots + X_n^2 \sim \chi^2(n)$ 。“iid”表示随机变量相互独立而且服从相同的分布。
- R中有关函数为 `pchisq(x, df)`, `dchisq`, `qchisq`, `rchisq`。

t分布

- 随机变量 X 取值于 $(-\infty, \infty)$, 密度为

$$f(x; n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in (-\infty, \infty)$$

则称 X 服从 n 个自由度的 t 分布, 记为 $t(n)$ 。

- 若 $\xi \sim N(0,1)$, $\eta \sim \chi^2(n)$, ξ 与 η 相互独立, 则

$$\frac{\xi}{\sqrt{\eta/n}} \sim t(n)$$

- 若 $X_1, X_2, \dots, X_n$ iid $N(\mu, \sigma^2)$, 则 $\frac{1}{\sigma^2} \sum_i (X_i - \bar{X})^2 \sim \chi^2(n-1)$,

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

其中 $S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$ 。

- R 中有关函数为 `pt(x, df)`, `dt`, `qt`, `rt`。

F分布

- 称正值随机变量 X 服从 $F(m, n)$ 分布 (m 和 n 为正整数), 若其密度为

$$\frac{1}{B(\frac{m}{2}, \frac{n}{2})} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m-2}{2}} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, x > 0$$

其中 B 为 Beta 函数,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx, \quad \alpha > 0, \beta > 0$$

- 若 $\xi \sim \chi_m^2$, $\eta \sim \chi_n^2$, ξ 与 η 独立, 则 $(\xi/m)/(\eta/n)$ 服从 $F(m, n)$ 。
- R 中有关函数为 `pf(x, df1, df2)`, `df`, `qf`, `rf`。

练习

- 同时掷两只骰子, 令 X 为掷得的点数和, 求 X 的概率分布, 求 EX 和 $\text{Var}(X)$ 。
- 设 $X_1, X_2, \dots, X_n$ iid $N(\mu, \sigma^2)$, μ, σ^2 未知。说明 $(\bar{X} - \mu)/\sigma$, $\sum_i (X_i - \bar{X})^2/\sigma^2$, $(\bar{X} - \mu)/(S/\sqrt{n})$ 的分布。
- 设 T 服从 $t(n)$ 分布, 用 R 表达式表示 T 的双侧 α 分位数, 即 x 使 $P(|T| > x) = 1 - \alpha$ 。

第二章 置信区间和假设检验

§2.1 总体与样本

§2.1.1 统计基本概念

总体

- 数理统计学是研究收集、整理、分析数据得到有意义结论的科学。
- 把要研究的对象全体称为**总体**。简单的总体用一个随机变量 X 表示，比如：随机抽出一块晶圆上的疵点个数；北京市一年级小学男生身高；北京市工作了两年的大学生的工薪。
- 关心的是总体的分布：晶圆上疵点个数等于0, 1, 2, ...的比例；身高在100厘米以下的比例，在100—110厘米的比例等；工资小于1000的比例，1000—2000的比例等。
- 分布可以用分布函数、分布概率函数、分布密度函数刻画，可以用平均值、中位数、标准差、极差等概括。

样本

- 从总体中抽取的有代表性的一组值。
- **简单随机样本**：独立重复试验的结果 $X_1, X_2, \dots, X_n$ 。每个 X_i 的分布与总体分布相同， $X_1, X_2, \dots, X_n$ 相互独立。 n 称为**样本量**。
- 比如，总体 X 为掷一枚硬币的结果， $X = 1$ 表示正面， $X = 0$ 表示反面。重复掷了 n 次的结果 $X_1, X_2, \dots, X_n$ 为总体的一组简单随机样本。
- 总体 X 表示从一批产品中随机抽取一个的结果，次品表示为1，正品表示为0。从这批产品中随机**有放回**地抽取 n 个， $X_1, X_2, \dots, X_n$ ，是 X 的一个简单随机样本。

总体参数和统计量

- 假设总体 X 服从分布 $F(x; \theta)$ ， F 的形式已知，但包含未知的**参数** θ 。从样本 $X_1, X_2, \dots, X_n$ 估计出参数 $\hat{\theta}$ 就可以了解总体分布。
- 从样本 $X_1, X_2, \dots, X_n$ 计算得到的量 $\phi(X_1, X_2, \dots, X_n)$ 叫做**统计量**，统计量的计算中不能用到未知量，即得到样本的观测值以后统计量必须能计算出具体数值。

- 常用的统计量:

$$\bar{X} = \frac{1}{n} \sum_i X_i$$

$$S^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

$\bar{X}$ 称为样本平均值, S 称为样本标准差。

- 在R中, 用`mean(x)`求向量 x 的平均值, `sd(x)`求向量 x 的样本标准差。

§2.1.2 探索性数据分析(EDA)

探索性数据分析(EDA)

- 对一个变量, 我们可以用一些**描述统计**的办法概括其分布情况。
- 对离散变量, 关心其取值集合以及取每个值的次数, 比例。例如: 性别, 取值为" F", " M", 取" F"和" M"的次数和比例。
- 对连续变量, 关心: 取值区间; 位置信息(平均值、中位数); 分散程度信息(标准差, 极差, 四分位间距), 分布密度是否对称或偏斜, 是否有较多离群值。

例: 离散变量

- 19个性别记录的分布情况。

```
x <- cl$sex
print(x)
table(x)
table(x)/length(x)
barplot(table(x))
```

例: 正态分布样本

- 生成了30个 $N(10, 2^2)$ 随机数, 计算其简单统计量, 做直方图、盒形图、正态QQ图:

```
x <- rnorm(30, 10, 2)
summary(x)
hist(x)
boxplot(x)
qqnorm(x);qqline(x)
```

- 盒形图的粗线是中位数，盒子下、上边缘为1/4和3/4分位数，“触须线”向下延伸到最小值，向上延伸到最大值，但长度不超过盒子长度1.5倍，超过部分用散点表示，为离群值。如`boxplot(rt(100,3))`。
- 正态QQ图画出的散点基本围绕一条直线时说明数据来自正态分布总体，否则说明总体不是正态分布。

例：指数分布样本

- 生成30个指数分布 $\text{Exp}(1)$ 随机数，并探索其分布：

```
x <- rexp(30)
summary(x)
hist(x)
boxplot(x)
qqnorm(x);qqline(x)
```

- 分布右偏：分布密度右边尾部较长。

Box-Cox变换

- 密度偏斜的数据模型难以建立，即使画图也有“扎堆儿”现象。Box-Cox变换常常可以把数据分布变得比较对称。
- 对右偏分布，常用对数变换或开平方根使其对称。如：

```
x <- exp(rnorm(30,0,1))
hist(x);locator(1)
y <- log(x)
hist(y);locator(1)

x <- rnorm(30,10,3)^2
hist(x);locator(1)
y <- sqrt(x)
hist(y)
```


- Box-Cox变换包含一系列变换，形式为

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$$

- $\lambda = 0$ 时做对数变换，否则做幂变换，如平方、开平方根、倒数。
- 在R中，用MASS包的**boxcox**函数寻找参数 λ 的值。用法如

```
library(MASS)
boxcox(x ~ 1)
```

函数画出估计参数 λ 用的profile似然，最大值点为应该使用的变换参数，并给出95%的置信限。

§2.2 参数估计

§2.2.1 点估计方法

参数估计方法

- 如果总体服从参数未知的分布 $F(x; \theta)$, 从样本得到参数的估计 $\hat{\theta}$ 就可以了解总体分布。
- 最大似然估计法: 设总体 X 的密度为 $f(x; \theta)$, 则样本 $X_1, X_2, \dots, X_n$ 的联合密度函数为

$$L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

把 $L(X_1, X_2, \dots, X_n; \theta)$ 看成 θ 的函数, 简记为 $L(\theta)$, 称为给定样本 $X_1, X_2, \dots, X_n$ 后参数 θ 的似然函数。若 $\hat{\theta}$ 使 $L(\theta)$ 达到最大值, 就称 $\hat{\theta}$ 为参数 θ 的最大似然估计(MLE)。

- 矩估计法: 用 $\bar{X}$ 估计 EX , 用 $\frac{1}{n} \sum_i (X_i - \bar{X})^2$ 估计 $\text{Var}(X)$, 并据此反解未知参数 θ , 这种估计方法叫做矩估计法。

最大似然估计求解

- $\ln L(\theta)$ 称为对数似然函数, 与似然函数 $L(\theta)$ 有相同的最大值点。 $\ln L(\theta) = \sum_{i=1}^n \ln f(x; \theta)$ 。
- 设 $\theta = (\theta_1, \theta_2, \dots, \theta_m)$, 最大值点应满足似然方程:

$$\begin{aligned} \frac{\partial \ln L}{\partial \theta_1} &= 0 \\ \frac{\partial \ln L}{\partial \theta_2} &= 0 \\ &\dots\dots\dots \\ \frac{\partial \ln L}{\partial \theta_m} &= 0 \end{aligned}$$

- 最好能得到方程的显示解。
- θ 为多维的时候, 有时可以先关于其中一个分量取最大值。
- R中的函数`optim`和`optimize`可以用数值方法求极大值点, `optimize`只能用于一元函数。

例: 正态分布参数估计

- 总体 $X \sim N(\mu, \sigma^2)$, 样本 $X_1, X_2, \dots, X_n$ 。

-

$$f(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

$$\ln f(x; \mu, \sigma^2) = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}(x - \mu)^2$$

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum (X_i - \mu)^2$$

- 写出似然方程后求解得 $\hat{\mu} = \bar{X}$, $\widehat{\sigma^2} = \frac{1}{n} \sum_i (X_i - \bar{X})^2$ 。
- 求正态参数MLE的另一方法是从 $\ln L$ 分步求最大值。从 $\ln L$ 表达式看出对任意 $\sigma > 0$, 要使 $\ln L$ 最大必须 $\sum_i (X_i - \mu)^2$ 最小, 易证明这当且仅当 μ 取 $\bar{X}$ 。
- 取定 $\mu = \hat{\mu} = \bar{X}$ 后, $\ln L(\hat{\mu}, \sigma^2)$ 是 σ^2 的一元函数:

$$\ln L(\bar{X}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum (X_i - \bar{X})^2$$

容易求其最大值, 也解得 $\widehat{\sigma^2} = \frac{1}{n} \sum_i (X_i - \bar{X})^2$ 。

- 例: 给了100个正态分布样本, 估计总体在[5, 15]之间取值的概率。

```
x <- rnorm(100,10,3)
mu <- mean(x)
sig <- sd(x)
p = pnorm(15,mu,sig) - pnorm(5,mu,sig)
cat("Prob in [5,15] = ", p, "\n")
```

- 为求最大似然估计还可以使用数值方法。
- 为了求对数似然函数最大值点, 因 $\sigma^2 > 0$, 所以作变换令 $\delta = \ln(\sigma^2)$, $\sigma^2 = e^\delta$ 。使用-2倍的对数似然函数, 去掉其中的常数值, 似然函数的R函数写法为

```
objf <- function(theta, x){
  mu <- theta[1]
  s2 <- exp(theta[2])
  n <- length(x)
  res <- n*log(s2) + 1/s2*sum((x - mu)^2)
  res
}
```

- 用`optim`函数来求极小值点。下面是一个模拟演示:

```
sim <- function(n=30){
  mu0 <- 20
  sigma0 <- 2
  x <- rnorm(n, mu0, sigma0)

  theta0 <- c(0,0)
  ores <- optim(theta0, objf, x=x)
  print(ores)
  theta <- ores$par
  mu <- theta[1]
  sigma <- exp(0.5*theta[2])
  cat("mu: ", mu, " ==> ", mu0, "\n")
  cat("sigma: ", sigma, " ==> ", sigma0, "\n")
}
```

- 函数`optimize`用来求一元函数最小值点。例如，在正态MLE中，对数似然函数中取定 $\mu = \bar{X}$ ，乘以 -2 ，去掉常数项，函数变成:

$$h(\sigma^2) = \ln(\sigma^2) + \frac{1}{\sigma^2} \sum (X_i - \bar{X})^2$$

- 为求 σ^2 的极小值点，可用如以下程序:

```
sim1 <- function(n=30){
  mu0 <- 20
  sigma0 <- 2
  x <- rnorm(n, mu0, sigma0)

  mu <- mean(x)
  ss <- sum((x - mu)^2)/length(x)
  objf <- function(delta,ss) log(delta) + 1/delta*ss
  ores <- optimize(objf, lower=0.0001,
                    upper=1000, ss=ss)
  delta <- ores$minimum
}
```

```

sigma <- sqrt(delta)

print(ores)
cat("mu: ", mu, " ==> ", mu0, "\n")
cat("sigma: ", sigma, " ==> ", sigma0, "\n")
}

```

§2.2.2 置信区间

标准误差

- 从样本 $X_1, \dots, X_n$ 得到未知参数 θ 的估计 $\hat{\theta}$, 称为**点估计**。只有点估计无法了解其精度。
- $\hat{\theta}$ 也是随机变量, 其分布称为**抽样分布**。 $\hat{\theta}$ 的抽样分布的标准差大小可以衡量估计精度。设 $SE(\hat{\theta})$ 是 $\hat{\theta}$ 的抽样分布的标准差的估计值, 称为 $\hat{\theta}$ 的**标准误差**(standard error)。与总体的标准差(standard deviation)区别。
- 增加对估计精度了解的另一工具是置信区间。

置信区间

- 为了估计未知参数 θ 并了解估计的可靠程度, 可以给出 θ 的一个最可能的范围。(如天气预报可以预报为“小到中雨”)。
- 设统计量 $L(X_1, \dots, X_n)$ 和 $U(X_1, \dots, X_n)$ 使得

$$P(\theta \in [L(X_1, \dots, X_n), U(X_1, \dots, X_n)]) \geq 1 - \alpha$$

其中 $0 < 1 - \alpha < 1$, 则称 $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$ 为未知参数 θ 的**置信度**为 $1 - \alpha$ 的**置信区间**。 $L(X_1, \dots, X_n)$ 称为**置信下限**, $U(X_1, \dots, X_n)$ 称为**置信上限**。

- 置信区间可以是单边的, 如 $[L(X_1, \dots, X_n), +\infty)$, 或 $(-\infty, U(X_1, \dots, X_n)]$ 。

正态分布的置信区间(1)

- 设 $X \sim N(\mu, \sigma^2)$, $X_1, \dots, X_n$ 为样本。
- 若 μ 未知, σ^2 已知, 则由 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ 服从标准正态分布可知

$$P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

其中 z_p 表示标准正态分布的 p 分位数, 在 R 中用函数 `qnorm(p)` 计算。 μ 的置信区间为 $\bar{X} \pm z_{1-\frac{\alpha}{2}} \sigma/\sqrt{n}$ 。

- $1 - \alpha = 0.95$ 时 μ 的置信区间为 $\bar{X} \pm 1.96\sigma/\sqrt{n}$ 。
- 例如, $n = 9$, $\sigma = 2$, $\bar{X} = 10$, 则95%置信区间为 $10 \pm 1.96 \times 2/\sqrt{9} = [8.69, 11.31]$ 。

正态分布的置信区间(2)

- 若 μ 和 σ^2 均未知, 有

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

设 λ 为 $t(n-1)$ 的 $1 - \frac{\alpha}{2}$ 分位数, 则 $P(|T| \leq \lambda) = 1 - \alpha$, μ 的置信度 $1 - \alpha$ 的置信区间为 $\bar{X} \pm \lambda S/\sqrt{n}$ 。

- 其中 S 为样本标准差。
- R中用`qt(p,n)`计算自由度为 n 的 t 分布的 p 分位数。
- 例如 $n = 9$, $\bar{X} = 10$, $S = 2$, $1 - \alpha = 0.95$, λ 可以用R函数调用`qt(1-0.05/2, 9-1)`求得为2.306, 于是 μ 的95%置信区间为 $10 \pm 2.306 \times 2/\sqrt{9} = [8.46, 11.54]$ 。
- 在R中用`mean(x)`计算平均值 $\bar{X}$, 用`sd(x)`计算样本标准差 S 。
- 在R中, 若 x 向量中为样本, 可以用`t.test(x, conf.level=置信度)`来计算方差未知时均值的置信区间。
- 例:

```
x <- c(11.67, 9.29, 10.45, 9.01, 12.67,
       16.24, 11.64, 7.73, 12.23)
t.test(x, conf.level=0.95)
```

- 得到 μ 的95%置信区间为 $[9.29, 13.14]$ 。

置信区间的频率解释

- 关于置信度95%的置信区间, 置信度的解释是: 同样的问题实际中反复遇到, 使用同样方法求得置信区间, 多次使用后参数真值在置信区间内的比例约为95%。
- 如果我们已经计算出了置信区间的值, 比如 $[9.29, 13.14]$, 不能使用“ μ 有95%的概率落在区间内”这样的说法, 因为未知的参数是固定不变的, 或者在区间内, 或者不在区间内。但可以不严格的说“我们有95%的把握保证 μ 在区间内”。
- Flash演示: 置信区间的频率解释。

正态分布的置信区间(3)

- 考虑 μ 和 σ^2 未知时 σ^2 的置信区间。
- $\frac{1}{\sigma^2} \sum (X_i - \bar{X})^2 \sim \chi^2(n-1)$ 。
- 取 λ_1 为 $\chi^2(n-1)$ 的 $\frac{\alpha}{2}$ 分位数, λ_2 为 $\chi^2(n-1)$ 的 $1 - \frac{\alpha}{2}$ 分位数, 则

$$P(\lambda_1 \leq \frac{1}{\sigma^2} \sum (X_i - \bar{X})^2 \leq \lambda_2) = 1 - \alpha$$

于是 σ^2 的 $1 - \alpha$ 置信区间为

$$[\sum (X_i - \bar{X})^2 / \lambda_2, \sum (X_i - \bar{X})^2 / \lambda_1]$$

- σ 的置信区间只要开平方根。
- 在R中用`qchisq(p,n)`求 $\chi^2(n)$ 的 p 分位数。

比例的置信区间

- 比例的点估计很容易, 比如抽查了100人, 其中30人喜欢看电影, 记 p 为喜欢看电影的人的比例, 则 p 的点估计为 $\hat{p} = 30/100 = 30\%$ 。
- 当样本量 n 较大时, $\hat{p}$ 有近似正态分布:

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}} \sim N(0, 1)$$

取正态分布的双侧 α 分位数(即 $1 - \frac{\alpha}{2}$ 分位数) λ , 可得 p 的近似 $1 - \alpha$ 置信区间为 $\hat{p} \pm \lambda \sqrt{\hat{p}(1 - \hat{p})/n}$ 。如置信度95%的置信区间为 $\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/n}$ 。

- 在R中用`prop.test`来估计比例的置信区间。
- 例如, 在 $n = 100$ 个人中, $x = 30$ 个人喜欢看电影。喜欢看电影的人的比例 p 的置信度为95%的置信区间可以如下计算:

```
prop.test(30, 100, conf.level=0.95)
```

结果为[21%, 40%]。

§2.3 假设检验

§2.3.1 假设检验的概念

假设检验的意义

- 例: 某产品合格率达99%则符合要求。在一批产品出厂前抽检100个, 废品数少于多少合格?
- 比较同龄的男、女生身高, 抽检男、女生各10人, 若男生平均身高150, 女生身高155, 男女生平均身高是否有**显著**差异? (思考: 150和150.01的比较; 150和160的比较)
- 关键是需要有客观的比较标准, 称为临界值。
- 从样本判断总体是可能有错误的。

零假设和对立假设

- 对总体研究某个假设是否成立, 这个假设称为**零假设** H_0 。我们关心的零假设的反面叫做**对立假设** H_a 。
- 比如, 比例是否达到0.99, 检验问题为

$$H_0 : p \geq 0.99 \longleftrightarrow H_a : p < 0.99$$

- 比如, 男、女生平均身高是否相等, 检验问题为

$$H_0 : \mu_{\text{男}} = \mu_{\text{女}} \longleftrightarrow H_a : \mu_{\text{男}} \neq \mu_{\text{女}}$$

检验统计量和p值

- 假设检验是“带概率的反证法”。只考虑零假设成立时的情况, 从样本计算一个和要检验的参数有关的统计量, 叫做**检验统计量**, 检验统计量的分布在 H_0 下已知。在 H_0 下, 如果发现统计量取值偏向于 H_a 发生的时候, 就否定 H_0 。
- 比如, 在比较男女生平均身高时, 以平均身高的差 d 为统计量。如果发现 $|d| = 10$, 这在 H_0 下发生可能性很小, 就否定 H_0 。
- 可以计算一个**p值**用来衡量 H_0 下统计量取值的偏僻程度, p值越小, 说明 H_0 越不可信。一般预先取定一个**检验水平** α , 如0.05, 0.10, 0.01等, 当且仅当p值小于 α 时否定 H_0 。

两类错误

- 假设检验从样本推断总体，可能发生两类错误：

		检验结果	
		H_0	H_a
真实情况	H_0	正确	第I类错误
	H_a	第II类错误	正确

- 第一类错误控制在 α 以下。
- 第二类错误没有控制，只能通过设计优良的检验方法和增大样本量来减少。
- 所以承认 H_0 的结论是意义不大的。
- 为了得到可靠的结论，尽可能把要做的结论当作对立假设。
- 作出某结论出错时后果严重，应尽可能把此结论作为对立假设。（如：审判中的无罪推断原则）

§2.3.2 单正态总体参数的检验

单正态总体参数的检验

- 设总体 X 服从 $N(\mu, \sigma^2)$ 分布， $X_1, X_2, \dots, X_n$ 为样本， $\bar{X}$ 为样本均值， S^2 为样本方差。考虑：
- 方差已知时对 μ 的检验，包括双侧、右侧、左侧。
- 方差未知时对 μ 的检验，包括双侧、右侧、左侧。
- 关于方差 σ^2 的检验，包括双侧、右侧、左侧。

方差已知时的Z检验

- 若方差 $\sigma^2 = \sigma_0^2$ 已知，则统计量

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}} \sim N(0, 1)$$

当 $\mu = \mu_0$ 成立。

- 均值的双侧检验：

$$H_0 : \mu = \mu_0 \longleftrightarrow H_a : \mu \neq \mu_0$$

设统计量 Z 的值为 a ，则检验的p值为

$$p\text{值} = P(|N(0, 1)| > |a|) = 2(1 - \Phi(|a|))$$

其中 $\Phi(x)$ 为标准正态分布函数。

- 注意在R中用`pnorm(x)`计算 $\Phi(x)$ 。

- 均值的右侧检验:

$$H_0 : \mu = \mu_0 \longleftrightarrow H_a : \mu > \mu_0$$

设统计量 Z 的值为 a , 则检验的p值为

$$\text{p值} = P(N(0, 1) > a) = 1 - \Phi(a)$$

- 均值的左侧检验:

$$H_0 : \mu = \mu_0 \longleftrightarrow H_a : \mu < \mu_0$$

设统计量 Z 的值为 a , 则检验的p值为

$$\text{p值} = P(N(0, 1) < a) = \Phi(a)$$

例子

- 微波炉辐射量指标服从 $N(\mu, 0.1^2)$, 要求 μ 不超过0.12。设抽检25台, 样本平均值 $\bar{X} = 0.13$, 产品辐射量是否超标?
- 假设检验问题为

$$H_0 : \mu = 0.12 \longleftrightarrow H_a : \mu > 0.12$$

- $Z = (0.13 - 0.12)/(0.1/\sqrt{25}) = 0.5$ 。
- p值为 $1 - \text{pnorm}(0.5) = 0.31$, 在0.05水平下不否定 H_0 , 认为产品辐射量没有显著超标。

方差未知时的t检验

- 若正态总体参数 μ 和 σ^2 都未知, 统计量

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1)$$

当 $X \sim N(\mu_0, \sigma)$ 时(σ^2 未知)。

- 均值的双侧检验:

$$H_0 : \mu = \mu_0 \longleftrightarrow H_a : \mu \neq \mu_0$$

设统计量 T 的值为 a , 则检验的p值为

$$\text{p值} = P(|t(n-1)| > |a|) = 2(1 - F(|a|, n-1))$$

其中 $F(x, n)$ 为 $t(n)$ 的分布函数。

- 在R中用`pt(x,n)`计算 $F(x,n)$ 。R函数`t.test`直接计算t检验。

- 均值的右侧检验:

$$H_0: \mu = \mu_0 \longleftrightarrow H_a: \mu > \mu_0$$

设统计量 T 的值为 a , 则检验的p值为

$$\text{p值} = P(t(n-1) > a) = 1 - F(a, n-1)$$

- 均值的左侧检验:

$$H_0: \mu = \mu_0 \longleftrightarrow H_a: \mu < \mu_0$$

设统计量 T 的值为 a , 则检验的p值为

$$\text{p值} = P(t(n-1) < a) = F(a, n-1)$$

例子

- 例如, 食盐包装机把食盐包装为500克每袋, 随机抽检了9袋食盐重量, 结果为:

490, 506, 508, 502, 498, 511, 510, 515, 512

- 为检验包装机是否正常工作, 进行检验

$$H_0: \mu = 500 \longleftrightarrow H_a: \mu \neq 500$$

- 计算程序为:

```
x <- c(490, 506, 508, 502,
       498, 511, 510, 515, 512)
t.test(x, mu=500, alternative="two-sided")
```

得p值0.059, 在0.05水平下不否定零假设, 可以认为包装机工作正常。

- 做单侧检验时`alternative="greater"`或`"less"`。

方差的检验

- 总体方差和标准差代表了测量精度、加工精度等。
- 在 μ 和 σ^2 都未知时, 统计量

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

若 $\sigma^2 = \sigma_0^2$ 。

- 方差的双侧检验:

$$H_0: \sigma^2 = \sigma_0^2 \longleftrightarrow H_a: \sigma^2 \neq \sigma_0^2$$

设统计量 χ^2 的值为 a , $\chi^2(n-1)$ 的分布函数为 $F(x)$, 则检验的p值为

$$\text{p值} = 2 \min\{F(a), 1 - F(a)\}$$

- 方差的右侧检验:

$$H_0: \sigma^2 = \sigma_0^2 \longleftrightarrow H_a: \sigma^2 > \sigma_0^2$$

设统计量 χ^2 的值为 a , $\chi^2(n-1)$ 的分布函数为 $F(x)$, 则检验的p值为

$$\text{p值} = 1 - F(a)$$

- 方差的左侧检验:

$$H_0: \sigma^2 = \sigma_0^2 \longleftrightarrow H_a: \sigma^2 < \sigma_0^2$$

设统计量 χ^2 的值为 a , $\chi^2(n-1)$ 的分布函数为 $F(x)$, 则检验的p值为

$$\text{p值} = F(a)$$

- R中自由度 n 的 χ^2 分布函数为`pchisq(x,n)`。

例子

- 某种保险丝要求强电流下熔化时间方差不超过80。抽检了10根，熔化时间为:

42, 65, 75, 78, 59, 71, 57, 68, 54, 55

- 假设检验:

$$\sigma^2 = 80 \longleftrightarrow \sigma^2 > 80$$

- 计算程序:

```
x <- c(42,65,75,78,59,71,57,68,54,55)
n <- length(x)
chi2 <- (n-1)*var(x)/80
p <- 1-pchisq(chi2,n-1)
cat("Chi-squared = ", chi2, " p-value = ", p, "\n")
```

- 得p值为0.13, 超过0.05, 认为熔化时间的方差不超过80。

§2.3.3 两正态总体参数的检验

两正态总体参数的检验

- 设总体 X 与 Y 独立, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 分别有样本 $X_1, \dots, X_{n_1}$ 和 $Y_1, \dots, Y_{n_2}$, 分别有样本统计量 $\bar{X}$, S_1^2 和 $\bar{Y}$, S_2^2 。
- 考虑假定 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知情况下 μ_1 和 μ_2 的比较: 双侧、右侧、左侧。
- 考虑方差的比较: 双侧、右侧、左侧。

两样本t检验

- 设 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知。在 $\mu_1 = \mu_2$ 时统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S_{\text{pool}}} \sim t(n_1 + n_2 - 2)$$

其中

$$S_{\text{pool}}^2 = \frac{1}{n_1 + n_2 - 2} \left[\sum_i (X_i - \bar{X})^2 + \sum_j (Y_j - \bar{Y})^2 \right]$$

- 两样本均值的双侧检验:

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_a: \mu_1 \neq \mu_2$$

设统计量 T 的值为 a , 则检验的p值为

$$\text{p值} = P(|t(n_1 + n_2 - 2)| > |a|) = 2(1 - F(|a|))$$

其中 $F(x)$ 为 $t(n_1 + n_2 - 2)$ 的分布函数。

- 在R中用`pt(x, n)`计算 $t(n)$ 的分布函数 $F(x)$ 。R函数`t.test`直接计算两样本t检验。
- 两样本均值的右侧检验:

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_a: \mu_1 > \mu_2$$

设统计量 T 的值为 a , 则检验的p值为

$$\text{p值} = P(t(n_1 + n_2 - 2) > a) = 1 - F(a)$$

其中 $F(x)$ 为 $t(n_1 + n_2 - 2)$ 的分布函数。

- 两样本均值的左侧检验:

$$H_0: \mu_1 = \mu_2 \longleftrightarrow H_a: \mu_1 > \mu_2$$

设统计量 T 的值为 a , 则检验的 p 值为

$$p\text{值} = P(t(n_1 + n_2 - 2) < a) = F(a)$$

其中 $F(x)$ 为 $t(n_1 + n_2 - 2)$ 的分布函数。

- 设向量 x 和 y 分别保存了 X 和 Y 的样本, 用`t.test(x,y, alternative="two-sided")`进行双侧两样本 t 检验。 `alternative="greater"`或`"less"`做右侧或左侧检验。
- `t.test(x,y, var.equal=FALSE)`可以放松对两总体方差相等的限制。
- `boxplot(list(x=x,y=y))`可以画 x 和 y 的并排盒形图, 直观比较其位置和分散程度。

例子

- 设甲、乙两台机床分别加工某种轴承, 轴承直径分别服从正态分布, 假定两总体方差相等。各随机抽取若干样品得如下数据:

甲:20.5, 19.8, 19.7, 20.4, 20.1, 20.0, 19.0, 19.9

乙:20.7, 19.8, 19.5, 20.8, 20.4, 19.6, 20.2

- 要检验两台机床加工的直径有无显著差异。假设检验:

$$\mu_1 = \mu_2 \longleftrightarrow \mu_1 \neq \mu_2$$

- 计算程序:

```
x <- c(20.5, 19.8, 19.7, 20.4,
       20.1, 20.0, 19.0, 19.9)
y <- c(20.7, 19.8, 19.5, 20.8,
       20.4, 19.6, 20.2)
t.test(x, y)
```

- 得 p 值为0.41, 超过0.05, 认为两台机床加工的轴承直径没有显著差异。

两独立总体方差比较

- 考虑

$$H_0: \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_a: \sigma_1^2 \neq \sigma_2^2$$

- 在 H_0 下, 统计量

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

- R中`var.test(x,y)`可对方差进行双侧检验。`var.test(x,y, alternative="greater")`或`"less"`可进行右侧或左侧检验。

例子

- 在上面的两台机床比较中, 比较加工精度, 即方差有无显著差异。
- 假设为

$$\sigma_1^2 = \sigma_2^2 \longleftrightarrow \sigma_1^2 \neq \sigma_2^2$$

- R程序:

```
var.test(x, y)
```

- 结果p值为0.76, 超过 $\alpha = 0.05$, 不否定零假设, 认为两台机床加工精度没有显著差异。

§2.3.4 成对t检验

成对t检验

- 设 X 和 Y 是同一个个体的两个属性, 比如服药前血压和服药后血压。 X 和 Y 不独立。
- 要比较 X 和 Y 的均值, 只要计算 $Z = X - Y$, 设 Z 服从 $N(\mu, \sigma^2)$, 检验 Z 的均值 μ 。这是单总体均值检验问题。
- 在R中, 用`t.test(x, y, paired=TRUE)`进行成对t检验。加`alternative="greater"`或`"less"`可以进行右侧或左侧检验。

例子

- 为考察针织品在不同温度下漂白对强度的影响，取了8件样品，然后每件样品分为两半，分别用70°C和80°C的水漂白，结果如下：

样品号	1	2	3	4
70°C下强度	20.5	18.8	19.8	20.9
80°C下强度	17.7	20.3	20.0	18.8
样品号	5	6	7	8
70°C下强度	21.5	19.5	21.0	21.2
80°C下强度	19.0	20.1	20.0	19.1

- 检验为

$$\mu_X = \mu_Y \longleftrightarrow \mu_X \neq \mu_Y$$

- R程序为

```
x <- c(20.5, 18.8, 19.8, 20.9,
       21.5, 19.5, 21.0, 21.2)
y <- c(17.7, 20.3, 20.0, 18.8,
       19.0, 20.1, 20.0, 19.1)
t.test(x, y, paired=TRUE)
```

- p值为0.11, 超过 $\alpha = 0.05$, 认为两种温度下漂白对强度没有显著影响。

§2.3.5 正态性检验

正态性检验

- 以上单总体、两总体、成对比较都是基于正态分布的，如何获知某样本 $X_1, X_2, \dots, X_n$ 是否来自正态分布总体呢？
- 正态性检验:

总体 X 服从正态分布 $\longleftrightarrow$ 总体 X 不服从正态分布

- 拒绝零假设时结论较可靠，但接受零假设的结论也可以用。
- 在R中，用`shapiro.test(x)`做Shapiro-Wilk正态性检验。用盒形图、QQ图、直方图直观查看分布。

例子

- 例子:

```
x <- c(20.5, 19.8, 19.7, 20.4,
       20.1, 20.0, 19.0, 19.9)
hist(x);locator(1)
boxplot(x);locator(1)
qqnorm(x);qqline(x)
shapiro.test(x)
```

- 结果p值为0.52, 大于 $\alpha = 0.05$, 可以认为样本来自正态总体。

§2.3.6 比例的检验

单样本比例的检验

- 设总体 X 为两点分布 $b(1, p)$, p 为成功概率, 样本为 $X_1, X_2, \dots, X_n$ 。实际上只关心 $\sum_i X_i$, 即成功次数。
- p 是总体中某种特性的比例, 比如“喜欢看电影的人的比例”。
- 考虑假设检验:

$$H_0 : p = p_0 \longleftrightarrow H_a : p \neq p_0$$

$$H_0 : p = p_0 \longleftrightarrow H_a : p > p_0$$

$$H_0 : p = p_0 \longleftrightarrow H_a : p < p_0$$

- 在R中, 当样本量 n 不大(如 $n < 30$)时, 可以使用`binom.test(nsucc, ntest, p0)`进行检验, 加`alternative="greater"`或`"less"`可以进行右侧或左侧检验。
- 样本量较大时, 在 H_0 下统计量

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

近似服从标准正态分布。在R中用`prop.test(nsucc, ntest, p0)`进行检验, 加`alternative="greater"`或`"less"`可以进行右侧或左侧检验。

例子

- 某产品的优质品率一直保持在40%，近期抽查了12件，其中优质品5件，问在0.05水平能否认为优质品率保持不变？

- 假设:

$$p = 0.4 \longleftrightarrow p \neq 0.4$$

- R程序:

```
binom.test(5, 12, 0.4)
prop.test(5, 12, 0.4)
```

- 精确检验的p值等于1，不否定零假设，认为优质品率保持不变。此问题样本量小，不应使用近似的prop.test。

例子

- 某大学随机调查120名男同学，发现有35人喜欢看武侠小说。能否认为该大学有四分之一男同学喜欢看武侠小说？

- 假设:

$$p = 0.25 \longleftrightarrow p \neq 0.25$$

- R程序:

```
prop.test(35, 120, 0.25)
```

- 检验结果p值为0.34, 超过 $\alpha = 0.05$, 不否定零假设, 可以认为该大学男同学中喜欢看武侠小说的比例与四分之一没有显著差异。

两样本比例比较

- 考虑两个总体比例的比较, 比如男生和女生中喜欢看电影的比例的比较。
- $X \sim B(1, p_1)$, $Y \sim B(1, p_2)$, 样本 $X_1, \dots, X_{n_1}$ 和 $Y_1, \dots, Y_{n_2}$ 。对样本只需要试验数和成功数。

- 考虑假设检验:

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 \neq p_2$$

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 > p_2$$

$$H_0 : p_1 = p_2 \longleftrightarrow H_a : p_1 < p_2$$

- 在 n_1 和 n_2 较大时, 在 H_0 下统计量

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})\hat{p}(1-\hat{p})}}$$

近似服从标准正态分布。其中 $\hat{p}_1$ 为 X 中成功比例, $\hat{p}_2$ 为 Y 中成功比例, $\hat{p}$ 为 X 和 Y 样本合并在一起的成功比例。

- 在R中, 用`prop.test(c(nsucc1, nsucc2), c(n1,n2))`来检验两个比例是否相等。加`alternative="greater"`或`"less"`可以进行右侧或左侧检验。

例子

- 随机抽查了102个男生, 135个女生, 有23个男生和25个女生家中有计算机。在0.05水平下检验男女生家庭有计算机的比例是否相同。
- 假设检验:

$$p_1 = p_2 \longleftrightarrow p_1 \neq p_2$$

- R程序:

```
prop.test(c(23,25), c(102,135))
```

- 结果p值为0.55, 超过0.05, 男女生家庭有计算机的比例没有显著差异。

§2.4 练习

练习

- 写出样本方差的计算公式。
- 在R中, 令`x <- rgamma(30,2,1)`, 简述 x 的分布情况。
- 令 x 如下定义:

```
x <- c(11.67, 9.29, 10.45, 9.01, 12.67,  
       16.24, 11.64, 7.73, 12.23)
```

求其均值、标准差，假定总体为正态分布，估计总体在[6,10]内取值的概率；求 μ 和 σ 的90%置信区间。

- 某工厂需要订购的某零部件合格率 p 达到95%。去供货商处提货的检验员从产品中抽检 n 个检验是否 $p \geq 0.95$ 。应如何选取零假设和对立假设？
- 编写用于进行Z检验的函数，输入为保存了样本的向量 x ，用于比较的均值 μ_0 ，已知的标准差 σ_0 ，以及选择双侧、右侧、左侧的 $alternative$ ，输出检验 p 值。
- 假设对随机选取的8位病人使用A药物，对随机选取的6位病人使用B药物，过一段时间后测量每位病人体细胞中的药物含量，A、B药物的测量数据见下表：

A药物:	1.23	1.42	1.41	1.62	1.55	1.51
	1.60	1.76				
B药物:	1.76	1.41	1.87	1.49	1.67	1.81

在0.10水平下检验B药物的含量是否高于A药物的含量。

- 有若干人参加了一个减肥锻炼，在一年后测量了他们的身体脂肪含量，结果如下(身体脂肪含量的百分数)：
- | | | | | | | | |
|------|------|----|----|----|------|------|----|
| 男性组: | 13.3 | 19 | 20 | 8 | 18 | 22 | 20 |
| | 31 | 21 | 12 | 16 | 12 | 24 | |
| 女性组: | 22 | 26 | 16 | 12 | 21.7 | 23.2 | 21 |
| | 28 | 30 | 23 | | | | |

比较这些人中男性和女性的身体脂肪含量有无显著差异(检验水平0.05。)

- 下表为某基础统计课程两次考试的学生成绩。两次考试考同样的知识。试比较这两次考试难易程度有无显著差异(检验水平0.05)。

学号	1	2	3	4	5	6	7	8	9	10
第一次	93	88	89	88	67	89	83	94	89	55
第二次	98	74	67	92	83	90	74	97	96	81
学号	11	12	13	14	15	16	17	18	19	20
第一次	88	91	85	70	90	90	94	67	87	83
第二次	83	94	89	78	96	93	81	81	93	91

- 下面是一组草原隼的鸟巢高度的数据, 试检验其分布是否正态。

15	3.5	3.5	7	1	7
5.75	27	15	8	4.75	7.5
4.25	6.25	5.75	5	8.5	9
6.25	5.5	4	7.5	8.75	6.5
4	5.25	3	12	3.75	4.75
6.25	3.25	2.5			

第三章 方差分析

第三讲内容概要

- 两组比较进阶。
- 多组比较—单因素方差分析
- 两因素方差分析
- 协方差分析
- 正交设计

§3.1 两组比较进阶

两组比较进阶

- 两个正态总体均值的比较，差的比较
- 两个总体位置比较的非参数方法
- 成对比较的非参数方法
- 分布一致性检验：卡方检验
- 两个变量独立的卡方检验
- 两个比例的比较：卡方检验和Fisher精确检验。

正态总体的检验

- 两个独立的正态总体均值的比较：见第二讲第3.3小节。在R中用`t.test`检验。
- 两个正态独立的总体方差的比较：见第二讲第3.3小节。在R中用`var.test`检验。
- 成对比较，假设差值正态：见第二讲第3.4小节。在R中用`t.test`检验。

§3.1.1 位置检验的非参数方法

独立两总体位置比较的Wilcoxon秩和检验

- 非参数方法是不需要假定总体理论分布的方法，使用范围广，但针对性差，在参数方法与非参数方法都适用的情形，非参数方法效率较差。
- “秩”就是名次，由低到高排列。为比较独立的两个组 X 大小有无显著差异，只需比较两组样本的平均名次有无显著差异。
- 在R中用`wilcox.test(x, y)`做Wilcoxon秩和检验，比较两组大小有无显著差异。

Wilcoxon秩和检验例子

- 考虑两台机床加工同一种零件，考察两台加工的直径大小有无显著差异。
- 输入数据及检验的程序为:

```
x <- c(20.5, 19.8, 19.7, 20.4,
       20.1, 20.0, 19.0, 19.9)
y <- c(20.7, 19.8, 19.5, 20.8,
       20.4, 19.6, 20.2)
wilcox.test(x, y)
```

- p值为0.60，不显著，认为两台机床加工的直径大小没有显著差异。
- 两样本t检验的p值为0.41，小于非参数方法的p值，说明非参数方法效率较低。

成对比较的符号检验

- 设 (X, Y) 是同一个体的两个测量值，一般是相关的随机变量。靠检验 $Z = X - Y$ 来比较两个变量。
- 如果 Z 服从正态分布，只要用成对t检验即可(本质上是单总体的均值检验)。
- 否则，设 Z 的中位数为 M ，考虑假设检验问题

$$H_0 : M = 0 \longleftrightarrow H_a : M \neq 0$$

- 拒绝 H_0 时认为两个变量大小有显著差异。
- 也可以提出单侧检验问题。
- 设 Z 的样本为 $Z_1, Z_2, \dots, Z_n$ 。符号检验的想法是比较正号和负号的多少，如果正号比负号多出很多，则 X 较大，如果负号比正号多出很多，则 Y 较大，如果正负号个数相差无几，则 X 与 Y 大小没有显著差异。
- 以“取正号”作为成功，“取符号”作为失败，可以把 Z 的正负号看成两点分布，来检验成功概率是否等于 $\frac{1}{2}$ 。
- 在R中，用`binom.test(sum(z>0), sum(z !=0), p=0.5)`来检验。如果要做单侧检验，加上`alternative="greater"` 或`alternative="less"`选项。

符号检验例子

- 考虑纺织品在70°C和80°C的强度比较的例子。我们用符号检验重新做双侧检验：

```
x <- c(20.5, 18.8, 19.8, 20.9,
       21.5, 19.5, 21.0, 21.2)
y <- c(17.7, 20.3, 20.0, 18.8,
       19.0, 20.1, 20.0, 19.1)
t.test(x, y, paired=TRUE)

z <- x - y
binom.test(sum(z>0), sum(z != 0), p=0.5)
```

- p值为0.73，不拒绝零假设，认为两种温度下强度没有显著差异。
- 注意成对t检验的p值为0.11，这说明两种检验方法相比符号检验效率较低。

成对比较的符号秩检验

- 为了比较相关的 (X, Y) 的大小，使用类似秩和检验的想法。
- 令 $Z = X - Y$ ，样本 $Z_1, \dots, Z_n$ ，把 Z 的样本分为正号组和负号组，计算 $|Z|$ 的秩，比较正号组和负号组各自的平均秩。
- 平均秩有显著差异则 X 和 Y 的大小有显著差异。
- 符号秩检验利用了 Z 的绝对值大小而符号检验仅利用 Z 的符号，所以符号秩检验比符号检验有效。
- 在R中用`wilcoxon.test(x, y, paired=TRUE)`检验。

符号秩检验例子

- 仍考虑纺织品在70°C和80°C的强度比较的例子。
- 做符号秩检验：

```
wilcox.test(x, y, paired=TRUE)
```

- p值为0.14，比成对t检验的p值略大，但明显小于符号检验的p值。

§3.1.2 卡方检验

单总体分布检验

- 拟合优度卡方检验用来检验样本是否来自某个总体。
- 例如，掷100次骰子，各点数的次数为：

点数	1	2	3	4	5	6
次数	18	13	17	21	15	16

- 欲检验骰子是否正常。
- 模型：令 X 为掷一次的点数，则 X 为在 $\{1, 2, 3, 4, 5, 6\}$ 内取值的随机变量。问题化为假设检验

$$H_0 : P(X = 1) = P(X = 2) = \cdots = P(X = 6) \\ \longleftrightarrow H_a : \text{概率不全相等}$$

- 设 x 为各个结果的次数， p 为各个结果在零假设下的比例，用R检验的程序为`chisq.test(x, p=p)`。
- 例：

```
x <- c(18, 13, 17, 21, 15, 16)
p <- rep(1/6, 6)
chisq.test(x, p)
```

- 结果 p 值为0.82，不拒绝零假设，认为骰子没有问题。

单个比例的卡方检验

- 单个比例的假设检验也可以用卡方检验法。设总体 X 为表示某事件成功与失败的随机变量，成功概率为 p 。要检验

$$H_0 : p = p_0 \longleftrightarrow H_a : p \neq p_0$$

设试验了 n 次，成功了 s 次，只要用R程序`chisq.test(c(s, n-s), c(p0, 1-p0))`。

两总体独立性检验

- 设 X, Y 为研究的两个变量，要检验

$$H_0 : X \text{与} Y \text{相互独立} \longleftrightarrow H_a : \text{不独立}$$

- 设 X 取值集合为 $\{a_1, \dots, a_r\}$, Y 取值集合为 $\{b_1, \dots, b_c\}$, (X, Y) 的一个样本量为 n 的样本中，组合 (a_i, b_j) 发生的次数为 n_{ij} 。
- 数据可以列表为

	b_1	b_2	$\dots$	b_c	行和
a_1	n_{11}	n_{12}	$\dots$	n_{1c}	$n_{1\cdot}$
a_2	n_{21}	n_{22}	$\dots$	n_{2c}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_r	n_{r1}	n_{r2}	$\dots$	n_{rc}	$n_{r\cdot}$
列和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot c}$	n

- 设上表输入为一个矩阵`tab`，以 X 的各取值为行名，以 Y 的各取值为列名，元素是各 n_{ij} 的值。R中检验独立性用`chisq.test(tab)`。

独立性检验例子

- 例如: 研究吸烟与肺癌的关系。调查了63个肺癌患者和43个健康人，得下表:

	吸烟	不吸烟
病人	60	3
健康人	32	11

- 检验吸烟与否和患肺癌是否独立:

```
tab <- matrix(c(60, 3,
                32, 11),
              nrow=2, ncol=2, byrow=TRUE,
              dimnames=list(c("病人", "健康人"),
                            c("吸烟", "不吸烟")))
chisq.test(tab)
```

- p值为0.005，结果显著，所以两者不独立。

- 注意，不能得出吸烟导致肺癌的结论。
- 这个检验也可以看成是病人与健康人中吸烟比例的比较，所以独立性卡方检验也可以用于比较两个总体中某种属性的比例。
- 一般地，以行作为两个组，以列作为事件发生与否，可以用独立性卡方检验方法检验两个组事件发生概率是否相等。

Fisher精确检验

- 假设独立性检验中行和列和固定，非随机。可以用`fisher.test(tab)`来检验行和列的独立性。
- 如吸烟与肺癌的例子。结果p值为0.003。
- 也可以看成是病人与健康人中吸烟比例的比较，所以Fisher精确检验也可以用于比较两个总体中某种属性的比例。

§3.2 单因素方差分析

单因素方差分析

- 方差分析介绍
- 单因素方差分析
- 多重比较
- 方差齐性检验
- 非参数的Kruskal-Wallis检验

§3.2.1 单因素方差分析

方差分析

- 方差分析研究**因素**对**指标**的影响。
- 单因素方差分析是两组比较的推广。如，比较不同年龄的记忆力，分60岁以下组和60岁以上两个组，比较两个组能记住的数字位数多少。可以用两样本t检验。但是，如果年龄组分成了15岁以下、16—60、60以上三个组，就不能再用两组比较的方法，需要使用单因素方差分析。
- 各组的指标均值有显著差异，则用来分组的**因素**对指标就有显著影响。比如，各年龄组记忆数字的平均位数如果有显著差异，就说明因素“年龄”对指标“记忆位数”有显著影响。
- 方差分析要回答的问题是：因素（分组）对指标的大小有无显著影响，如果有显著影响，那些组的指标较好。

数学模型

- 设因素A有 m 种不同取值（称为因素的**水平**）。在每个水平下，独立重复了 r 次试验，得到试验结果：

因素水平(分组)	试验结果
1	$Y_{11}, Y_{12}, \dots, Y_{1r}$
2	$Y_{21}, Y_{22}, \dots, Y_{2r}$
$\vdots$	$\vdots$
m	$Y_{m1}, Y_{m2}, \dots, Y_{mr}$

- 设各 Y_{ij} 相互独立，且

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, j = 1, \dots, r, \quad i = 1, \dots, m$$

其中 $\{\varepsilon_{ij}\}$ iid $N(0, \sigma^2)$ 。 $\{\alpha_1, \dots, \alpha_m\}$ 叫做因素A的主效应, $\sum_i \alpha_i = 0$ 。

- 假设检验问题

$$H_0: \alpha_1 = \cdots = \alpha_m = 0 \longleftrightarrow H_a: \text{不全为零}$$

- 计算各组的平均值，记作 $\bar{Y}_1, \bar{Y}_2, \dots, \bar{Y}_m$ 。比较平均值之间的差异大小，用组间平方和

$$S_A = r \sum_{i=1}^m (\bar{Y}_i - \bar{Y})^2$$

代表。

- 把组间平方和与组内平方和相比较:

$$S_e = \sum_{i=1}^m \sum_{j=1}^r (Y_{ij} - \bar{Y}_i)^2$$

- 用统计量

$$F = \frac{S_A/(m-1)}{S_e/(m(r-1))}$$

检验组间差异是否显著， F 大时认为组间有显著差异，即因素对指标有显著影响。

单因素方差分析的R程序

- 设因素A和指标Y的观测值分别在数据集d的两列中，变量名为A和y，则可以用R程序 `aov(y ~ A, data=d)` 来进行方差分析。
- 因素在R中用因子(factor)表示，用 `factor(x)` 生成因子。

例子

- 生产酱色的过程需要先进行除杂，比较5种不同除杂方法的除杂量，每种除杂方法试验4次，结果为:

除杂方法(水平)	除杂量(指标观测值)				组均值
1	25.6	22.2	28.0	29.8	26.4
2	24.4	30.0	29.0	27.5	27.7
3	25.0	27.7	23.0	32.2	27.0
4	28.8	28.0	31.5	25.9	28.6
5	20.6	21.2	22.0	21.2	21.3

- R程序:

```
A <- factor(rep(1:5, each=4))
y <- c(25.6, 22.2, 28.0, 29.8,
      24.4, 30.0, 29.0, 27.5,
      25.0, 27.7, 23.0, 32.2,
      28.8, 28.0, 31.5, 25.9,
      20.6, 21.2, 22.0, 21.2)
d <- data.frame(A, y)
plot(y ~ A, data=d)
summary(aov(y ~ A, data=d))
tapply(d$y, d$A, mean)
```

- 从图形看出，第5种方法指标明显低于其它四种方法。
- 结果的方法分析表给出了检验的p值0.016，在0.05水平下认为五种方法的除杂量有显著差异，或称不同方法对除杂量有显著影响。
- 分组计算了均值，但不确定哪些均值是有显著差异的。

§3.2.2 多重比较

均值的多重比较

- 如果各组之间有显著差异，那些水平之间是有显著差异的，那些水平之间是没有显著差异的？
- 可以进行多次两两比较，称为**多重比较**，但比较多次，第一类错误会增大。
- 多重比较的第一类错误概率分为**单次比较错误率**和**总错误率**。
- 为得到较多的显著差异结果，可以控制单次比较错误率。
- 为得到的结果可信，应该控制总错误率。

用R做多重比较

- 在R中用`pairwise.t.test(y, A, p.adjust="none")`进行各水平间的两两比较检验，控制单次比较错误率。如

```
pairwise.t.test(y, A, p.adjust="none")
```


在0.05水平下，找到有显著差异的水平对有(1,5), (2,5), (3,5), (4,5)，即方法5和前四种方法有显著差异，而前四种方法之间无显著差异。

- 用`pairwise.t.test(y, A)` 控制总错误率。如

```
pairwise.t.test(y, A)
```

这时有显著差异的水平对为(1,5),(2,5),(4,5)。

- 用`pairwise.t.test(y, A, p.adjust="fdr")` 控制错误发现率，即显著结果中错误显著的比例。控制FDR比控制总错误率的保守性低，可以找到较多显著差异。如

```
pairwise.t.test(y, A, p.adjust="fdr")
```

这时有显著差异的水平对为(1,5),(2,5),(4,5)。

Tukey的同时置信区间方法

- 控制总错误率的另一种方法是计算所有差 $\{\alpha_i - \alpha_k, i < k\}$ 的同时置信区间。差 $\alpha_i - \alpha_k$ 的置信区间不包含零则说明第*i*水平和第*k*水平有显著差异。
- 在R中，用`TukeyHSD(aov(y ~ A, data=d))`计算同时置信区间。如

```
TukeyHSD(aov(y ~ A, data=d))
```

结果发现(2,5), (4,5) 有显著差异。

§3.2.3 方差齐性检验

方差齐性检验

- 方差分析模型中要求各组的误差方差相同。为检验

$$H_0 : \text{各组方差相同} \longleftrightarrow H_a : \text{不同}$$

- 可以用Bartlett检验。如

```
bartlett.test(y ~ A, data=d)
```

结果p值为0.13，不显著。

- 可以用car包中的Levene检验:

```
require(car)
leveneTest(y ~ A, data=d)
```

结果p值为0.19，不显著。

Welch检验

- 方差分析的前提条件:

- 各组独立
- 正态分布
- 各组方差相等

其中“各组独立”是不可缺的。

- 如果“方差相等”条件不满足，可以使用Welch检验。如

```
oneway.test(y ~ A, data=d)
```

p值为0.0026，各组有显著差异。

§3.2.4 非参数Kruskal-Wallis检验

非参数Kruskal-Wallis检验

- 如果指标不服从正态分布，或根本是**有序数据**，比如调查问卷中回答强烈反对、反对、无意见、赞成、强烈赞成，可以用非参数的Kruskal-Wallis检验来比较组间有无显著差异。
- 想法是把所有指标观测值计算**秩**，比较各组的平均秩。
- 在R中，用`kruskal.test(y, A, data=d)`进行Kruskal-Wallis检验。如

```
kruskal.test(y ~ A, data=d)
```

结果p值为0.042，比经典方差分析的p值0.016大。

§3.3 两因素方差分析

多因素方差分析

- 进一步考虑多个因素 $A, B, C, \dots$ 对指标 Y 的影响。
- 需要回答:
 - 那些因素的主效应对指标有显著影响?
 - 因素之间有无交互作用?
 - 显著的因素的最好水平?

两因素方差分析

- 设指标 Y , 有两个因素 A, B , A 有 s 个水平, B 有 t 个水平。每个水平组合 (i, j) 重复试验 r 次。模型为

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

其中 μ 为总平均, $\{\alpha_i\}$ 称为因素 A 的主效应, $\{\beta_j\}$ 为因素 B 的主效应, $\{\gamma_{ij}$ 为因素 A 与因素 B 的交互作用效应。

- 参数满足条件

$$\begin{aligned} \sum_i \alpha_i &= 0, & \sum_j \beta_j &= 0 \\ \sum_i \gamma_{ij} &= 0, & \sum_j \gamma_{ij} &= 0 \end{aligned}$$

- $\{\varepsilon_{ijk}\}$ iid $N(0, \sigma^2)$ 。

主效应和交互作用效应

- 如果所有 $\gamma_{ij} \equiv 0$, 则模型称为可加模型。这意味着水平 (i, j) 的平均值 EY_{ijk} 只取决于 A 的水平 i 的平均, 和 B 的水平 j 的平均, 而与 (i, j) 搭配无关。
- 如果 $\gamma_{ij} \equiv 0$ 不全为零, 模型存在交互作用。
- 比如, 假设 $A = 1$ 时指标均值比总平均高 1 ($\alpha_1 = 1$), $B = 2$ 时指标均值比总平均低 2 ($\beta_2 = -1$), 在没有交互作用时, $(1, 2)$ 组合指标均值比总平均低 1。如果 $(1, 2)$ 组合时指标均值不是比总平均低 1, 而是高 1, 则存在交互作用 $\gamma_{12} = 2$ 。

假设检验

- 检验A的主效应:

$$H_0 : \alpha_1 = \cdots = \alpha_s = 0 \longleftrightarrow H_a : \text{不全为零}$$

考虑平方和

$$S_A = rt \sum_{i=1}^s (\bar{Y}_{i..} - \bar{Y})^2$$

- 检验B的主效应:

$$H_0 : \beta_1 = \cdots = \beta_t = 0 \longleftrightarrow H_a : \text{不全为零}$$

考虑平方和

$$S_B = rs \sum_{j=1}^t (\bar{Y}_{.j.} - \bar{Y})^2$$

- 检验交互作用效应:

$$H_0 : \gamma_{ij} \equiv 0 \longleftrightarrow H_a : \text{不全为零}$$

考虑平方和

$$S_{AB} = r \sum_{i=1}^s \sum_{j=1}^t (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y})^2$$

- 与纯误差平方和比较:

$$S_e = \sum_{i=1}^s \sum_{j=1}^t \sum_{k=1}^r (Y_{ijk} - \bar{Y}_{ij.})^2$$

R程序

- 在R中, 把数据保存为一个数据框d, 各列包括A、B和y。每一行是一次试验的水平组合和指标值。方差分析检验的程序为aov(y ~ A + B + A:B, data=d)。
- 如果检验结果发现交互作用不显著, 可以改进模型为可加模型: aov(y ~ A + B, data=d)

例子

- 检验毒药毒性的试验。因素A是三种不同的毒药, 因素B是四种不同治疗方案。因素A、B完全搭配有12种情况, 随机把48只老鼠分配到12组中, 每组4个试验结果, 指标为存活时间。
- 输入数据:

```
rats <- data.frame(
y = c(0.31, 0.45, 0.46, 0.43, # (1,1)
      0.82, 1.10, 0.88, 0.72, # (1,2)
      0.43, 0.45, 0.63, 0.76, # (1,3)
      0.45, 0.71, 0.66, 0.62, # (1,4)

      0.36, 0.29, 0.40, 0.23, # (2,1)
      0.92, 0.61, 0.49, 1.24, # (2,2)
      0.44, 0.35, 0.31, 0.40, # (2,3)
      0.56, 1.02, 0.71, 0.38, # (2,4)

      0.22, 0.21, 0.18, 0.23, # (3,1)
      0.30, 0.37, 0.38, 0.29, # (3,2)
      0.23, 0.25, 0.24, 0.22, # (3,3)
      0.30, 0.36, 0.31, 0.33), # (3,4)
Toxicant=factor(rep(1:3, each=4*4)),
Cure=factor(rep(rep(1:4, each=4), 3)))
```

- 用盒形图比较因素各水平:

```
opar <- par(mfrow=c(1,2))
plot(y ~ Toxicant + Cure, data=rats)
par(opar)
```

可以看出第三种毒药的毒性最大。四种治疗方案中第二种生存时间长一些,但也不一定显著。

- 做有交互作用的方差分析:

```
res <- aov(y ~ Toxicant + Cure
           + Toxicant:Cure, data=rats)
summary(res)
```

发现在0.05水平下交互作用效应不显著。

- 删去交互作用效应:

```
res2 <- aov(y ~ Toxicant + Cure, data=rats)
summary(res2)
tapply(rats$y, rats$Toxicant, mean)
tapply(rats$y, rats$Cure, mean)
```

两个因素的主效应都是高度显著的。最强毒效是第三种毒药。最好疗效是第二种治疗方案。

§3.4 协方差分析

协方差分析

- 方差分析中，除了因素对指标的影响外，其他的影响都归入随机误差项中。
- 有些情况下，存在对指标 Y 有影响的变量 X ，但我们无法控制 X 的取值，称这样的变量为**协变量**。
- 协方差分析研究扣除协变量影响后，指标与因素之间的关系。

模型

- 设指标 Y 受因素 A 和一个协变量 X 的影响。因素 A 有 m 个水平。试验结果的模型为

$$\begin{aligned}
 Y_{ij} &= \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \varepsilon_{ij}, \\
 j &= 1, \dots, n_i, \quad i = 1, \dots, m \\
 \varepsilon_{ij} &\text{ iid } N(0, \sigma^2) \\
 \sum_i n_i \alpha_i &= 0
 \end{aligned}$$

- 我们不关心 X 的影响，只希望检验

$$H_0: \alpha_1 = \dots = \alpha_m = 0 \longleftrightarrow H_a: \text{不全为零}$$

R程序

- 在R中，安装HH程序包后，用`ancova`可以进行协方差分析。也可以用`lm`做协方差分析。
- `ancova`使用如`ancova(y ~ X + A, data=d)`，其中 y 为指标， X 为协变量， A 为因素，用R的因子保存。
- `lm`使用如`summary(lm(y ~ X + A, data=d))`。

例子

- 例如，研究三种饲料对猪的催肥效果，饲料为因素 A ，指标为增重 Y ，考虑初始体重 X 作为协变量。每种饲料分配了8头猪进行试验。
- 读入数据:

```
d1 <- data.frame(A=1,  
x=c(15, 13, 11, 12, 12, 16, 14, 17),  
y=c(85, 83, 65, 76, 80, 91, 84, 90))  
d2 <- data.frame(A=2,  
x=c(17, 16, 18, 18, 21, 22, 19, 18),  
y=c(97, 90, 100, 95, 103, 106, 99, 94))  
d3 <- data.frame(A=3,  
x=c(22, 24, 20, 23, 25, 27, 30, 32),  
y=c(89, 91, 83, 95, 100, 102, 105, 110))  
d <- rbind(d1, d2, d3)  
d$A <- factor(d$A)
```

- 用HH包中的ancova分析:

```
require(HH)  
ancova(y ~ A + x, data=d)
```

结果显示因素A(饲料)高度显著。协变量(初始体重)也对增重量有高度显著的影响, 原因显然。

- 用lm分析:

```
anova(lm(y ~ A + x, data=d))
```

显示类似的结果。

§3.5 正交设计

正交设计

- 因素较多时，完全搭配所有组合需要大量试验。
- 比如，有 n 个因素，每个因素2个水平，完全搭配所有组合需要 2^n 次试验。 $n = 10$ 时，需要1024次试验。
- 正交试验法假定所有因素只有主效应(有些情况下可以考虑少量的交互作用)，然后巧妙设计试验方案使得所有因素两两搭配次数相同。
- 试验方案从一些预先设计好的正交表中选取。

 $L_8(2^7)$ 表

下面的 $L_8(2^7)$ 表用8次试验最多可以安排7个两水平因素的主效应:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 2 & 2 & 2 \\ 1 & 2 & 2 & 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 2 & 2 & 1 & 1 \\ 2 & 1 & 2 & 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 2 & 1 & 2 & 1 \\ 2 & 2 & 1 & 1 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 & 1 & 2 \end{pmatrix}$$

因素个数不足7个时可以选取其中的部分列组成试验方案。

例：烟灰砖试验的正交设计

- 研究用烟灰制砖问题，指标为折断力，考察的因素为成型水分(A)、碾压时间(B)、一次碾压料重(C)，各取3个水平。
- 完全试验需要 $3 \times 3 \times 3 = 27$ 次试验。假设模型为可加模型(只有主效应的模型)，使用如下的 $L_9(3^4)$ 表可以把试验次数减少到9次。

 $L_9(3^4)$ 正交表

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 3 & 3 \\ 2 & 1 & 2 & 3 \\ 2 & 2 & 3 & 1 \\ 2 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \\ 3 & 2 & 1 & 3 \\ 3 & 3 & 2 & 1 \end{pmatrix}$$

- 把因素A、B、C放在 L_9 表的前三列，删去第四列，作为试验方案。试验时可以把行次序打乱。
- 数据输入为数据集并绘图:

```
d <- data.frame(
  A = factor(rep(1:3, each=3)),
  B = factor(rep(1:3, 3)),
  C = factor(c(1,2,3, 2,3,1, 3,1,2)),
  y = c(16.9, 19.1, 16.7, 19.8, 23.7,
        19.0, 25.0, 20.4, 23.1))
```

- 绘图比较因素各水平:

```
opar <- par(mfrow=c(1,3))
plot(y ~ A + B + C, data=d)
par(opar)
```

- 图形提示因素A显著，B不显著，C显著。
- 做方差分析，仅考虑主效应:

```
summary(aov(y ~ A + B + C, data=d))
```

因素都不显著。

- 但根据以往经验的随机误差大小可以判断因素都显著。计算均值:

```
tapply(d$y, d$A, mean)
tapply(d$y, d$B, mean)
tapply(d$y, d$C, mean)
```

*A*的最好水平为3, *B*的最好水平为2, *C*的最好水平为3。

练习

练习

- 为研究溶菌酶水平在患胃溃疡的病人与正常人之间有无显著差异, 测量了一组病人和一组正常人的溶菌酶水平, 结果见下表。试检验两者的溶菌酶水平有无显著差异(水平0.05)。

胃溃疡	0.2	10.4	0.3	10.9	0.4	11.3	1.1	12.4
病人组:	2.0	16.2	2.1	17.6	3.3	18.9	3.8	20.7
	4.5	24.0	4.8	25.4	4.9	40.0	5.0	42.2
	5.3	50.0	7.5	60.0	9.8			
对照组:	0.2	5.4	0.3	5.7	0.4	5.8	0.7	7.5
	1.2	8.7	1.5	8.8	1.5	9.1	1.9	10.3
	2.0	15.6	2.4	16.1	2.5	16.5	2.8	16.7
	3.6	20.0	4.8	20.7	4.8	33.0		

提示: 要考虑分布是否正态。

- 假设对随机选取的8位病人使用A药物, 对随机选取的6位病人使用B药物, 过一段时间后测量每位病人体细胞中的药物含量, A、B药物的测量数据见下表:

A药物:	1.23	1.42	1.41	1.62	1.55	1.51
	1.60	1.76				
B药物:	1.76	1.41	1.87	1.49	1.67	1.81

在0.10水平下检验B药物的含量是否高于A药物的含量。

- 为了考察两种测量萘含量的液体层析方法: 标准方法和高压方法的测量结果有无显著差异, 取了10份试样, 每份分为两半, 一半用标准方法测量, 一半用高压方法测量, 每个试样的两个结果如下表, 试检验这两种化验方法有无显著差异(水平0.05):

标准:	14.7	14.0	12.9	16.2	10.2
	12.4	12.0	14.8	11.8	9.7
高压:	12.1	10.9	13.1	14.5	9.6
	11.2	9.8	13.7	12.0	9.1

- 为了研究药物补钙对高血压是否有疗效, 随机选取了10个人服用补钙药物, 11个人服用安慰剂, 预先记录这些人的血压。12周后测量每人的血压并减去原来的血压, 得到如下的血压变化数据:

补钙组:	7	-4	18	17	-3	-5
	1	10	11	-2		
安慰剂组:	-1	12	-1	-3	3	-5
	5	2	-11	-1	-3	

在0.10水平下检验服用补钙药物与服用安慰剂相比是否血压降低更多。

- 在一个双盲试验(受试者和操作者都不知道分组情况)中研究了咖啡因对受试者反映能力的影响。选了30个大学生进行按键速度测试,把这30人随机分为三组,每组10人,分别服用三种不同剂量的咖啡因(0 mg, 100 mg, 200 mg)。服药后记录每人每分钟按键次数。数据如下:

咖啡因剂量	每分钟按键数				
0 mg	242	245	244	248	247
	248	242	244	246	242
100 mg	248	246	245	247	248
	250	247	246	243	244
200 mg	246	248	250	252	248
	250	246	248	245	250

1. 对三个组的数据作并列的盒形图,看各组之间有无显著差异。
 2. 用方差分析表检验不同剂量的三组的按键次数有无显著差异(0.10水平)并解释结果。
- 为试制某种化工产品,在三种不同温度、四种不同压力下试验,每一水平组合重复两次,得到产品的收率数据如下(%):

温度	压力			
	1	2	3	4
1	52, 57	42, 45	41, 45	48, 45
2	50, 52	47, 45	47, 48	53, 30
3	63, 58	54, 59	57, 60	58, 59

试在0.05水平下进行方差分析并简述结果。

- 为了考察法院判决是否与被告种族有关,调查了326位被告的判决情况:

	黑人	白人
有罪	17	19
无罪	149	141

试在0.05水平下检验判决结果与被告种族是否独立。

- 下表为100位被调查者的性别及颜色偏好情况。

性别	颜色偏好		
	红	蓝	绿
男	32	14	4
女	25	17	8

试在0.05水平下检验颜色偏好是否与性别有关。

- 下表为200个婴儿的喂养方法(牛奶、母乳或并用)及母亲的经济状况的调查情况。试在0.05水平下检验喂养方法是否与母亲的经济状况有关。

喂养方法	经济状况			
	贫穷	下	中	上
牛奶	30	15	11	12
母乳	7	18	19	29
并用	5	23	7	19

第四章 统计模型

§4.1 相关与回归

相关与回归

- 相关分析
- 一元线性回归
- 曲线拟合方法
- 多元线性回归
- Logistic回归

§4.1.1 相关分析

相关分析

- 考虑连续型随机变量之间的关系。相关系数定义为

$$\rho(X, Y) = \frac{E[(X - EX)(Y - EY)]}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

又称Pearson相关系数。

- $-1 \leq \rho \leq 1$ 。 ρ 接近于+1表示 X 和 Y 有正向的相关； ρ 接近于-1表示 X 和 Y 有负向的相关。
- 相关系数代表的是**线性**相关性，对于 X 和 Y 的其它相关可能反映不出来，比如 $X \sim N(0, 1)$, $Y = X^2$, 有 $\rho(X, Y) = 0$ 。
- 给定样本 $(X_i, Y_i), i = 1, 2, \dots, n$, 样本相关系数为

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

探索性分析

- 用散点图和散点图矩阵直观地查看变量间的相关。
- 例如，线性相关的模拟数据的散点图:

```
nsamp <- 30
x <- runif(nsamp, -10, 10)
y <- 20 + 0.5*x + rnorm(nsamp, 0, 0.5)
plot(x, y)
```


- 二次曲线相关的模拟数据散点图:

```
y2 <- 0.5*x^2 + rnorm(nsamp,0,2)
plot(x, y2)
```

- 指数关系的例子:

```
y3 <- exp(0.2*(x+10)) + rnorm(nsamp,0,2)
plot(x, y3)
```

- 对数关系的例子:

```
y4 <- log(10*(x+12)) + rnorm(nsamp,0,0.1)
plot(x, y4)
```

R程序与相关

- 设变量 X 和 Y 的样本存放于R向量 x 和 y 中，用`cor(x,y)`计算样本相关系数。
- 为检验相关系数是否等于零，可用`cor.test(x,y)`。
- 例如:

```
cor(x, y)
cor.test(x, y)
```

得 x 与 y 的相关系数为0.9780，检验的 p 值为 $2.2e-16$ ，95%置信区间为[0.9537, 0.9896]。

- 再例如:

```
cor(x, y2)
cor.test(x, y2)
```

得 x 与 y_2 的样本相关系数为0.1629， p 值为0.3897，不显著。

§4.1.2 一元回归分析

一元回归分析

- 考虑两个变量 Y 与 X 的关系, 希望用 X 值的变化解释 Y 值的变化。 X 称为自变量(independent variable), Y 称为因变量(response variable)。
- 假设模型

$$Y = a + bX + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- 设观测值为 $(X_i, Y_i), i = 1, 2, \dots, n$, 假设观测值满足上模型

$$Y_i = a + bX_i + \varepsilon_i, \quad \varepsilon_i \text{ iid } \sim N(0, \sigma^2)$$

最小二乘法

- 直观上看, 要找最优的直线 $y = a + bx$ 使得直线与观测到的点最接近。例如:

```
plot(x, y)
abline(lm(y ~ x), col="red", lwd=2)
```

- a, b 的解用最小二乘法得到:

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

参数估计公式

- 最小二乘解表达式为

$$\hat{b} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = r_{xy} \frac{S_y}{S_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

其中 r_{xy} 为 X 与 Y 的样本相关系数, S_x 与 S_y 分别为 X 和 Y 的样本标准差。

- 随机误差方差 σ^2 的估计取为

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i (y_i - \hat{a} - \hat{b}x_i)^2$$

- 标准误差: 为衡量 $\hat{a}$ 和 $\hat{b}$ 的估计精度, 计算 $SE(\hat{a})$ 和 $SE(\hat{b})$ 。

估计结果

- 拟合值(预测值)

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \quad i = 1, 2, \dots, n$$

- 残差

$$\hat{e}_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n$$

- $\sum_i \hat{e}_i^2$ 称为残差平方和, 残差平方和小则拟合优。

回归有效性检验

- 当 $b = 0$ 时, 模型退化为 $Y = a + \varepsilon$, X 不出现在模型中, 说明 Y 与 X 不相关。检验

$$H_0 : b = 0 \longleftrightarrow H_a : b \neq 0$$

- 可以用 $\hat{b}/\text{SE}(\hat{b})$ 计算一个 t 统计量进行检验, 或把因变量的总变差 $\sum_i (y_i - \bar{y})^2$ 分解为模型平方和 $\sum_i (\hat{y}_i - \bar{y})^2$ 和残差平方和 $\sum_i (y_i - \hat{y}_i)^2$ 两部分, 看模型平方和与残差平方和之比是否足够大。

R中回归分析的计算

- 设数据保存在数据框 d 中, 变量名为 y 和 x , 用 `lm(y ~ x, data=d)` 计算回归结果, 如:

```
res <- lm(y ~ x); res
```

结果只有回归系数。需要用

```
summary(res)
```

显示较详细的结果, 如 $H_0 : b = 0$ 的检验结果 p 值为 $< 2.2e - 16$ 。

- 又如对 `class` 数据集, 建立体重对身高的回归方程:

```
d <- read.csv("class.csv", header=TRUE)
lm1 <- lm(weight ~ height, data=d); summary(lm1)
plot(weight ~ height, data=d)
abline(lm1, col="red", lwd=2)
```

身高对体重的影响是显著的(p 值 $7.89e-7$)。

简单回归诊断

- `plot(res)` (`res`为回归结果变量)可以做简单回归诊断。
- 第一个图是残差对预测值散点图, 散点应该随机在0线上下波动, 不应该有曲线模式、分散程度增大模式、特别突出的离群点等情况。
- 第二个图是残差的正态QQ图, 散点接近于直线时可以认为模型误差项的正态分布假定是合理的。
- 第三个图是误差大小(标准化残差绝对值的平方根)对拟合值的图形, 可以判断方差齐性假设(方差 σ^2 不变)。
- 第四个图是残差对杠杆量图, 并叠加了Cook距离等值线。杠杆量代表了回归自变量对结果的影响大小, 超过 $4/n$ 的值是需要重视的。Cook距离考察删去第 i 个观测对回归结果的影响。

预测区间

- 设 X 取 x_0 , Y 的预测值为 $\hat{y}_0 = \hat{a} + \hat{b}x_0$ 。置信水平为 $1 - \alpha$ 的置信区间为

$$\hat{y}_0 \pm \lambda \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_i (x_i - \bar{x})^2}}$$

- 用`predict(res)`得到 $\hat{y}_i, i = 1, \dots, n$ 的值, `res`表示回归结果变量。用`predict(res, interval="prediction")`同时得到预测的置信区间, 需要的话加入`level=`选项设定置信度。

控制

- 如果需要把 Y 的值控制在 $[y_l, y_u]$ 范围内, 问如何控制 X 的范围, 可以求解 x_0 的范围使上面的置信区间包含在 $[y_l, y_u]$ 内。
- 近似地可以解不等式

$$\begin{aligned}\hat{a} + \hat{b}x_0 - z_{1-\frac{\alpha}{2}} \hat{\sigma} &\geq y_l \\ \hat{a} + \hat{b}x_0 + z_{1-\frac{\alpha}{2}} \hat{\sigma} &\leq y_u\end{aligned}$$

其中 $z_{1-\frac{\alpha}{2}}$ 为标准正态分布双侧 α 分位数 (用`qnorm(1-alpha/2)`计算)。

§4.1.3 曲线拟合

非线性回归模型

- 线性回归模型可以看成非线性回归模型的特例:

$$Y = f(X) + \varepsilon$$

其中 $f(x)$ 为未知的回归函数。

- 参数方法: 假定 $f(x)$ 具有某种形式, 如
 - $f(x) = a + bx$: 线性回归;
 - $f(x) = a + bx + cx^2$: 二次多项式回归;
 - $f(x) = Ae^{bx}$: 指数模型, 等等。
- 二次多项式回归可以令 $X_1 = x, X_2 = x^2$, 变成二元回归模型来解决, 程序如`lm(y ~ x + I(x^2), data=d)`。指数模型可以令 $z = \ln Y$, 模型化为 $z = a + bx$ 。
- 有一些曲线模型可以通过变换化为线性回归。

非参数曲线拟合

- 为了得到一般性的 Y 与 X 的曲线关系 $f(x)$ 的估计, 可以使用样条函数。样条函数是光滑的分段三次多项式。
- 用样条函数估计 $f(x)$ 的准则是曲线接近各观测值点 (x_i, y_i) , 同时曲线足够光滑。
- 在R中用`smooth.spline`函数进行样条曲线拟合。如

```
nsamp <- 30
x <- runif(nsamp, -10, 10)
x <- sort(x)
y <- 10*sin(x/10*pi)^2 + rnorm(nsamp,0,0.2)
plot(x, y)

require(splines)
res <- smooth.spline(x, y)
lines(spline(x, res$y), col="red")
```

其中`res`的元素`y`为拟合值, 用`spline(x,y)`从一组散点输出光滑曲线以便用`lines`函数绘图。

§4.1.4 多元线性回归

多元线性回归

- 模型

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

- 数据放在一个数据框中，有一列Y和p列自变量。
- 模型估计仍使用最小二乘法，得到系数估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 及误差方差估计 $\hat{\sigma}^2$ 。
- 为了检验整个回归模型是否都无效，考虑假设检验：

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

- 为了检验某一个自变量 X_j 是否对因变量的解释有贡献（在模型中已经包含了其它自变量的情况下），检验

$$H_0 : \beta_j = 0$$

R程序

- 在R中用`lm(y ~ x1 + x2 + x3, data=d)`这样的程序来做多元回归，数据集为d，自变量为x1,x2,x3三列。
- 比如，class数据集中体重对身高和年龄的回归：

```
lm2 <- lm(weight ~ height + age, data=d)
summary(lm2)
```

发现关于年龄的系数为零的检验p值为0.69, 不显著，说明在模型中已经包含身高的情况下，年龄不提供对体重的额外信息。

- 但是如果体重对年龄单独建模的话，年龄的影响还是显著的：

```
lm3 <- lm(weight ~ age, data=d)
summary(lm3)
```

- 模型中不显著的自变量应该逐一剔除。可以用`step`函数进行逐步回归变量选择，如：

```
lm4 <- step(lm(weight ~ height + age + sex, data=d))
summary(lm4)
```

预测

- 得到回归模型结果res后, 要对原数据框中的观测值做预测, 只要使用`predict(res)`。
- 为了使用得到的模型结果res对新数据做预测, 建立包含了自变量的一组新的观测值的数据框dp, 用`predict(res, newdata=dp)`做预测。

不同截距项的模型

- 分类变量作为R的因子(factor), 可以用在回归模型中。比如, 为了表示男生和女生的体重有不同, 可以在以体重为因变量的回归中加入自变量性别:

```
lm5 <- lm(weight ~ height + sex, data=d)
summary(lm5)
```

结果中的sexM项表示以女生为基数, 男生体重的平均增加量。这一项不显著。

§4.1.5 Logistic回归

Logistic回归

- 当因变量Y是零壹变量时, 即Y表示分两类的类别, 取值1和0, 我们关心的是 $P(Y = 1)$ 。这是一个 $[0, 1]$ 区间内的值。如果把Y当作一般因变量做线性回归, 会给出不合理的结果, 比如负值, 另外线性回归假定误差项为正态分布在这里也不适用。
- 为此考虑广义的回归模型(广义线性模型):

$$Y \sim F(y; \theta)$$

$$g(\theta) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- 特别地, 定义logit函数

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right)$$

- Logit模型:

$$Y \sim b(1, p)$$

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

R程序

- 进行logistic回归的R程序如`glm(y ~ x1+x2, family=binomial, data=d)`, 其中`y`为取0或1的向量, 输入数据集为`d`。
- `y`也可以是一个两列的矩阵, 第一列为成功数, 第二列为失败数。
- 例如, “remiss.csv”中保存了癌症病人康复的数据, 变量`remiss`为康复与否(1为康复, 0为未康复), 另外的6个变量为可能影响康复概率的自变量。
- 程序:

```
remiss <- read.csv("remiss.csv", header=TRUE)
res1 <- glm(remiss ~ cell+smear+infil+li
            +blast+temp, family=binomial,
            data=remiss)
summary(res1)
```

- 以`p`值0.30为界限, 逐步删去不显著的自变量:

```
res2 <- glm(remiss ~ cell+smear+infil+li
            +temp, family=binomial,
            data=remiss)
summary(res2)

res3 <- glm(remiss ~ cell+infil+li
            +temp, family=binomial,
            data=remiss)
summary(res3)

res4 <- glm(remiss ~ cell+li+temp,
            family=binomial, data=remiss)
summary(res4)
```

- 或可用逐步回归:

```
ress <- step(glm(remiss ~ cell+smear+infil+li
                +blast+temp, family=binomial,
                data=remiss))
summary(ress)
```


§4.2 多元分析

多元分析

- 主成份分析
- 因子分析
- 聚类分析
- 典型相关分析

§4.2.1 主成份分析

主成份分析

- 假设有多个变量 $Y_1, Y_2, \dots, Y_p$ 。设法用若干个线性组合(称为主分量)浓缩这 p 个变量的信息。
- 问题归结到求 $Y_1, \dots, Y_p$ 的协方差矩阵或相关系数阵的特征值和特征向量, 用特征向量作为线性组合系数。
- 保留前几个主分量, 使得前几个主分量能够解释原来变量的70%–90%以上的变差。
- 如果原来的变量 $Y_1, \dots, Y_p$ 大小是可比的, 用协方差矩阵来得到主成份。否则需要把原始变量标准化, 用相关系数阵来得到主成份。

R程序

- R中用princomp做主成份分析。如

```
nsamp <- 30; s <- 1
p01 <- runif(nsamp, -10, 10)
p02 <- runif(nsamp, 0, 5)
x1 <- 0.5*p01 + 0.1*p02 + rnorm(nsamp, 0, s)
x2 <- 0.1*p01 - 0.5*p02 + rnorm(nsamp, 0, s)
x3 <- 0.5*p01 - 0.1*p02 + rnorm(nsamp, 0, s)
x4 <- -0.1*p01 + 0.5*p02 + rnorm(nsamp, 0, s)
res1 <- princomp(~x1+x2+x3+x4, cor=FALSE, scores=TRUE)
summary(res1)

res2 <- princomp(~x1+x2+x3+x4, cor=TRUE)
summary(res2)
```

计算主成份得分

- 用`predict(res)`计算各主成份值（称为主成份得分）。如

```
pr1 <- predict(res1)
plot(d[,1], pr1[,1])
```

- 对新的数据, 假设数据放在数据框`newd`中, 用`predict(res, newdata=newd)`计算新数据的主成份得分。

§4.2.2 因子分析

因子分析

- 设有 $Y_1, Y_2, \dots, Y_p$ 变量, 假设这些变量有若干个共同的解释因子:

$$Y_i = \mu_i + \lambda_{i1}f_1 + \dots + \lambda_{ik}f_k + \varepsilon_i$$

其中 $f_1, \dots, f_k$ 为公共因子, ε_i 为特殊因子。

- 希望找到有解释意义的公共因子集合。

R程序

- 在R中用`factanal`函数做因子分析。
- 数据“scores.csv”中包含100个学生的学号、语文、数学、英语、物理、化学、生物的成绩。希望做主成份分析。

```
scores <- read.csv("scores.csv", header=TRUE)
dd <- scores[,2:7]
res <- factanal(dd, factors=2)
res
```

- 从结果看出, 第一主成份代表物理、化学、生物, 第二主成份代表语文、数学、英语。

§4.2.3 判别分析

判别分析

- 设因变量 Y 为分类变量, Y 与自变量 $X_1, \dots, X_p$ 有关, 希望建立规则, 由自变量值判别 Y 的类取值。
- Fisher判别法可以建立自变量的线性组合, 用该线性组合判别 Y 的类属。
- R中用MASS包中的lda函数做Fisher判别分析。
- 例如, R中自带的数据iris中包含了Fisher的鸢尾花数据, 有三种不同类型的鸢尾花, 变量Species中为种类, 用四个自变量对种类建立判别分析公式。
- 程序示例:

```
data(iris)
plot(Sepal.Length ~ Species, data=iris)
require(MASS)
res <- lda(Species ~ Sepal.Length+Sepal.Width
           +Petal.Length+Petal.Width, data=iris)
res
```

- 用predict(res)对原数据做判别, 用predict(res, newdata=newd)对新数据进行判别。如:

```
pr <- predict(res)$class
table(iris[, "Species"], pr)
```

可以看出仅判错了3例。当然, 对新数据, 结果一般要差一些。

§4.2.4 聚类分析

聚类分析

- 聚类分析和判别分析都涉及到分类问题。区别在于, 判别分析是从一些已知分类样例中找到判别规则; 聚类分析是从未知分类的样例中根据距离远近得到分类。
- 判别分析属于“有导师学习(supervised learning)”, 聚类分析属于“无导师学习(unsupervised learning)”。

- 聚类分析中的“系统聚类(hierarchical clustering)”方法把 n 个样本点初始看成 n 个类, 然后逐次进行类合并, 每次把距离最近的两个类合并。
- 最后保留多少个类主要靠认为判断。

样本点的距离

- 距离分为样本点之间的距离和两类之间的距离。
- 两个样本点 $\mathbf{x} = (x_1, \dots, x_p)$ 和 $\mathbf{y} = (y_1, \dots, y_p)$ 之间的距离定义有:
 - Manhattan距离 $d = \sum_k |x_k - y_k|$ 。
 - 欧氏距离(Euclidean) $d = \sqrt{\sum_k (x_k - y_k)^2}$ 。
 - Minkowski距离 $d = (\sum_k |x_k - y_k|^\gamma)^{1/\gamma}$ ($\gamma > 0$)。
 - Chebyshev距离(R中用maximum表示) $d = \max_{k=1,2,\dots,p} |x_k - y_k|$ 。
 - 马氏(Mahalanobis)距离 $d = (\mathbf{x} - \mathbf{y})' S^{-1} (\mathbf{x} - \mathbf{y})$, 其中 S 为 $\mathbf{X}$ 的协方差阵。
 - 兰氏(canberra)距离 $d = \frac{1}{p} \sum_{k=1}^p \frac{|x_k - y_k|}{x_k + y_k}$, 要求 $x_k > 0, y_k > 0$ 。
- 在R中, 用`dist(x, method="距离名")`计算存放在数据框 x 中的各个观测(行)的距离矩阵, 输出为下三角矩阵。
- 如

```
data(iris)
di <- dist(iris[,1:4], method="euclidean")
print(as.matrix(di)[1:10,1:10])
```

类间距离

- 类平均(average linkage);
- 重心法(centroid);
- 中位距离(median);
- 最长距离(complete);
- 最短距离(single);
- 离差平方和(ward);
- 密度估计法(density)。

R程序

- 在R中，先用dist函数计算得到距离矩阵d，然后以距离矩阵d为输入调用hclust函数，用法为hclust(d, method="类距离")。
- 用plot(res)对聚类结果绘图，可以由此判断分成几个类合适。用rect.hclust(res, k=类数) 在聚类图中显示分割成指定类数的效果。
- 用cutree(res, k=类数)得到分类结果。
- 比如，对iris数据聚类：

```
res <- hclust(di, method="complete")
plot(res, labels=FALSE)
rect.hclust(res, k=3)

clus <- cutree(res, k=3)
table(iris[, "Species"], clus)
```

- 结果中, setosa被完美地归入了新类1中；virginica被基本完美地归入了新类2中；versicolor的新分类很不好，一半进入新类3，但有一半被归入virginica的类。

§4.2.5 典型相关分析

典型相关分析

- 两个变量间的线性相关性用相关系数衡量。
- 一个变量Y和多个变量 $X_1, \dots, X_p$ 之间的线性相关性可以用复相关系数平方衡量。
- 把复相关系数推广到一组变量 $Y_1, \dots, Y_q$ 和一组变量 $X_1, \dots, X_p$ 之间的线性相关性，称为**典型相关分析**(canonical correlation)。
- 例如，肉、蛋、奶、大米、面粉的价格与这些产品的供应量之间的关系；运动员的体力测试指标(反复横向跳、纵跳、背力、握力)与运动能力测试指标(耐力跑、跳远、投球)之间的关系，等等。
- 典型相关分析的思想是，构造 $Y_1, \dots, Y_q$ 的线性组合 $U = a_1 Y_1 + \dots + a_q Y_q$ 和 $X_1, \dots, X_p$ 的线性组合 $V = b_1 X_1 + \dots + b_p X_p$ ，使得 U, V 的相关系数最大，定义这个相关系数为两组变量之间的典型相关系数。

- 与主成份分析类似,可以在找出第一对最大相关的线性组合 U, V 之后,再寻找与原来的组合不相关的组合使互相之间的相关最大。
- 这样有助于找到为了研究 $Y_1, \dots, Y_q, X_1, \dots, X_p$ 如何最好地降维。如果仅对 $X_1, \dots, X_p$ 做主成份分析,不能很好地选取和 $Y_1, \dots, Y_q$ 相关的信息。

R程序

- 在R中用`cancor(x, y)`来计算典型相关分析。
- 比如,在iris数据集中,计算花萼长、宽与花瓣长、宽之间的典型相关系数:

```
data(iris)
res <- cancor(iris[,1:2], iris[,3:4])
print(res)
```

可见,第一典型相关就达到0.94。

- 结果中给出了典型相关分量(U 和 V)的计算公式。用`xcoef`的第一列做线性组合可以得到 x 的分量,用`ycoef`的第一列做线性组合可以得到 y 的分量。

R程序

- 在数据“jobs.csv”中包含关于工作性质和工作满意度的一些测量数据。工作满意度的测量值包括Career(对提职的满意度)、Supervis(对主管的满意度)、Finance(对收入的满意度)。工作性质的测量值包括Variety(工作内容丰富)、Feedback(主管的反馈多少)、Autonomy(工作自主性)。
- 要研究工作性质与工作满意度的相关性。
- 程序:

```
jobs <- read.csv("jobs.csv", header=TRUE)
res <- cancor(jobs[,1:3], jobs[,4:6])
print(res)
```

可见,第一典型相关就达到0.92。

§4.3 时间序列分析

时间序列分析

- 平稳时间序列
- 自回归模型
- 白噪声检验
- 单位根检验
- 预报方法

§4.3.1 时间序列

时间序列

- 时间序列是随时间记录的数据的数学模型，比如每小时记录的机器温度，每天的股指，等等。
- 时间序列表示为 $\{X_t, t = 1, 2, \dots\}$ 。
- 在R中，用`ts(x)`把一个向量转换为一个时间序列对象。
- 平稳时间序列:

$$EX_t \equiv \mu$$

$$\text{Cov}(X_t, X_{t+k}) = \gamma_k, k = 0, 1, 2, \dots$$

即：水平不变；前后的线性相关性不变。

自相关函数

- 自相关函数(ACF):

$$\rho_k = \text{corr}(X_t, X_{t+k}) = \frac{\gamma_k}{\gamma_0}$$

在R中用`acf(x)`做ACF图。

- 白噪声: 如果 $\varepsilon_t, t = 1, 2, \dots$ 期望为零，方差为统一的 σ^2 ，但 $\text{Cov}(\varepsilon_t, \varepsilon_s) = 0, t \neq s$ ，称这样的平稳列为白噪声。
- 白噪声的自相关函数满足: $\rho_k \equiv 0, k \geq 1$ 。
- R中ACF图中有上下两条水平虚线，是在假定白噪声的情况下每个 $\hat{\rho}_k$ 的置信限。如果ACF除 $k = 0$ 处以外都在线内则可以认为序列是白噪声。记为 $\text{WN}(0, \sigma^2)$ 。

北京地区洪灾数据

- 在R矩阵d.flood中包含了北京地区1949—1964每年的洪水受灾面积和成灾面积。
- 两个序列转换为时间序列:

```
flood.area1 <- ts(d.flood[, "area1"],  
                  frequency=1,  
                  start=1949)  
flood.area2 <- ts(d.flood[, "area2"],  
                  frequency=1,  
                  start=1949)  
plot(flood.area1)  
lines(flood.area2, col="red", lty=2)
```

- 做自相关函数图:

```
acf(flood.area1)
```

序列可能为白噪声。

-

居民用煤消耗季度数据

- 输入北京市居民用煤消耗季度数据:

```
coal.consumption <-  
  ts(c(  
    6878.4 , 5343.7 , 4847.9 , 6421.9 ,  
    6815.4 , 5532.6 , 4745.6 , 6406.2 ,  
    6634.4 , 5658.5 , 4674.8 , 6445.5 ,  
    7130.2 , 5532.6 , 4989.6 , 6642.3 ,  
    7413.5 , 5863.1 , 4997.4 , 6776.1 ,  
    7476.5 , 5965.5 , 5202.1 , 6894.1 ),  
    frequency=4, start=c(1991,1))  
plot(coal.consumption)
```


从图中看出明显的季节性(周期性)波动。

- 做ACF:

```
acf(coal.consumption)
```

明显不是白噪声。

§4.3.2 时间序列的检验

白噪声检验—Durbin-Watson

- 回归分析中要求因变量(或随机误差项)不能带有自相关。R中car包的durbinWatsonTest()函数对回归结果做残差的Durbin-Watson检验，零假设为残差无序列自相关。如:

```
> require(car)
> durbinWatsonTest(lm(flood.area1 ~ 1))

lag Autocorrelation D-W Statistic p-value
  1          0.2218556      1.383029   0.216
Alternative hypothesis: rho != 0
```

结果可以认为洪灾受灾面积为白噪声。

白噪声检验—Box-Pierce

- Box.test()函数可以进行Box-Pierce白噪声检验和Ljung-Box白噪声检验。如

```
> Box.test(flood.area1)

Box-Pierce test

data: flood.area1
X-squared = 0.7875, df = 1, p-value = 0.3749
```

做Box-Pierce检验, 结果可以认为洪灾受灾面积为白噪声。

- 做Box-Ljung检验:

```
> Box.test(flood.area1, type="Ljung-Box")

Box-Ljung test

data:  flood.area1
X-squared = 0.945, df = 1, p-value = 0.331
```

承认白噪声。

单位根检验

- 单位根检验是关于序列不平稳的检验。
- R中tseries包中的`adf.test()`函数做Dickey-Fuller单位根检验。零假设是序列有单位根, 序列不平稳, 对立假设是没有单位根, 序列是平稳的。
- 如:

```
> require(tseries)
> adf.test(flood.area1)

Augmented Dickey-Fuller Test

data:  flood.area1
Dickey-Fuller = -3.2158, Lag order = 2,
p-value = 0.1092
alternative hypothesis: stationary
```

可以认为受灾面积序列平稳(取检验水平0.15)。

- 用煤消耗季度数据:

```
> adf.test(coal.consumption)

Augmented Dickey-Fuller Test

data:  coal.consumption
Dickey-Fuller = -5.2275, Lag order = 2,
p-value = 0.01
alternative hypothesis: stationary
```

可以认为平稳(取检验水平0.15)。

§4.3.3 时间序列分析模型

时间序列分解

- 时间序列的分解:

$$X_t = T_t + S_t + e_t$$

- 分解为趋势、季节和随机部分。
- 趋势可以用以 t 为自变量的线性回归、曲线回归。
- 季节项可以用因子作为自变量进行回归。
- 随机部分可能有序列自相关, 需要用平稳时间序列模型建模。典型的模型有自回归(AR)模型、滑动平均(MA)模型、ARMA模型和非平稳的ARIMA模型。

AR模型

- AR(1)模型:

$$X_t = \rho X_{t-1} + \varepsilon_t$$

其中 $\varepsilon_t \sim \text{WN}(0, \sigma^2)$; $|\rho| < 1$ 。这是零均值的模型。

- AR(p)模型:

$$X_t = a_1 X_{t-1} + \cdots + a_p X_{t-p} + \varepsilon_t$$

- 偏自相关系数(PACF): $\{\psi(k), k = 1, 2, \dots\}$, 为 X_t 与 X_{t+k} 在排除了 $X_{t+1}, X_{t+2}, \dots, X_{t+k-1}$ 的影响后的相关。
- AR(p)的偏相关系数在 p 后截尾:

$$\psi(k) = 0, k > p$$

可以作为建模估计 p 的方法。

R程序

- 在R中用`pacf(x)`做偏相关系数图。图形中两条水平虚线内的值可以认为是零。如:

```
pacf(coal.consumption)
```

从图中看出可以用AR(3)建模。

- 用`arma(x, order=c(p,0,0))`建立AR(p)模型。如

```
res <- arma(coal.consumption, order=c(3,0,0))  
summary(res)
```

- 模型方程为

$$X_t = 3932 + 0.64X_{t-1} - 0.93X_{t-2} + 0.65X_{t-3} + \varepsilon_t,$$
$$\varepsilon_t \sim \text{WN}(0, \sigma^2)$$

- 用`fitted(res)`计算拟合值。拟合情况绘图:

```
pr <- fitted(res)  
plot(coal.consumption)  
lines(pr, col="red", lty=2)
```

第五章 质量管理

§5.1 质量概述

质量概述

- 本讲内容来自《质量管理与质量控制》(第7版), J.R.Evans 和W.M. Lindsay, 人民大学出版社。
- 质量管理历史
- 质量定义
- 全面质量管理
- 质量与竞争优势
- 质量的三个层面

质量的重要性

- 质量是制造业的三个重要议题之一: 生产率、成本、质量。
- 质量的重要性是根本性的。
- 质量是重要的竞争优势源泉。

质量的历史

- 手工业时代, 匠人的个人行为。
- 可互换零件引发工业革命, 也引起了对质量控制的需求。主要靠最后检验来控制质量。

质量的历史(续1)

- 20世纪早期:
 - 科学管理之父Frederick W. Taylor 提出把计划和执行分开, 提高了生产率, 但也使得非质量部门人员对质量不重视。
 - 第二次工业革命的领导人之一Henry Ford Sr.与20世纪初提出了“全面质量管理”, 但只在日本得到了贯彻。
 - 1920年代贝尔系统的几位质量先驱提出了“统计质量控制”(SQC), 包括Walter Shewart, Harold Dodge, George Edwards, W. Edwards Deming。Deming是日本全面质量管理的关键人物之一。
 - 统计质量控制是统计方法在质量控制中的应用, 超越了检验的范畴, 关注的是识别和解决引起缺陷的那些问题。
 - 二战期间, 美军采用统计抽样程序, 对供应商指定严格标准, 培养了质量方面的专家, 使SQC在制造业被接受。

质量的历史(续2)

- 在1950年代A.V. Feigenbaum提出全面质量控制。
- 二战后，美国的朱兰(Juran)和戴明(Deming)把SQC介绍到了日本，日本制造业全面推广SQC，到1970年代日本产品质量超过了美国。
- 美国1980-1990年代全社会对质量问题引起重视，质量书籍、咨询、培训变得丰富。
- 全面质量管理(Total Quality Control)的提出。超越了制造质量管理，包含整个组织过程的质量的管理。质量升级为“卓越绩效(performance excellence)”。提出了“六西格玛”质量基本原则。
- 质量方面当前和未来的挑战：全球化；创新、创造、变革；外包；消费者期望的提高，不仅仅是质量要求；价值创造；质量观的改变，由过程模型相系统方法论演变。

质量的定义

- 不同人员对质量有不同观点，比如评判的观点（大家一致认为宝马、劳力士等质量高），产品性能、价格（性能高、价格高的质量就好？），用户适用（营销人员的观点），性价比（设计人员应该考虑），生产规范（生产人员通常采用）。
- 现代质量理论认为，质量就是满足或超越顾客的需要。顾客包括外部顾客和内部顾客。组织的每一个人都应该对质量负责。

质量作为一种管理框架

- 研究发现，受顾客赞赏的产品系列不仅仅是遵循生产质量管理，而需要重视满足顾客期望，通过市场调查来确定顾客的需求，使用基于顾客的质量绩效指标，在企业的所有职能领域都有正式的质量控制系统。
- 全面质量(TQ)就是组织中每一个成员为了理解、满足并超越顾客的期望而进行的坚决而持续的改进活动。
- 全面质量的原则：
 - 集中注意于顾客和利益相关者。
 - 组织中的每个成员的参与和团队合作。
 - 以持续改进和学习所制成的过程导向。
- 全面质量的实施包括基础架构、各部门的质量实践和工具。工具包括大量的图表和统计方法，用以计划工作活动，收集资料，分析结果，监控进展以及解决问题。

质量与竞争优势

- 调查发现：产品质量是决定企业利润的重要因素；
- 提供优质产品和服务的公司通常拥有很大的市场份额，并且是市场的早期进入者；
- 对于几乎所有的产品 and 市场形势来说，质量与高投资回报率是正向的显著相关；
- 实施质量改进战略通常会提高市场份额，但要付出短期利润率下降的代价；
- 高质量的生产者常常能够指定较高的价格。
- 大量证据表明，质量举措对于财务结果具有积极的作用。

全面质量的三个层面

- 组织层面：以满足外部顾客要求为战略。
- 过程层面：按部门划分，但也需要配合。
- 执行/岗位层面：产出标准必须以质量和顾客要求为基础，要求来自组织层面和过程层面。

§5.2 质量理念与框架

戴明的质量理念

- 戴明是1920-1930年代统计质量管理的先驱，但他的质量管理思想在日本得到贯彻。他强调管理层对质量的认识才是最重要的。戴明认为变异引起质量下降，提出永无止境的循环改进思想。
- 戴明的知识体系包括：对系统的重视；对变异的理解；知识的理论；心理学。

戴明的知识体系

- 系统：组织的生产是一个系统的过程，必须兼顾各利益相关者（股东、雇员、顾客、社会和环境）。比如，采购成本最低导致质量下降。
- 变异：多种微小复杂因素的影响被建模为随机变量，变异通常是由系统设计中的固有因素所引起的，这些因素无法轻易地控制，所以变异是不可避免的，但可以设法减少变异。变异的减少使生产商和消费者都受益。统计方法是确定和量化变异的首要工具。
- 知识的理论：管理者需要理论指导。经验值是在描述而不能被检验或确认，只靠经验无助于管理。于此相对照，理论有助于理解因果关系，这种因果关系能够应用于预测和理性的管理决策。不理解理论而只是肤浅地使用一些软件结果或遵从别人的建议导致失败。
- 心理学：人可以收到外在或内在的激励，但是，最强有力的激励因素是内在的。

戴明的14点管理建议

1. 创建一个愿景并作出承诺。
2. 接受新的理念，学习新的质量原理。
3. 理解检验。要鼓励工人自己负责，把检验作为改进质量的信息收集工具。
4. 停止单纯依据成本做决策。建议企业与少数供应商建立长期的合作关系。仅凭价格就频繁更换供应商加大了原料的变异程度。
5. 持续不断地改进。
6. 开展培训。
7. 进行领导，而非监督。

- 8 驱除恐惧。恐惧包括恐惧报复、恐惧失败、恐惧无知、恐惧失去控制、恐惧变革等。因为担心完不成指标或因为机器的问题受到指责而不去报告质量问题，因为不愿别人得到更高绩效而不愿合作，等等。
- 9 优化团队的努力。部门间的内部竞争、劳资矛盾都破坏质量。培训和雇员参与可以解决这方面问题。
- 10 停止说教。质量问题主要根源于系统而非员工，统计思考和培训才是应有的改进质量途径。
- 11 取消数量定额和目标管理。工人会牺牲质量去完成目标，工人没有动力去改善质量。没有途径的数量指标只会招致挫折和怨恨。
- 12 清除影响人们工作自豪感的障碍。
- 13 鼓励教育与自我改进。要给员工以进步的机会。
- 14 采取行动。

朱兰的理念

- Joseph Juran 1920年代在西方电气公司做统计质量管理，在1950年代想日本人传授质量原理。朱兰更希望在经理们所熟悉的系统内来提高质量。
- 主张使用质量成本核算和分析来关注质量问题。
- 质量定义为“适用性”。
- 质量计划、质量控制、质量改进。
- 质量控制包括确定应当控制什么，建立测量单位以客观地评价数据，确立绩效标准，测量实际绩效，解释实际绩效与标准之间的差异，并对这种差异采取措施。
- 朱兰的质量改进程序包括证明改进的必要性，识别具体的改进项目，组织对于项目的支持，诊断问题的根源，提出纠正措施，验证纠正措施在实际运行条件下的有效性，建立控制措施以巩固成果。

克劳斯比的理念

- 菲利普.B.克劳斯比曾担任国际电话电报公司(ITT)的质量副总裁，并创建质量方面的学院。
- 克劳斯比的质量管理定律：质量意味着符合要求，而非完美；质量问题应该由产生问题的部门来解决；质量无需付费，一次就把事情做好是最节省的；测量绩效的唯一指标就是质量成本，即“不符合”的代价；唯一绩效标准就是“零缺陷”（ZD）。

其他的质量理念

- 费根鲍姆提出全面质量管理。质量领导：管理重点应在计划而非对失败的被动反映。现代质量技术：将工程师、一线员工和办公室职员在过程中整合起来学习新技术解决质量问题。组织的承诺：将质量融入经营计划中。
- 石川馨是日本质量革命的先驱者。推动从高层领导到一线员工的全体员工的广泛参与，主张用简单直观的工具来收集和分析事实数据，运用统计技术和团队精神作为实现全面质量的基础。石川馨的一些观点：质量始于教育，终于教育；质量工作的第一步就是了解顾客的要求；当检验不再必要时就达到了质量控制的理想状态；消除问题的根源而非状态；质量控制是所有员工和所有部门的责任；不要混淆手段与目标；质量第一，要着眼于长远的收益；营销是质量的入口，也是出口；当下级如实汇报事实时，上级不得发怒；公司内95%的问题可以用简单的分析工具和问题解决工具来解决；没有变异性的数据是假数据。
- 田口玄一，阐述了减少变异在经济学上的价值。
- 制造业中的公差限时有缺陷的，如 0.500 ± 0.020 ，人们认为在公差内的任何值都是可行的，并认为只要在公差限内就能保证质量。但是，只有实际值接近于标称值才能得到更好质量，而0.799与0.481并没有本质差别。
- 田口用相对于设计目标值的变异来衡量质量，并把其转化为一个经济“损失函数”，代表用货币表示的变异成本。

质量管理奖与框架——鲍德里奇奖

- 国际上对质量管理最具影响力的框架是美国马尔科姆·鲍德里奇国家质量奖和ISO 9000国际标准认证，近年来，六西格玛也演化成一个独特的质量管理框架。
- 马尔科姆·鲍德里奇国家质量奖的卓越绩效评价准则建立了一个框架，任何一个组织都可以用来整合其全面质量的原则和实践。卓越绩效评价准则包括领导、战略计划、顾客与市场、测量分析和知识管理、人力资源、过程管理、经营结果。给出了详细的条目和重点。质量被“绩效”替代。
- 鲍德里奇计划在2000年投资1.19亿美元，估计汇报为246.5亿美元，回报比例达207:1。

其它质量奖

- 日本于1951年设立戴明奖，公司范围的质量管理(CWQC)是其中一部分。为CWQC建了框架。
- 欧洲质量奖。管理框架称为卓越经营模式。有16个国家。
- 加拿大卓越经营奖。
- 澳大利亚卓越经营奖。

ISO 9000:2000

- 国际标准化组织(ISO)正式通过质量标准。
- ISO 9000 规定了质量体系的标准，它基于这样的认识，即某些普遍的管理惯行(practice)的特征是可以标准化的，一个精心设计、有效实施并仔细管理的质量体系提供了输出将满足顾客期望和要求的信心。
- 指定标准为了满足五方面目标：达到、保持和追求持续改进的符合要求的产品（包括服务）质量；改进运营活动的质量以持续满足顾客和利益相关者明示或隐含的需要；为内部的管理层和员工提供质量满足要求、改进正在实现的信心；为顾客和其他利益相关者提供其产品满足质量要求的信心；提供质量体系要求被满足的信心。

ISO 9000: 2000标准的架构

- 基于全面质量的基本准则，注重于建立、文件化和实施程序来确保生产和服务提供过程运行和绩效的一致性，其目的是实现持续的改进。主要由三个文件构成：
 - ISO 9000 — 基本原理与词汇表；
 - ISO 9001 — 要求，针对质量管理体系的具体要求。包括管理职责、资源管理、产品实现，以及测量、分析与改进。
 - ISO 9004 — 绩效改进指南。不是必须遵守的要求。
- ISO 9000:2000的八项原则：以顾客为关注焦点；领导作用；全员参与；过程方法；管理的系统方法；持续改进；基于事实的决策方法；与供方护理的关系。
- ISO9000认证可能花费数万美元，但许多机构从实施ISO9000中取得了显著的益处，包括实现了更高的顾客满意度和顾客保留率、更高质量的产品、生产率的提高等。

六西格玛

- 六西格玛可以看作一种通过关注对于顾客最为重要的那些输出以及实现组织的财务回报，致力于找出和消除制造和服务过程中的缺陷和差错的经营改进工具。
- 六西格玛通过由团队应用基本的和高级的质量改进和控制工具来实施，这些团队成员要接受相关的培训，能够提供基于事实的决策信息。
- 六西格玛一词是基于统计中正态分布与制造业中公差定义，若制造业中的公差定为 $\mu_0 \pm 6\sigma$ ，则只要制造时真实均值 μ 落在 $\mu_0 \pm 1.5\sigma$ 内，制造的产品超出公差的概率就只有百万分之3.4。意思是在内部指定更高的质量要求。
- 六西格玛的成功使用者包括摩托罗拉、通用电气、德州仪器、联合信号、波音、3M、家得宝(Home Depot)、卡特彼勒、IBM、施乐、花旗、雷神(Raytheon)、美国空军空战指挥中心等。

六西格玛的核心理念

- 在整体战略目标的指导下，从关键业务过程和顾客要求出发来思考问题；
- 重视公司的倡议人，他们负责用户项目，支持团队活动，帮助克服变革阻力并获取资源。
- 重视诸如百万机会缺陷数(defects per million opportunities, dpmo) 这样的定量的测量指标，这样的测量指标可以应用于公司的所有方面，包括制造、工厂、行政、软件等。
- 尽早识别聚焦于业务成果的过程中的指标，据此来提供激励和明确责任。
- 在充分培训的基础上，以项目团队的方式来提高利润，减少非增值活动，实现运转周期的所见。
- 培养能够应用改进工具并领导团队的具有充分能力的过程改进专家。
- 设定富有挑战性的改进目标。

六西格玛的质量框架

- 六西格玛为实施全面质量体系提供了一个蓝图，尤其是人的参与，过程管理方法的应用、变异分析和统计方法、程序化解决方法、基于事实的管理等。与全面质量管理相比的特点为：

- 六西格玛由业务领导者来推动；六西格玛是真正跨职能的；六西格玛使用的工具更先进，依托于一套先进的统计方法以及一个结构化的问题解决方法论DMAIC—定义、测量、分析、改进和控制；六西格玛关注财务底线并要求有一个可验证的投资回报。
- 六西格玛大大提升了统计学和统计思考在质量改进中的重要性。

§5.3 过程管理

过程管理

- 大多数质量问题是过程管理不当产生的，很少有问题是员工产生的。
- 过程管理(process management)涉及对以下这些方面所必需的活动的计划与管理，即实现关键业务过程的高绩效，识别改进质量和营运极小以及顾客满意的机会。
- 常见的业务过程: 了解顾客和市场，战略计划，研究开发，采购，开发新产品和服务，完成顾客订单，管理信息，测量和分析绩效，培训员工，等等。

过程管理的范围

- 领先的公司会在整个价值链上识别影响顾客满意的过程。这些过程通常被分解为两大类，即价值创造过程和支持过程。价值创造过程由外部顾客需要驱动，支持过程则由内部顾客需要驱动。
- 过程管理由三个方面的活动所组成：设计、控制和改进。设计包括识别过程和对过程的文件化。控制是使过程的绩效指标围绕在一个平均水平上下波动，而改进意味着将绩效指标的平均水平提高。
- 过程需要可重复、可测量，从而可以揭示过程绩效的变化模式。
- 简单的质量改进过程包括识别问题、分析原因、选择解决方案、计划实施、评价效果等阶段。

产品设计过程

- 设计的改进可以降低成本、提高质量。
- 典型的产品开发过程包括六个阶段：提出创意、初步概念开发、产品/过程开发、量产、市场导入、市场评价。
- 产品设计会显著影响生产的成本(直接和间接的劳动力、原材料以及管理费用)，影响产品的再设计、保证和现场维修的成本，也会影响产品生产的效率以及输出的质量。

关于设计的一些建议

- 使零件数最少：减少零件和装配图纸，减少复杂的装配，减少必须控制的质量特性，减少失效的零件数。

- 减少零件种类：减少相似零件的变异(降低装配错误率)。
- 稳健性设计(田口方法)：减少对零件变异的敏感性。
- 取消调整：避免装配调整差错，取消效率高的可调整部件。
- 尽量使易于装配并能防误：零件不会被错装，当零件丢失时一眼就可以看出，零件本身设计有装配工艺，零件是自保护的，无须用力装配。
- 应用可重复并易理解的过程：零件质量易于控制，装配质量容易控制。
- 选择能够经受过程作业的零件：减少零件损伤，减少零件降等使用。
- 设计要便于有效而充分地测试。
- 零件布局合理：减少搬运和装配时对零件的损伤。
- 消除以批准产品的设计变更：减少变更差错和多个版本。

过程控制

- 控制是确保符合要求，当需要改正问题时及时改正以维持稳定效能的活动。
- 过程控制的重要性源于：过程管理方法是有效地对过程进行日常管理的基础；如果不能使过程处于控制状态，就谈不上长期的改进。
- 控制系统的三个组成部分：(1) 标准或目标；(2) 测量结果的方法；(3) 将实际结果与标准对比，并以反馈信息作为改正行动的基础。
- 有效的质量控制系统包括：所有关键过程的文件化程序；关于设备和工作环境的明确规定；见识和控制关键质量特性的方法；设备的批准过程；工艺标准，如书面标准、样品或图表；保养活动。

§5.4 绩效测量与信息管理

绩效测量与信息管理

- 有效的运营信息使管理层发现那些项目创造价值、那些项目损失价值。
- **数据** 反映了来自某种测量过程的事实。**测量**(measurement)是对产品、服务、过程及其他业务活动的绩效唯独进行量化的行为。**指标**(measures)和**指征**(indicators)是指测量的数字信息。
- 关于企业各个职能领域的一致、准确而即使的数据为评估、控制和改进过程、产品和服务提供了实时的信息，从而有助于企业实现经营目标并满足不断变化的顾客需要。
- **信息**(information)是关于一项业务或一个组织的数据。信息来源于对数据的分析。

信息的战略价值

- 信息不仅仅在生产过程中需要，在组织战略制定，评估行动计划有效性，持续改进等方面也是需要的。
- 主要做法：
- 建立一套综合的绩效指标，以反映内外部顾客要求以及驱动企业的关键因素。
- 应用比较性的信息和数据，以改进整体绩效和竞争地位。
- 持续改进信息资源及其在组织内的使用。
- 运用合理的分析工具进行分析，并且运用这些分析结果来支持战略计划和日常决策。
- 让每个人都参与测量活动，并且确保绩效信息在整个组织中是广泛可见的。
- 确保数据和信息准确、可信、及时、安全以及保密。
- 确保硬件和软件系统可靠易用，所有需要数据和信息的员工都能得到它们。
- 系统地管理组织的知识，识别并分享最佳惯行。

绩效测量的范围

- 传统的投资回报率等绩效指标不准确，并且强调数量甚于质量。
- 为了使决策实现组织满足或超越顾客期望的总目标，最大限度地利用有限的资源，除了传统的财务绩效和会计指标，企业还需要其他方面的数据和信息，包括顾客与市场、人力资源的有效性、产品和服务的质量，以及其他方面的关键因素。
- 卡普兰和诺顿的平衡计分卡是一页包含了非财务绩效指标和关键财务目标的摘要报告。包含四个方面：财务方面、内部运营、顾客方面、创新和学习方面。

设计有效的绩效测量系统

- 绩效测量系统的目的为：为持续改进提供方向和支持；识别趋势和进展；使因果关系更加易于理解；使得能够与标杆进行绩效对比；认识过去、现在和将来。
- 在设计绩效测量系统时，组织必须考虑使绩效指标能够支持最高管理层的绩效评审，使整个组织的计划活动指向组织的整体健康状况，还要使绩效指标支持日常的运营和决策。

设计绩效测量系统的使用建议

- 指标越少越好。集中于测量关键的少数变量，而非无关紧要的多数变量。
- 测量指标应当与关键成功要素联系起来。
- 测量指标应当包含过去、现在、未来的组合以确保组织考虑到所有三个方面。
- 测量指标应当以顾客、股东和其他重要利益相关者的需要为根据。
- 测量应当自上而下展开。
- 多重指标可以整合成单一指标以更好地全面评价绩效。
- 指标应当根据研究而非主观臆断来确定目标。

过程层面的测量

- 良好的测量指标需要SMART: 简单(Simple)、可测量(Measurable)、可执行(Actionable)、与顾客需求或者其他方面相关联(Related)、及时(Timely)。

- 产品和服务的质量指标聚焦于制造和服务过程的结果。常用的制造质量的指标是单位不合格数量(nonconformities per unit) 或单位缺陷数(defects per unit)。在服务中使用每个机会的差错数(errors per opportunity)。通常用千分率或百万分率, 如每百万机会缺陷数(defects per million opportunities, dpmo)。
- 缺陷分为严重、重要、轻微, 可以用不同权重综合为一个缺陷指标。

绩效数据的分析和使用

- 数据必须经过分析转化为信息才能被高层利用。分析值得是为了提供有效决策的基础而对事实和数据进行的检查。比如:
- 检查关键绩效指标的 trends 和变化;
- 与其他的业务单位、竞争对手或行业最佳标杆的绩效作比较;
- 计算均值、方差以及其他的统计量;
- 使用相关分析和回归分析等复杂的统计工具找出不同绩效指标间的关系。
- 使用Excel之类的图表汇总就能得到许多有用的结论, 使用高级的统计分析软件可以有更好的经营结果。

关联分析

- 管理者必须理解关键业务绩效指标之间的联系。如
- 产品和服务质量的改进与关键的顾客指标如何相关, 这些指标包括顾客满意度、顾客保有率、市场份额等。
- 员工的安全、缺勤及流失等方面的改进所带来的财务方面的益处。
- 教育和培训的益处和成本。
- 产品和服务的质量、运营绩效指标和总的财务绩效之间的关系。
- 顾客满意度和顾客保有率对利润的影响。
- 顾客满意度的变化导致的市场份额的变化。
- 员工满意度对顾客满意度的影响。

关联分析的好处

- 找出薄弱的或误导性的绩效指标。
- 使管理层的注意力集中在确实能带来变化的关键绩效指标上。
- 预测顾客满意水平之类的绩效。
- 设立绩效目标标准。
- 促使营销和运营这些领域协调它们的数据分析活动。
- 比竞争对手更快地作出明智的决策。
- 看出被竞争对手忽略的绩效指标之间的关系。
- 在良好的数据分析和基于事实的管理的基础上加强组织的沟通。

数据挖掘

- 数据挖掘是从海量企业数据中发现未知关系的计算机技术。
- 包括聚类分析、判别分析、判别树、支持向量机、神经网络、关联分析等方法。
- 其应用相对便宜，但也存在过多的输出结果需要人为解释，不能产生因果性结论等缺点。

质量成本

- 质量成本(Cost of quality, COQ)传统定义为检验方面的成本，但实际上与质量有关的成本可以占到销售收入的20-40%。
- 与质量相关的成本不仅与生产活动有关，与采购和顾客服务等辅助部门也有关系。
- 大部分的成本来自不良的质量，是可以避免的。
- 所以，质量成本指的是避免不良质量或由不良质量而产生的后果的成本。
- 质量成本的方法需要把质量成本转换为高层易用的金钱的语言。

质量成本的组成

- 预防成本：避免不合格产品的发生以及防止不合格产品流入顾客手中而进行的投资。包括质量计划成本、过程控制成本、信息系统成本、培训及一般管理成本。
- 鉴定成本：主要是试验和检验成本、维护检验仪器的成本、过程测量和控制成本。
- 内部故障成本：包括废品和返工成本、纠正措施成本、降级成本、过程故障。
- 外部故障成本：包括顾客投诉和退货成本、产品召回和维护成本、产品责任成本。
- 增加预防成本比检验确保更有效。
- 质量成本的各个类别很少会平均分布，和可能70%–80%的内部故障成本仅仅是由一两个制造问题导致的。

§5.5 六西格玛的原理

§5.5.1 六西格玛的统计基础

六西格玛介绍

- 通过聚焦于对顾客最为重要的特性，识别和消除过程中的差错或缺陷的原因，六西格玛已经由一个单纯的质量测度指标演化成一套加速改进、实现前所未有绩效水平的综合测量。
- 六西格玛方法论旨在把组织的关键产品和过程的缺陷水平降至百万分之几的程度。
- 为此必须有效地应用统计原理和各种工具来诊断质量问题并促进改进。

六西格玛的统计基础

- 顾客所接收到的任何错误或差错都被称为缺陷(defect)或不合格(nonconformance)。某个过程或过程步骤的产出称作单位产出(unit of work)。单位产出缺陷数(defects per unit, DPU)便是衡量质量的一个指标：

$$\text{单位产出缺陷数} = \frac{\text{所发现的缺陷数}}{\text{所生产的产出单位数}}$$

- 这样的指标只针对最终产品，而非生产过程。也难以用于复杂过程，特别是服务过程。
- 不同过程出错的几乎数有很大差别。
- 定义百万机会缺陷数(defects per million opportunities, dpmo)

$$\text{dpmo} = \frac{\text{所发现的缺陷数}}{\text{出错机会数}} \times 1,000,000$$

六西格玛指标的统计来源

- 六西格玛意味着dpmo最多为3.4这样一个质量水平。
- 设产品设计公差为 $[\mu_0 - 6\sigma, \mu_0 + 6\sigma]$ 。实际生产产品指标 $X \sim N(\mu, \sigma^2)$ 。生产中，要求控制 $\mu \in [\mu_0 - 1.5\sigma, \mu_0 + 1.5\sigma]$ 。这时， $P(X \notin [\mu_0 - 6\sigma, \mu_0 + 6\sigma]) = 3.4\text{dpmo}$ 。
- 当过程(X)标准差(σ)的6倍等于设计公差(12σ)的二分之一，且其均值偏离目标值最大不超过1.5个标准差时，即为六西格玛质量水平。

- 仿照以上设定，若过程标准差为 σ ，公差的二分之一为 $k\sigma$ ，过程均值 μ 偏离设计均值 μ_0 a 个标准差，dpmo的计算公式为

$$\begin{aligned} \text{dpmo} &= P(X > \mu_0 + k\sigma | \mu = \mu_0 + a\sigma) \times 10^6 \\ &= P\left(\frac{X - \mu_0 - a\sigma}{\sigma} > \frac{k\sigma - a\sigma}{\sigma}\right) \\ &= \Phi(a - k) \end{aligned}$$

- 所以，dpmo=3.4可以由不同的中心偏移倍数 a 和半公差倍数 k 组成。比如， 0.5σ 中心偏移与 5σ 质量水平组合， 1.0σ 中心偏移与 5.5σ 质量水平组合， 1.5σ 中心偏移与 6σ 质量水平组合。
- 在产品公差不变条件下，提高质量水平 $k\sigma$ 意味着减小产品变异性，在产品变异性不易减小的情况下，可以控制过程均值(μ)更接近设计值(μ_0)，即减小 a 。

dpmo与偏移量、质量水平的关系编程

```
aarr <- seq(0, 2, by=0.25)
karr <- seq(3, 6, by=0.5)
dpmo.tab <- outer(aarr, karr,
                  function(a, k) pnorm(a-k)*1E6)
rownames(dpmo.tab) <- format(aarr)
colnames(dpmo.tab) <- format(karr)
print(dpmo.tab, digits=6)
```

dpmo与偏移量、质量水平的关系表格

a	k=3.0	k=3.5	k=4.0	k=4.5	k=5.0	k=5.5	k=6.0
0.00	1350	233	32	3.4	0.3	0.019	0.001
0.25	2980	577	88	11	1.0	0.076	0.004
0.50	6210	1350	233	32	3.4	0.290	0.019
0.75	12224	2980	577	88	11	1.0	0.076
1.00	22750	6210	1350	233	32	3.4	0.290
1.25	40059	12224	2980	577	88	11	1.0
1.50	66807	22750	6210	1350	233	32	3.4
1.75	105650	40059	12224	2980	577	88	11
2.00	158655	66807	22750	6210	1350	233	32

当 $k - a = 4.5$ 时, dpmo恰好为3.4。

dpmo的推广应用

- dpmo和六西格玛最初用于制造业,后来把百万机会缺陷数推广到一般过程,如产品开发、新业务收购、顾客服务、会计等。
- 固定 $a = 1.5$,可以用 k 来衡量质量水平。比如,某银行经常账户报告中的错误数,在1000份报告中发现了12个差错, dpmo=12,000,质量水平在3.5~4西格玛之间,与6西格玛质量水平差距巨大。如果手机系统的质量水平是4西格玛,相当于每个月有4个多小时无法通话。一个4西格玛水平的送货车每三辆车就有一个错包,而6西格玛水平则每5000辆车才有一个错包。
- 并非所有过程都需要在六西格玛水平下运行。从2或3西格玛水平提高到4西格玛水平相对容易,进一步提高则需要付出更多努力并需要应用更加高级的统计工具。

§5.5.2 六西格玛项目的选择

六西格玛项目的选择

- 质量人员一开始可以解决一个对于顾客或经营绩效确实具有影响的业务问题,成功完成一个基本的六西格玛项目。技能提高后可以解决更大范围的问题。
- 选择合适的问题是六西格玛活动中最困难的挑战之一。选择项目应有好的财务汇报和较高的成功可能性。可以用一些简单的评分过程对项目排序。

提出问题

- **问题**指应该达成的状态与实际达到的状态之间的差异,且这种差异重要得足以使某人认为应当加以纠正。如
- 符合性问题。系统本来正常运行,但由于某种原因失常,造成质量或客户服务水平下降。需找出失常的原因。
- 非结构性的绩效问题。指的是没有“正常状态”的明确定义。比如,“销售不良”。需要用创造性方法解决。
- 效率问题,如成本和生产率不令人满意,虽然质量可接受。
- 产品设计问题。要设计达到和超越顾客期望的产品。
- 过程设计问题。比如设计全新的过程或对现有过程的重大修改。

§5.5.3 六西格玛的问题解决

问题解决

- 问题解决(problem solving)是指将实际发生的状态转变为应当达成的状态所涉及的相关活动。
- 六西格玛项目的目标通常是使组织达到前所未有的绩效水平,通过实施系统化的问题解决措施为组织和顾客增加价值。
- 成功的质量改进和经营绩效改进取决于识别和解决问题的能力。要求管理者具有定量思维能力,用统计的方法解决问题。
- 六西格玛解决问题的方法论是DMAIC, define(定义), measure(测量), analyze(分析), improve(改进), control(控制)。
- 一个制度化的问题解决过程为全体雇员彼此间,尤其作为跨职能团队的成员间的沟通提供了共同的语言和工具。

DMAIC方法论

- 定义。要具体,这区别于提出问题。确定项目范围。明确对绩效影响最大的关键质量问题(CTQ),确定相关的绩效指标,核算成本,确定应达到的标准。
- 测量。测量影响CTQ的内部过程。必须理解造成质量问题的内部过程的因果关系。
- 分析。着重于缺陷、差错或过度的变异所发生的原因。
- 改进。要求举要高度的创造性,不怕犯错误,有名的方法如“头脑风暴法”。
- 控制。维持已经实现的改进。包括在修正的过程中采取措施确保关键变量保持在可接受范围内。可以用核对表或定期状态评审表确保改进持续有效,或用统计过程控制图监测。

分析的方法

- 分析缺陷的原因,通常表现为:
- 缺乏关于过程如何运行的知识。
- 缺乏关于过程应当如何运行的知识,包括对顾客的期望和过程目标的理解。

- 缺乏对过程中应用的材料和设备的控制。
- 工作中的疏忽造成的差错。
- 浪费和复杂性，如过程中的不必要步骤、过量的库存等。
- 零部件的设计和生產不够细致，设计规范不良，进货和原型的检验不充分。
- 不理解满足规范要求的过程能力。
- 缺乏培训。
- 仪器仪表的校准测试不良。
- 环境特性不佳，如光照、温度、湿度、净度、噪音等。

根本原因

- 为纠正缺陷，需要确定最优可能造成差错和过度变异的关键因素，称为**根本原因**(root causes)。根本原因定义为“导致某种缺陷发生的条件(或相关联的一系列条件)，这些条件一旦得到适当的纠正，就能够永久地避免该缺陷在过程产出的同类或随后的产品或服务中再次发生”。
- 查找根本原因有一种问“五个问什么”的方法。“五”是泛指。以一个因果链来不定地定义问题，探究症状的源头。
- 例如，丰田的五个为什么的例子。一台机器经常由于保险丝熔断而停机。换一根保险丝不解决根本问题。为什么保险丝熔断？因为轴承润滑不充分。为什么轴承润滑不充分？因为润滑泵有毛病。为什么润滑泵有毛病？因为泵轴有磨损。泵轴为什么磨损？因为有粉末渗漏到了轴上，这才是问题的根本原因。在润滑泵上安装了一个防止粉末深入的滤网解决了问题。

工具和方法

- DMAIC方法论的两个特征：对顾客要求的重视；统计工具和方法的应用。
- 六西格玛把传统的统计工具应用从工程领域推广到了全面质量，六西格玛的课程包括技术性内容、项目管理和领导方面的内容的混合。
- 课程内容示例：
 - 初级统计工具(描述统计、统计思考、假设检验、相关分析、简单回归)；

- 高级统计工具(试验设计、方差分析、多元分析);
- 产品设计和可靠性(质量功能展开、失效模式和影响分析);
- 测量(过程能力、测量系统分析);
- 过程控制(控制方案、统计过程控制);
- 过程改进(过程改进的计划、过程测绘、防误措施);
- 实施和团队活动(组织有效性、团队评估、促进工具、团队发展)。

§5.6 统计思考和应用

统计思考

- 统计学是一门设计“收集、组织、分析、解释以及呈现数据”的科学。
- 从现场搜集到的原始数据并不能提供质量控制和质量改进所必需的信息，还必须对数据加以组织、分析和解释。统计学为从数据中获取有意义的信息提供了有效的途径，从而使得管理人员和员工能够控制和改进过程。
- 统计概念在质量管理中的重要性再怎么强调都不过分。必须理解统计学在管理决策中所扮演的重要角色。
- 统计思考(statistical thinking)是基于以下原则的学习和行动哲学：
 - 所有工作发生在有相互联系的过程组成的系统中。
 - 变异存在于所有过程中。
 - 了解和减少变异是成功的关键。

变异

- 在致力于减少变异前，必须理解变异的本质。
- 任何生产过程都有变异源。例如，不同批次的材料在强度、厚度等方面不同。切割工具在强度和成分方面存在内在变异。制造过程中，工具会被磨损，震动会引起机器参数发生变化。电压波动会使动力发生变化。操作工人无法确保每次都将零部件一模一样地放入工作夹具内，身体和情绪的压力也会影响工人操作的一致性。测量量具和人的监测能力也不能确保始终如一。
- 单个的变异是随机的，但所有变异源共同发生作用时，其“分布”服从稳定的规律，能通过统计进行解释和预测。

变异的原因

- 产品和生产系统的设计决定了变异的一般性原因(common causes)。由一般性原因产生的变异通常占生产过程产出的变异的80%~95%。
- 生产过程的其他变异是由特殊原因(special causes)引起的，称为变异的可归因原因(assignable causes)。特殊原因是由不属于过程固有的外部源引起的，表现出偶然性，破坏了一般性原因的随机模式，使用统计方法可以识别它们，纠正的方法通常也很经济。

- 只是由一般性原因影响的系统称为稳定系统。理解稳定系统以及变异的特殊原因和一般性原因的差异,对于管理任何系统而言都是至关重要的。
- 要区分一般性原因和特殊原因,分别采取不同的处理措施。对两类原因分辨错误不能解决问题,并增加了不必要的变异。
- 对变异的原因分析不明,采取错误的措施可能会增大变异。

变异带来的运营问题

- 变异增加了不可预测性。不了解系统变异就很难预测系统未来的绩效。
- 变异减少了产能利用。为了完成计划经理们会增加生产负荷,不合理地使用了产能。
- 变异会导致相关部门的不必要的调整。
- 变异掩盖了根本原因。
- 变异使人们很难在早期发现潜在的问题。因为数据本身就是在随机波动的,所以发生错误引起的波动容易被忽略。

统计基础

- 随机变量和概率分布。大部分生产和商业过程不服从正态分布。
- 抽样。包括简单随机抽样,分层抽样,系统抽样,集群抽样,判断抽样等。一个好的抽样方案应当以最低的成本抽取到一个最能代表总体特种的样本,并且要符合该研究所确定的精确度和可靠性目标的要求。
- 不合理设计的抽样造成系统误差,包括:总是偏高或偏低(bias);非可比数据;不严格的趋势预测;武断地从相关推出因果;不合适的抽样。

统计方法

- 描述统计(descriptive statistics)和统计推断(statistical inference)。
- 数据收集。
- 描述统计包括中心位置和分散程度,频数分布,直方图。
- 预测性统计,如相关分析、回归分析、判别分析、聚类分析等。
- 统计推断如置信区间、假设检验、方差分析、实验设计。
- 要区分静态总体(样本观测之间独立同分布)和动态总体(时间序列)。
- 试验设计与方差分析可以定位对关键指标有真正影响的因素并找到最好水平。

§5.7 六西格玛设计

六西格玛设计

- 现在产品的质量和可靠性已经比二次大战时期有了长足的进步，但故障和服务混乱依然存在。
- 大多数问题是由于设计不良或设计过程的不充分造成的。
- 六西格玛设计(design for six sigma, DFSS)是用于产品开发过程中的一整套工具和方法论，目的是确保产品和服务满足顾客需要和实现性能目标，确保用于制造产品和服务的过程达到六西格玛能力。
- DFSS 包括四个主要活动：概念开发、设计开发、设计优化、设计验证。

§5.7.1 概念开发中的工具

概念开发中的方法论

- 概念开发(concept developement)是指英语科学的、工程的和商业的知识，创建一个既满足顾客需要有满足生产或服务提供要求的基本功能设计的过程。具有高度创造性，可以用“头脑风暴法”这类做法。先识别潜在的创意，再运用成本收益分析、风险分析以及其他方法来对这些创意进行评价，最后，依据加权评价矩阵来确定最佳概念。
- 一个基本的功能设计开发包括将顾客需要转化为可测量的技术要求，以及进一步转化为详尽的设计规范。
- 满足技术要求的两个工具：质量功能展开和概念工程。

质量功能展开

- 功能设计的难点之一是顾客与工程师使用不同的语言。
- 日本人提出质量功能展开(quality function deployment, QFD)方法，用以在整个设计过程和生产系统的设计中满足顾客要求。QFD是一个计划过程，以顾客要求知道整个组织，使用一种矩阵图来展示数据和信息。
- 在QFD指导下，驱动公司所有运营活动的都是顾客之声，而非高层管理者的命令，或者设计工程师的观点和愿望。在设计方面，QFD始终贯彻顾客需要，而不是先设计出样品再根据顾客反映改进。
- QFD 通过改善价值链上的所有参与者之间，如营销和设计之间、设计和生产之间，以及采购和供应之间的沟通和团队合作，而使公司收益。
- QFD使用质量屋工具来展示数据，评估计划。

概念工程

- 概念工程(concept engineering, CE)是发现和利用顾客要求,从而对满足顾客要求的优越产品和服务概念进行选择的聚焦过程。
- 该过程包括以下五个步骤:
 - 1. 理解顾客的环境。
 - 2. 把理解转换为要求。
 - 3. 把上述内容变成可操作的。
 - 4. 概念的产生。
 - 5. 概念选择。

§5.7.2 设计开发中的工具

产品规范

- 在详细设计过程中,产品规范是指把设计者的概念转变为生产设计并确保产出的经济、有效和高质量。
- 比如,微处理器生产,尺寸规范“0.514/0.588”表示了尺寸允许的范围是 $[0.514, 0.588]$,中间值0.551就是设计指标,容差为 ± 0.037 ,也可以写为 0.551 ± 0.037 。
- 生产规范有公称尺寸和容差构成。公称尺寸(nominal)是指生产寻求达到的理想尺寸或目标值。容差(tolerance)是指考虑与目标保持完全一致有难度而允许的偏差。
- 容差设计主要是确定尺寸允许的偏差。容差过窄增加生产成本,过宽则降低产品性能、耐用性和外观质量。
- 容差指定经常没有考虑到变异对产品性能、生产性和经济效果的影响。推挤推断、回归、实验设计等在设计开发中起到重要作用。
- 新介绍设计失效模式和影响分析、可靠性预测。

设计失效模式和影响分析

- 设计失效(故障)模式和影响分析(design failure mode and effects analysis, DFMEA)是指识别失效可能出现的所有方式,估计失效的影响和严重程度,从而建议采用正确的设计行动。组成:
 - 失效模式—每个要素或功能出现故障的方式。

- 失效对顾客的影响。
- 严重程度、出现的可能性和检测分级。
- 失效的潜在原因。故障通常源于不良设计，设计缺陷会在现场或使用现场或生产组装中引发问题。识别原因可能需要试验和严格的分析。
- 纠正行动和控制。控制可能包括设计改变、防误、更好的用户手册、管理责任和目标完成日期等。

§5.7.3 可靠性预测

可靠性预测

- 可靠性质产品在预期时间内正常运行的能力，是主要的质量指标之一。
- 可靠性是产品和过程设计的基本方面，高可靠性也为产品提供竞争优势。
- 生产过程中，自动化加大使用、设备的复杂性、微薄的利润，以及基于时间的竞争等使得生产过程的可靠性已经称为企业生存的关键议题。

可靠性定义

- 可靠性是一个在 $[0, 1]$ 区间取值的概率，比如，称产品1000小时可靠性为0.97，意思是100件产品运行1000小时平均有97件仍正常工作。
- 同样的可靠性，运行时间长的性能更好。
- 定义可靠性需要有失效的明确定义。产品和系统的失效分为功能失效和可靠性失效。可靠性失效是在运行了一段时间后的失效。
- 可靠性概念还需要指定运行环境。
- 通过定义产品的预期环境、性能特征、寿命等，生产者能够用设计和试验来先测量产品运行(或失败)的概率。试验分析能够更好地预测可靠性并改进产品和过程的设计。

可靠性测量

- 可靠性指在所考虑期限内的不失效比例。
- 设随机变量 $X > 0$ 为寿命， $F(t) = P(X \leq t)$ 是 X 的分布函数， $R(t) = P(X > t) = 1 - F(t)$ 是在工作 t 时间后的可靠性，称 $R(t)$ 为可靠性函数。
- 可靠性函数 $R(t)$ 与一个失效率函数一一对应：

$$h(t) = \frac{R'(t)}{R(t)}$$

是假定已经工作了 t 时间后，单位时间内失效的概率。

- 如果 X 服从指数分布, 其分布函数为 $F(t) = 1 - e^{-\lambda t}$, 可靠性函数为 $R(t) = e^{-\lambda t}$, 失效率函数恒定为 λ 。
- 对于不能维修的产品, 失效率的倒数称为失效平均时间(MTTF), 对于可维修的产品称为失效间平均时间(MTBF)。
- 指数分布的MTBF为 $1/\lambda$ 。
- 指数分布的失效率函数是常数, 许多电子原件通常在寿命周期早期表现为高而递减的失效率, 而后是一段相对低而恒定的失效率, 最后是递增的失效率。
- 计算可靠性指标可以数量化与可靠性有关的成本, 比如担保期的成本。

系统可靠性

- 串联、并联或混合连接而成的系统的可靠性可以通过各组成部分的可靠性计算。
- 串联系统不失效要求所有分系统不失效, 所以

$$R = R_1 R_2 \dots R_k$$

, R 为系统可靠性, $R_1, \dots, R_k$ 为串联的各单元的可靠性。

- 并联系统不失效只要至少一个分系统没有失效, 所以可靠性公式为

$$R = 1 - (1 - R_1)(1 - R_2) \dots (1 - R_k)$$

即不是所有分系统都失效的概率。

- 混合连接可以拆分计算。

§5.7.4 过程优化中的工具

过程优化中的工具

- 设计优化包括设定适当的容差以确保产品性能最优化, 使设计稳健(robust), 及对生产和使用环境不敏感。
- 工具有田口损失函数、可靠性优化等。

田口损失函数

- 田口损失函数(Taguchi loss function)是容差设计的一个科学方法。田口认为原来人们认为产品指标只要落在公差内就可行的想法是错误的,损失应该是一个二次函数,实际指标离设计值越远,尽管在公差内,损失越大。损失函数为

$$L(x) = k(x - T)^2$$

其中 x 是质量特性的实际值, T 是目标值, k 是常数。

- 常数 k 通过偏差相关成本加以估计。例如,某质量特性规范是 0.500 ± 0.020 ,如果超出范围,产品可能需要维修,费用为50美元,则 k 的估计方法为:

$$50 = k(0.020)^2$$

$$k = 50/0.0004 = 125000$$

平均损失

- 如果已知实际特性值作为随机变量的分布, $X \sim f(x)$, $f(x)$ 为密度或概率函数,则平均损失定义为

$$L = kE(X - T)^2 = k \int (x - T)^2 f(x) dx$$

- 平均损失用 X 的方差 σ^2 可以表示为

$$L = kE(X - T)^2 = k\sigma^2 + k(EX - T)^2$$

损失分为两个部分,由生产过程的均值与目标不一致造成的损失,和生产过程的变异性造成的损失。

单边的损失函数

- 越小越好的特性指标的损失函数可取为

$$L(x) = kx^2$$

- 越大越好的特性指标的损失函数可取为

$$L(x) = k/x^2$$

可靠性优化

- 优化可靠性的工具有:
- 标准化。使用已有多年的可靠性记录的原件,采用成熟的操作规范。
- 冗余。提供易损部件的备件,或进行并联设计。
- 失效的物理原因。

§5.7.5 设计验证中的工具

设计验证中的工具

- 设计验证有助于确保设计满足顾客需要并且使生产符合规范。
- 可靠性检验：可靠性需要试验确认，包括模仿环境条件确定产品特性、运行时间和失效模式。寿命加速试验(accelerated life testing)在恶劣或繁重工作条件下试验以提早发现问题。元件加强试验是指把集成电路暴露在高温中从而强迫潜在缺陷出现。潜在缺陷是指早期失效可能性较大，早期没有失效则可以很长时间以高可靠度运行。
- 测量系统评价。数据误差中一部分是测量误差，包括系统性误差（总是偏高或总是偏低），称为偏差(bias)，和随机误差。准确度(accuracy)衡量偏差大小，精确度(precision)衡量随机误差大小。

过程能力评价

- 过程能力(process capability)是指当系统有一般因素主导，及过程处于稳定状态时过程偏差的范围。
- 过程能力不仅对产品设计者和生产工程师重要，也是实现六西格玛绩效的关键。
- 了解过程能力有助于定量预测过程满足规范的程度，也有助于指定设备规范要求和需要的控制水平。
- 过程能力研究(process capability study)是指在具体运行条件下提供过程特性的详细信息的一项有计划的研究。要回答的典型问题包括：
 - 过程的中心在那里？
 - 过程的偏差多大？
 - 相关规范特性可否接受？
 - 多少输出比率将满足规范要求？
 - 偏差的原因是什么？
- 用频率分布、直方图和控制图评价过程能力。
- 在正态分布假定下，从样本中得到参数 μ 和 σ 的估计可以用来计算过程能力。

- 非正态时, 用分组频数统计和直方图直接估计绩效特性落入公差范围内的比例。
- 因为产品质量计算需要总体标准差 σ , 直接用样本标准差 S 代替会错误估计缺陷率。可以通过 S 和样本量 n 给出 σ 的置信下限和置信区间。 σ 的置信度为 $1-\alpha$ 的置信下限的R计算公式为 $\text{sqrt}((n-1)*S^2/\text{qchisq}(1-\alpha, n-1))$, 置信区间的下、上限分别为 $\text{sqrt}((n-1)*S^2/\text{qchisq}(1-\alpha/2, n-1))$ 和 $\text{sqrt}((n-1)*S^2/\text{qchisq}(\alpha/2, n-1))$ 。

过程能力指数

- 过程输出的分布因为相对与规范的位置(均值)和分布范围(标准差)而不同。自然变异和规范之间的关系可以用过程能力指数(process capability index) 加以量化:

$$C_p = \frac{UTL - LTL}{6\sigma}$$

其中[LTL, UTL]为容差下、上限, σ 为生产过程的标准差。给定公差后, 此指数越大, 要求控制的变异性越低, σ 要求越小。

- 实践中如果过程的实际均值漂移会造成公差范围超界增加, 一般建议 C_p 的安全的较低界限是1.5。许多公司要求供应商的 C_p 值等于或大于1.66。

一侧指数

- 以上讨论假定实际生产的均值 μ 等于目标值 μ_0 。要包括过程中新的信息, 经常使用一侧指数。

- 上侧指数

$$C_{pu} = \frac{UTL - \mu}{3\sigma}$$

- 下侧指数

$$C_{pl} = \frac{\mu - LTL}{3\sigma}$$

- 一侧指数

$$C_{pk} = \min(C_{pu}, C_{pl})$$

- 如果生产过程的 σ 很小, 那么有可能在 $\mu - \mu_0$ 已经较大的情形下 X 还落在公差范围内, 按照田口损失函数这样是不良状态, 所以 C_p 和 C_{pk} 不是唯一指标。
- 非正态分布时, 过程能力指数没有意义。

§5.8 过程改进工具

§5.8.1 过程改进的方法论

过程改进的方法论

- 六西格玛的DMAIC给出了一种方法论。
- 这里介绍其他流行方法，如戴明环。

戴明环

- 戴明环分四个阶段进行，即计划、执行、学习和行动(PDSA)，围绕“顾客满意”。
- 计划阶段包括研究当前环境和描述过程：投入、产出、顾客和供应商；理解顾客期望；收集数据；明确问题；测试成因理论；指定解决方案和行动计划。
- 执行阶段，计划在试验的基础上试行，比如在实验室中试验生产过程，或者与一个小组顾客合作，评估某个特定的解决方案并提供目标数据。试验得出的数据被收集和记录下来。
- 学习阶段通过评估结果、记录学习、决定是否有进一步可以采取的行动和可能的机会，来判断试验中的计划是否正确。通常第一个计划需要修改。修改后返回到执行阶段去评估。
- 在最后一个阶段行动中，改进变得标准化，称为“当前最佳惯行”。
- 回到计划阶段寻找其他改进机会。
- 戴明环既重视短期的持续改进，也重视长期的组织学习。

§5.8.2 过程改进的基本工具

过程改进的基本工具

- 六西格玛用于过程改进的工具具有：
- 流程图。用于定义和控制。
- 运行图和控制图。用于过程控制。
- 检查表。用于测量和分析。
- 直方图。用于测量和分析。
- 帕累托图。用于分析。

- 因果图。用于分析。
- 散点图。用于分析和改进。
- 日本人称为“全面质量的七种工具”(Seven QC tools)。

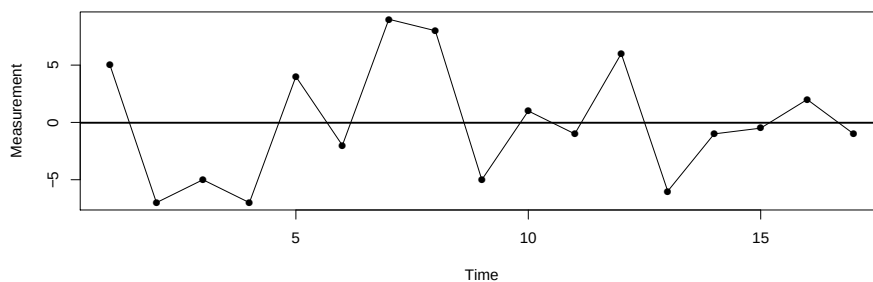
流程图

- 流程图画出一个过程的行动和次序。
- 流程图最好由身处过程之中的人——员工、主管、经理和顾客——来共同绘制。
- 流程图(flowchart map)或过程地图(process map)明确了一个过程中的活动次序或原材料和信息的流动方向。通过提供一幅完成任务所需的图画,帮助人们更好、更客观地了解过程。
- 流程图可以用回溯式策略绘制,从生产的产出开始向回提问。
- 流程图可以提供质量改进的方向,简化过于复杂流程可以提高绩效。

运行图和折线图

- 运行图(run chart)是一幅折线图,其中的数据按照时间顺序排列。又称时间序列曲线。
- 纵轴代表测量指标,横轴代表时间。
- 构造运行图,需要先确定要监测的测量指标。定期抽样为计算基本的统计指标如均值、标准差、次品率提供了数据。
- 运行图以一种易于理解和解释的图形方式,展示随着时间推移,某个过程或某些质量指标的绩效和偏差。它还能够指出过程随时间的变化和趋势,以及校正行动的效果。

运行图示例



构造运行图的步骤

- 收集数据，计算需要的统计指标，如平均值、标准差、缺陷率。
- 检验数据范围，确定纵坐标的限值，为新数据留出余地。
- 绘图，连线。
- 计算平均值，画一条水平线，称为图形的中心线(CL)。

运行图示例的程序

```
demo.run <- function(use.pdf=FALSE){  
  if(use.pdf){  
    pdf("run-chart.pdf", width=10, height=4)  
    on.exit(dev.off())  
  }  
  x <- c(5, -7, -5, -7, 4, -2, 9, 8, -5, 1,  
        -1, 6, -6, -1, -0.5, 2, -1)  
  mu <- mean(x)  
  plot(seq(along=x), x, type="p",  
        pch=16,  
        xlab="Time", ylab="Measurement")  
  lines(seq(along=x), x)  
  abline(h=mu, lwd=2)  
}  
demo.run()
```

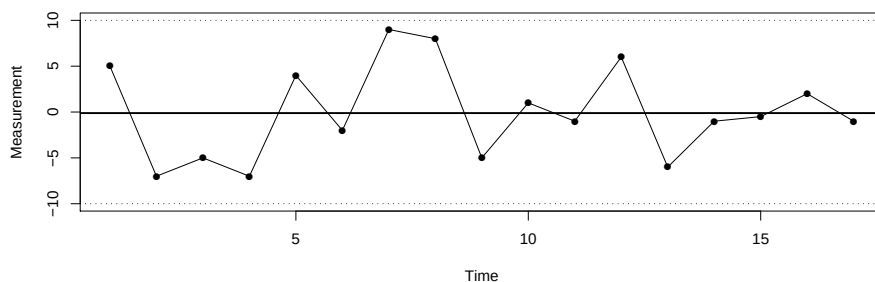
运行图的判读

- 如果数据点围绕中心点以一种稳定的随机模式波动，没有大的峰谷、走向或唯一，则说明过程是明显受到控制的。
- 如果存在异常的模式，则需要调查造成稳定性缺乏的原因，并采取校正行动。
- 运行图能够明确由于缺乏控制导致的混乱。

控制图

- 控制图仅在运行图基础上加入了代表控制上限(upper control limit, UCL)和控制下限(lower control limite, LCL)的两条水平线。
- 控制限根据统计原则选择,可以保证稳定状况下数据有99%落在控制限内。
- 如果样本值落在控制限以外,或者图中存在非随机的模式,则可能存在影响过程的特殊原因,过程是不稳定的。需要检查过程,采取适当的校正行动。
- 如果评估和校正同时进行,过程稳定性就能保持,质量得到保证。
- 控制图使得操作者能够在质量问题发生时就发现问题。

控制图示例



控制图示例的程序

```
demo.control <- function(use.pdf=FALSE){
  if(use.pdf){
    pdf("control-chart.pdf", width=10, height=4)
    on.exit(dev.off())
  }
  UCL <- 10; LCL <- -10
  mu <- mean(x)
  x <- c(5, -7, -5, -7, 4, -2, 9, 8, -5, 1,
        -1, 6, -6, -1, -0.5, 2, -1)
  plot(seq(along=x), x, type="p",
       pch=16, ylim=range(c(LCL, UCL, x)),
       xlab="Time", ylab="Measurement")
}
```

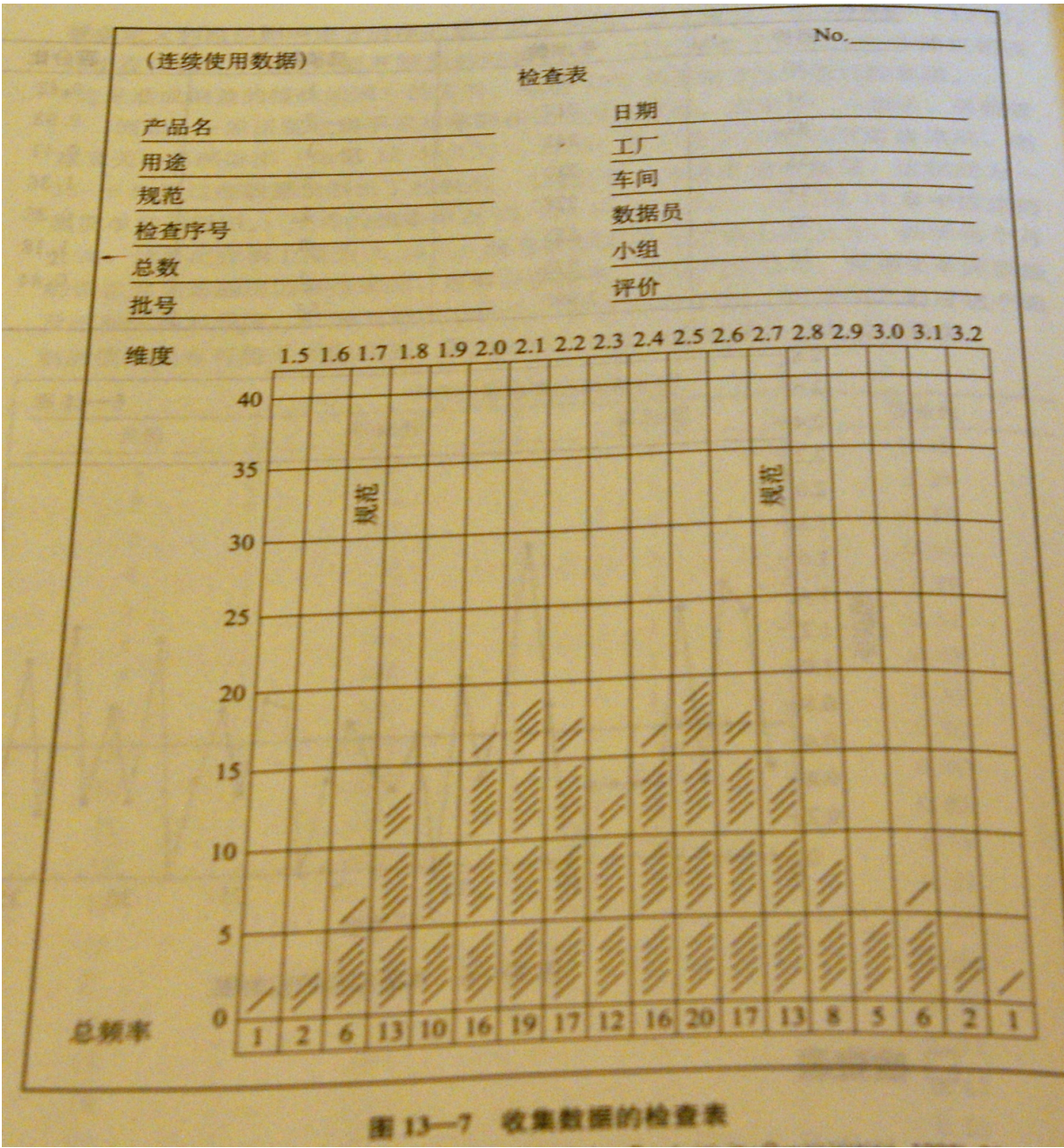


```
lines(seq(along=x), x)
abline(h=mu, lwd=2)
abline(h=c(UCL, LCL), lty=3)
}
demo.control()
```

检查表

- 检查表(check sheets)是搜集数据的简单工具。
- 在制造业中，类似下图的检查表能够很容易地被销售人员使用和解释。表中包含了规范限制等信息，使得不合格品的数量更容易被发现，并提供了关于过程质量的一个即时指标。例图中有较多次数超出规范，且超过上限情况更严重。

收集数据的检查表



不合格品检查表示例

- 下图记录一家数值生产厂商的不合格品的类型和数量。
- 能够被扩展到包含时间维度，从而可以随时间推移检测和分析数据，如果存在还可以发现趋势和模式。

检查表

产品: _____

制造阶段: 最后检查

缺陷类型: 划痕, 不完整, 畸形

检查总数: 2 530

评价: 全部检查品

日期: _____

工厂: _____

部门: _____

检查员: _____

批号: _____

订单号: _____

类型	检查	小计
表面划痕	### ### ### ### ### ### //	32
凹陷	### ### ### ### ///	23
不完整	### ### ### ### ### ### ### ### ### ///	48
畸形	////	4
其他	### ///	8
	总计	115
次品总数	### /	86

缺陷位置检查表示例

- 下图为缺陷位置检查表, 用以发现汽车挡风玻璃中的气泡的位置和形状。利用检查表发现位置主要在右侧, 进一步考察生产过程发现右侧压力较小。

气泡调查检查表

日期

产品类型

评价

直方图

- 对分类变量，直观展示各个取值的频数。
- 对连续变量，分组后展示各组频数，并反映分布范围、中心位置、分散程度、分布偏斜情况、有无离群值、是否多峰等信息。
- 直方图提供了关于样本的总体特征的线索，原本在数据表格中不易发现的模式变得明显了。

- 对离散数据观测值 x ，在R中用`barplot(table(x))`绘图。如

```
x <- sample(1:3, 100, replace=TRUE,  
            prob=c(0.5, 0.3, 0.2))
```

```
barplot(table(x))
```

- 对连续数据观测值 x ，在R中用`hist(x)`绘图。如

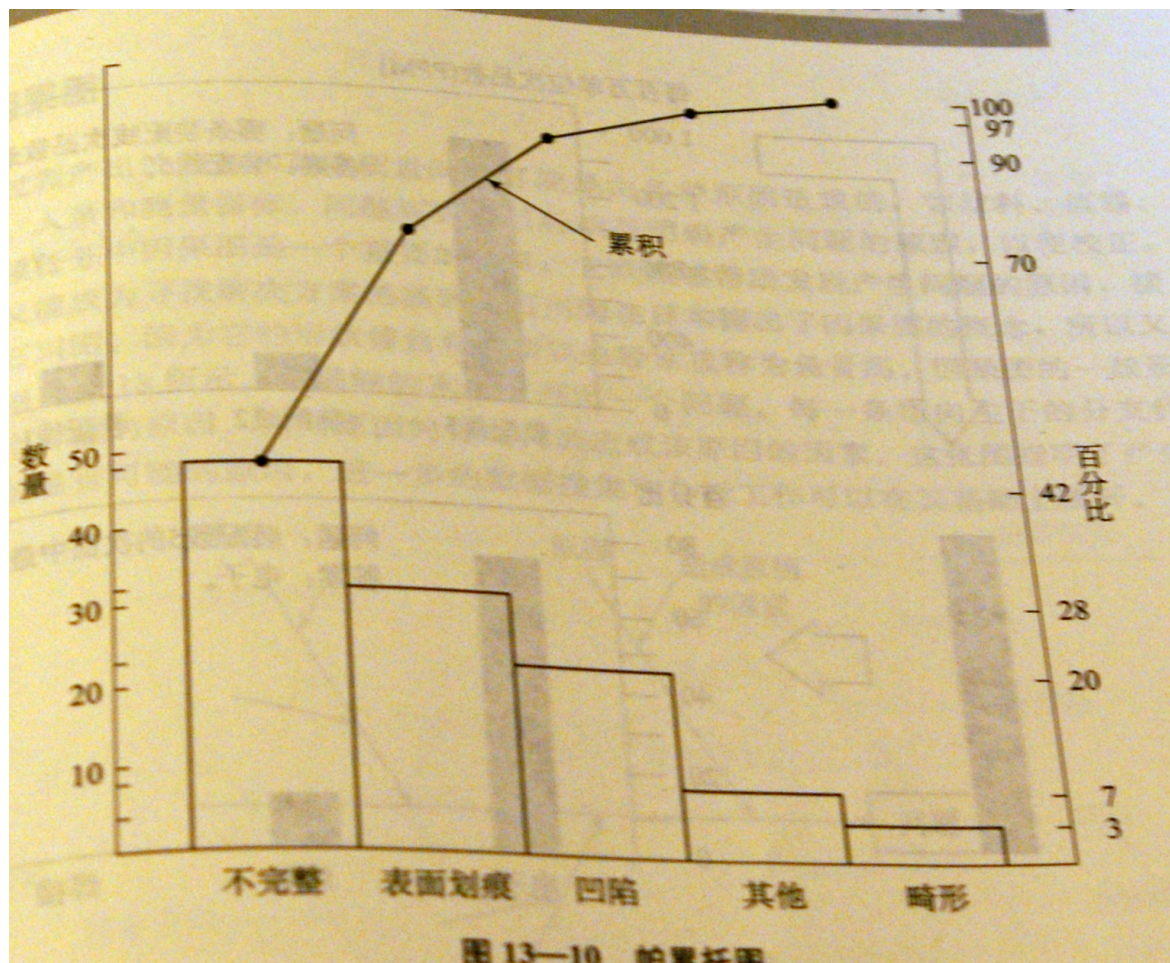
```
x <- exp(rnorm(100))  
hist(x)
```

帕累托图

- 帕累托(1848—1923)是一位意大利经济学家，他发现米兰85%的财富掌握在15%的人手中。
- 朱兰在对一家造纸厂的成本分析中发现，全部质量成本的61%可以归结为一个科目——“破损”。
- 帕累托分析清晰地将关键的少数从平凡的大多数中分离出来，提供了选择改进项目的方向。
- 我们平常所说的“抓重点”，“解决主要矛盾”，就是这种思想。
- 帕累托分布(Pareto distribution)显示了观察对象的特征从最高频率到最低频率的排序。帕累托图(Pareto diagram)是反映从最高频率数据到最低频率数据的直方图。

帕累托图示例：缺陷种类

- 下图用可视化的辅助清晰地展示了缺陷的相对比例，并能够用来发现改进的机会。成本最高、最重要的问题被凸显出来。



顾客投诉的帕累托图

- 下面的帕累托图分析了产品运输中产生的额外运费。发现产生高额运费的最常见原因是顾客要求。第二大原因是工时不足。

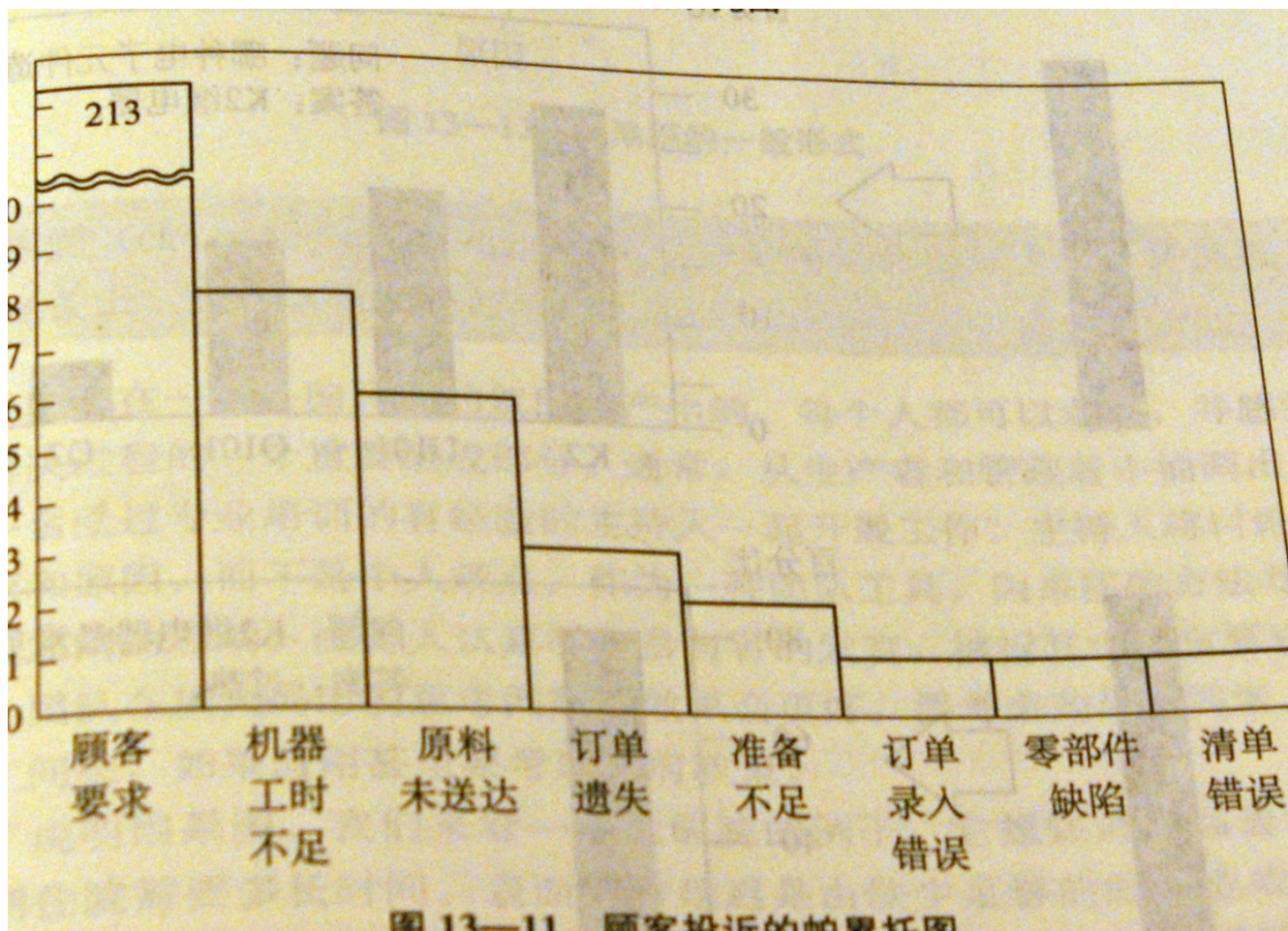
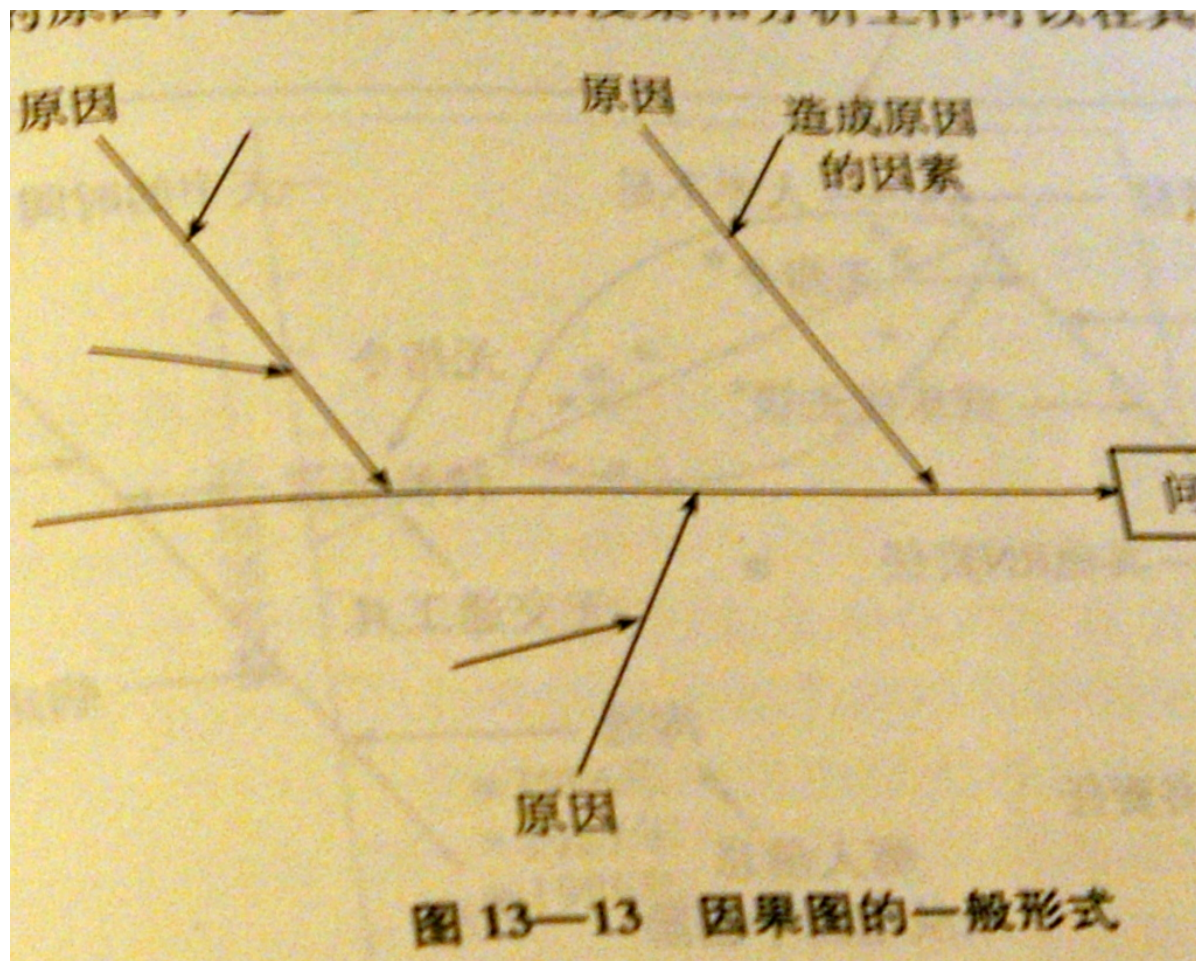


图 13—11 顾客投诉的帕累托图

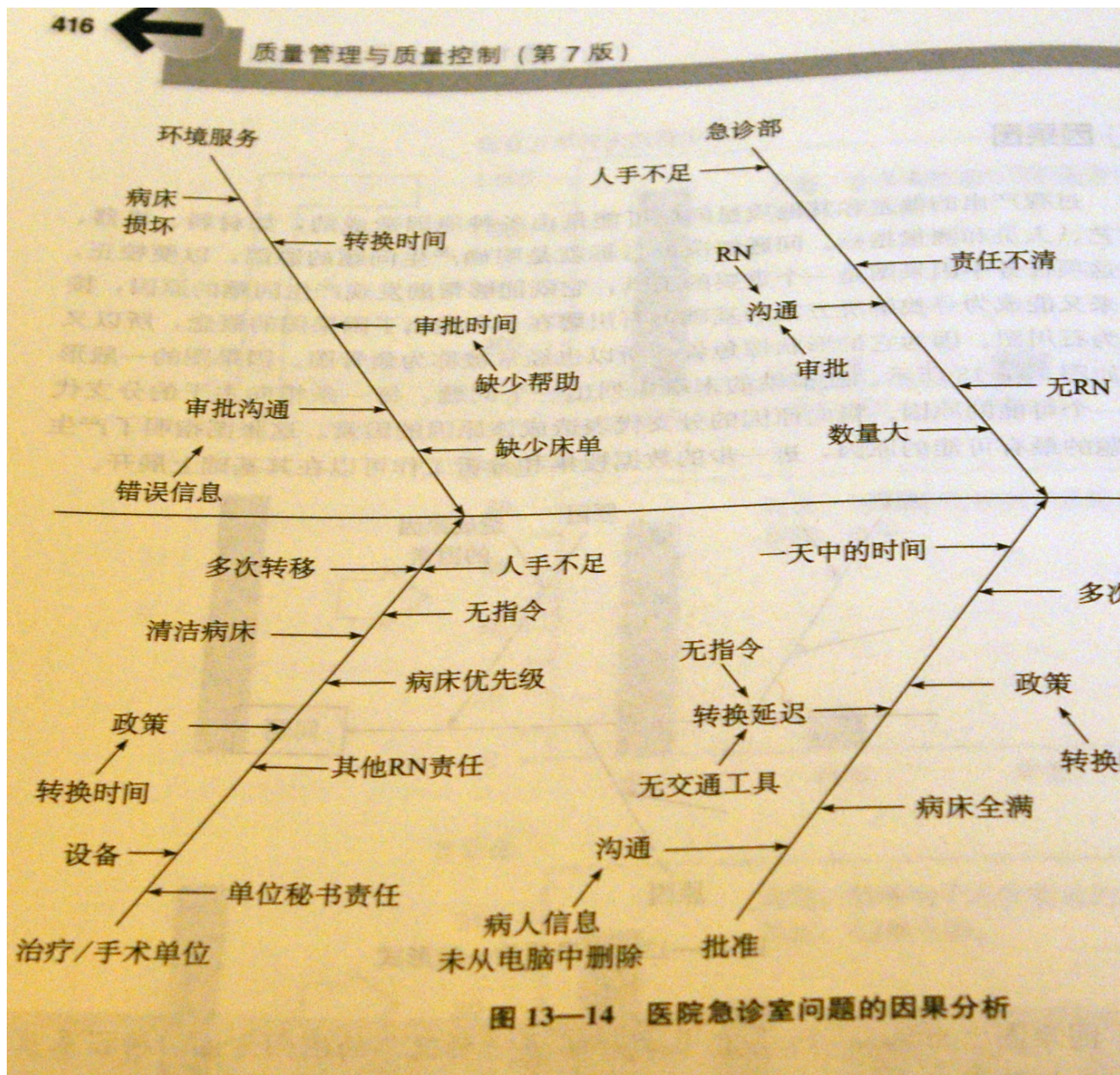
因果图

- 过程产出的偏差和其他质量问题可能是由各种原因造成的，如材料、机器、工艺、人员和测量指标。
- 因果图有助于发现产生问题的原因，并称为寻找解决方案的基础。
- 因果图又称为石川图、鱼骨图。
- 因果图(cause-and-effect diagram)是一种简单的图形方法，将因果呈现在一条链条上，并组织起变量之间的关系。
- 因果图的一般形式如下图，在横轴的末端，列出一个问题。每一条直线主干的分支代表一个可能的原因。指向原因(分支)的分支代表造成该原因的因素。

- 这张图指明了产生问题最有可能的原因，进一步的数据搜集和分析工作可以在其基础上展开。



因果图示例：医院急诊室的因果分析



散点图

- 散点图用处直观显示两个数值型指标之间的相关类型。
- 前面已经讲述。

过程模拟

- 采用计算机模拟的方法，先建立过程的动态模型，用计算机进行随机模拟，可以得到不同参数情况下的绩效指标期望值。

§5.9 统计过程控制

统计过程控制

- 统计过程控制(statistical process control, SPC) 是用于监视过程、识别特殊原因变异，并在适当的时候发出需要采取纠正措施信号的方法。
- 过程中只有一般性原因变异时，过程受控。
- 过程中如果存在特殊原因变异，过程几乎肯定会“失控”。
- 统计控制状态的定义是：随着时间的变化，过程的均值和方差都保持不变。这符合时间序列中的平稳性要求（平稳性要求更多一些）。
- SPC依赖于控制图。许多厂商要求供应商提供控制图。
- SPC要求过程存在可测量的变异，对于质量水平高到六西格玛的生产过程，SPC变得无效。
- 控制图也可以用于服务机构。

§5.9.1 质量控制测量指标

质量控制测量指标

- 质量控制测量指标分为计数值(attribute)和计量值(variable)。
- 计数值是两点分布结果的成功次数记录，常用分数或比例表示。比如，不合格项目比例、每个单位缺陷数、单位机会差错率等。
- 计量值数据是连续取值的，如长度、重量等。对于剂量数据，我们关心其符合规范的程度，通常用平均值、标准差这样的统计参数概括。
- 收集计数数据通常比收集计量数据容易，因为通过简单的检查或计数进行评价通常更快，而计量数据的收集则需要使用某种类型的测量装置。
- 统计中，计数数据不如计量数据提供的信息充分，所以计数检验不如计量检验效率高。

§5.9.2 R的qcc软件包

R的qcc软件包

- R的qcc包实现了控制图、累加图、能力分析、帕累托图等功能。
- 数据框pistonrings中包含了40次测量值，每次测量5个样品。

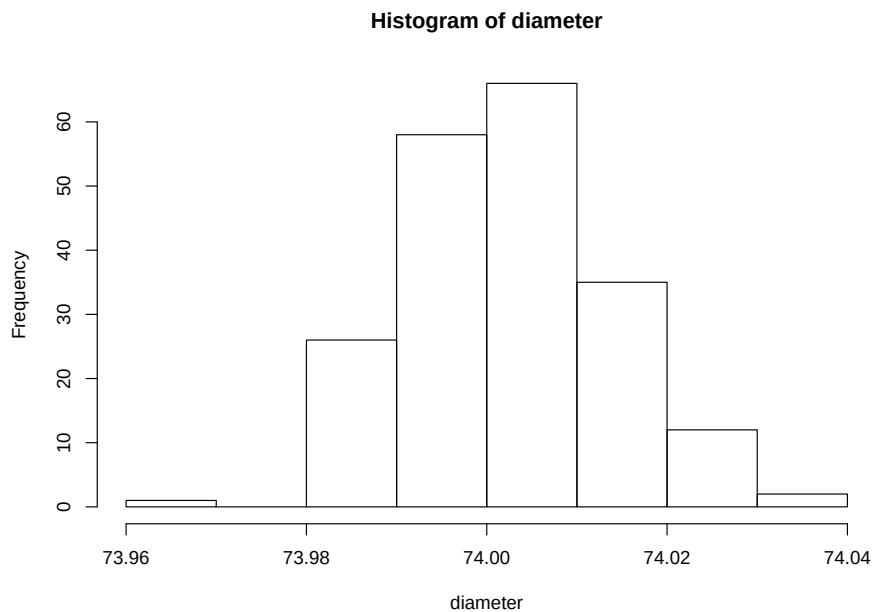
```
> require(qcc)
> data(pistonrings)
> attach(pistonrings)
> print(pistonrings[1:7,])
```

	diameter	sample	trial
1	74.030	1	TRUE
2	74.002	1	TRUE
3	74.019	1	TRUE
4	73.992	1	TRUE
5	74.008	1	TRUE
6	73.995	2	TRUE
7	73.992	2	TRUE

- 做观测值的直方图：

```
> hist(diameter)
```

- 直方图不能反映随时间的变化。

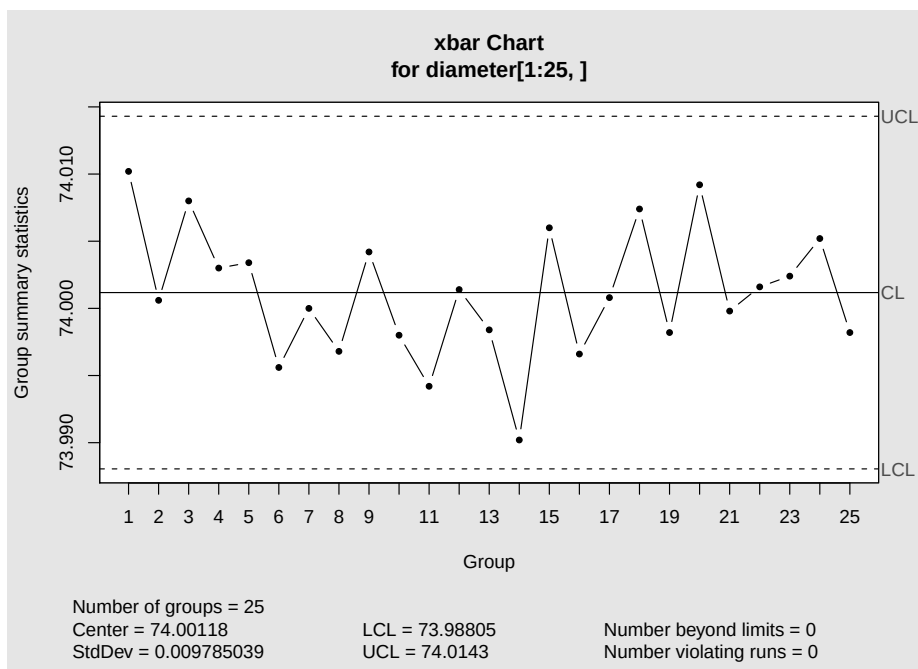


- 我们重排数据框，把一次观测的五个观测值放在一行中。

```
> diameter <- qcc.groups(diameter, sample)
> print(diameter[1:5,])
      [,1] [,2] [,3] [,4] [,5]
1 74.030 74.002 74.019 73.992 74.008
2 73.995 73.992 74.001 74.011 74.004
3 73.988 74.024 74.021 74.005 74.002
4 74.002 73.996 73.993 74.015 74.009
5 73.992 74.007 74.015 73.989 74.014
```

- qcc函数把数据框转换为质量控制所需数据结构，并可以做平均值的控制图。这里只用了前25次测量。

```
obj <- qcc(diameter[1:25,], type="xbar")
```



计算过程能力指数

- 从数据估计过程能力指数 C_p ，可以用qcc中的函数process.capability。
例如

```

> process.capability(obj,
+   spec.limits=c(73.95, 74.05))

Process Capability Analysis

Call:
process.capability(object = obj,
  spec.limits = c(73.95, 74.05))

Number of obs = 125          Target = 74
      Center = 74.00118      LSL = 73.95
      StdDev = 0.009785039    USL = 74.05

```

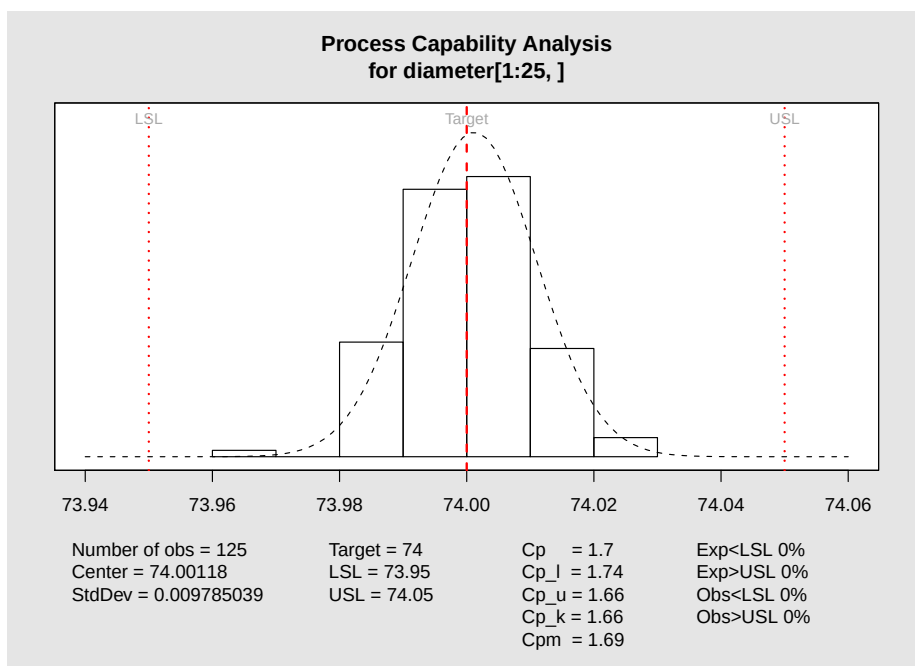
```

Capability indices:

      Value   2.5%  97.5%
Cp    1.703   1.491  1.915
Cp_l   1.743   1.555  1.932
Cp_u   1.663   1.483  1.844
Cp_k   1.663   1.448  1.878
Cpm    1.691   1.480  1.902

Exp<LSL 0%   Obs<LSL 0%
Exp>USL 0%   Obs>USL 0%

```



§5.9.3 能力与受控

能力与受控

- 受控是要求过程指标随时间的变化是稳定的，均值在中心上下随机波动，没有模式。
- 过程能力反映的是对标准差的控制， C_p 越大，过程的标准差越小。
- 期望的状态是既受控又有能力。

§5.9.4 计量值数据控制图

计量值数据控制图

- 常用均值控制图($\bar{x}$ 图)和极差控制图(R 图)。
- 均值控制图用于监视过程的中心；极差用于简便地测量变异，尤其是手工完成时。使用计算机分析时用标准差图代替。
- qcc包的qcc可以绘制均值控制图和标准差控制图。并可以用一个数据框作为训练，另一个数据框作为监测。比如，只适用pistonrings数据的前25次测量结果做均值控制图：

```
obj <- qcc(diameter[1:25,], type="xbar")
summary(obj)
```

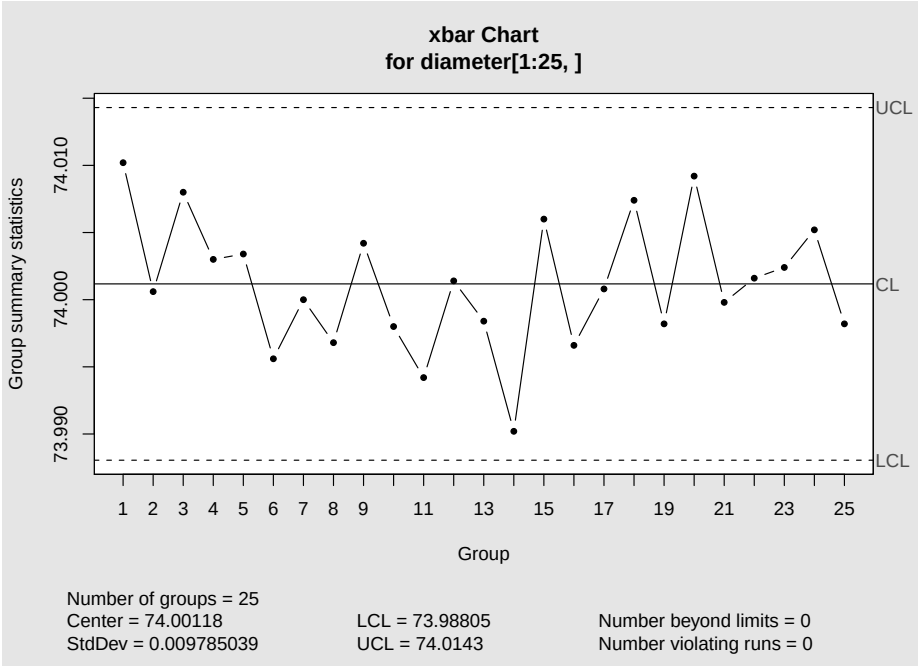
```
Call:
qcc(data = diameter[1:25, ], type = "xbar")

xbar chart for diameter[1:25, ]

Summary of group statistics:
      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 73.99   74.00   74.00   74.00   74.00   74.01

Group sample size: 5
Number of groups: 25
Center of group statistics: 74.00118
Standard deviation: 0.009785039

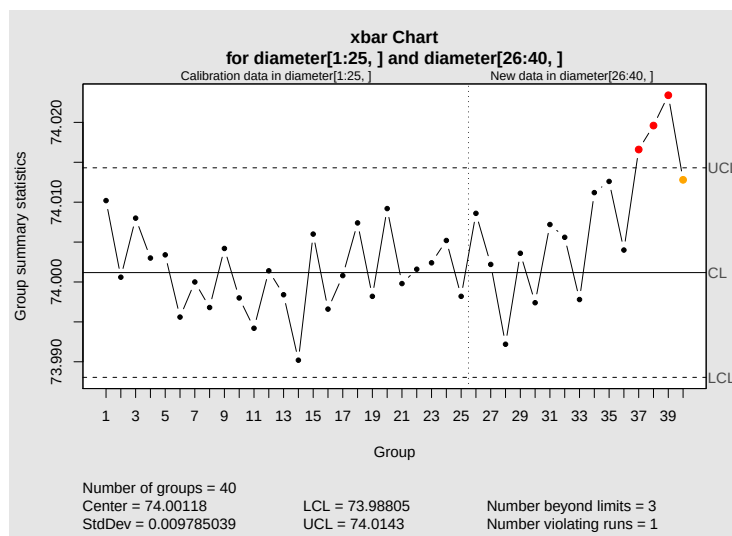
Control limits:
      LCL      UCL
73.98805 74.0143
```



- 用前25次测量建立的控制限监测后15次测量:

```
obj2 <- qcc(diameter[1:25,], type="xbar",
            newdata=diameter[26:40,])
```

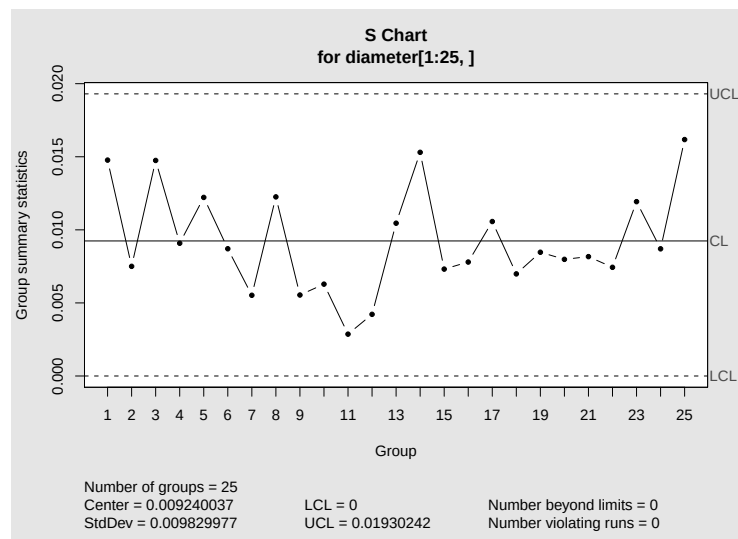
- 当过程处于统计控制状态时，点随机地落在控制限之间，且不呈现任何可以识别的形态。
- 结果有3个点落到了上限以外。



标准差的控制图

- 用前25次观测做标准差控制图:

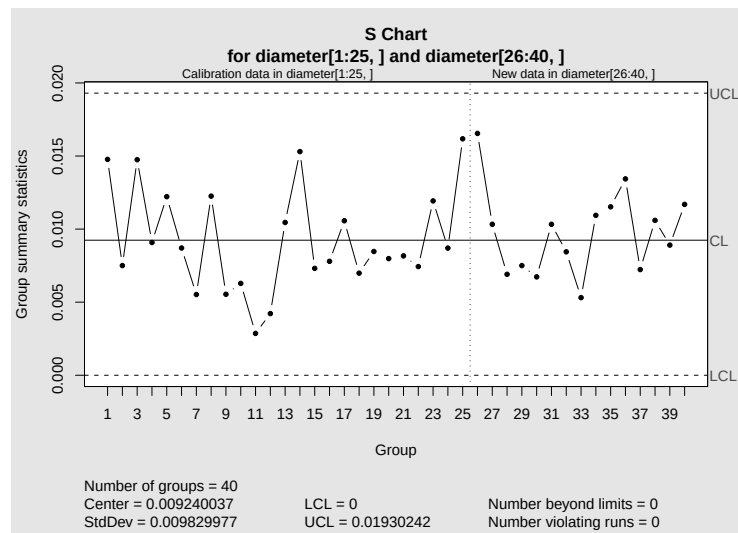
```
obj3 <- qcc(diameter[1:25,], type="S")
```



标准差的控制图

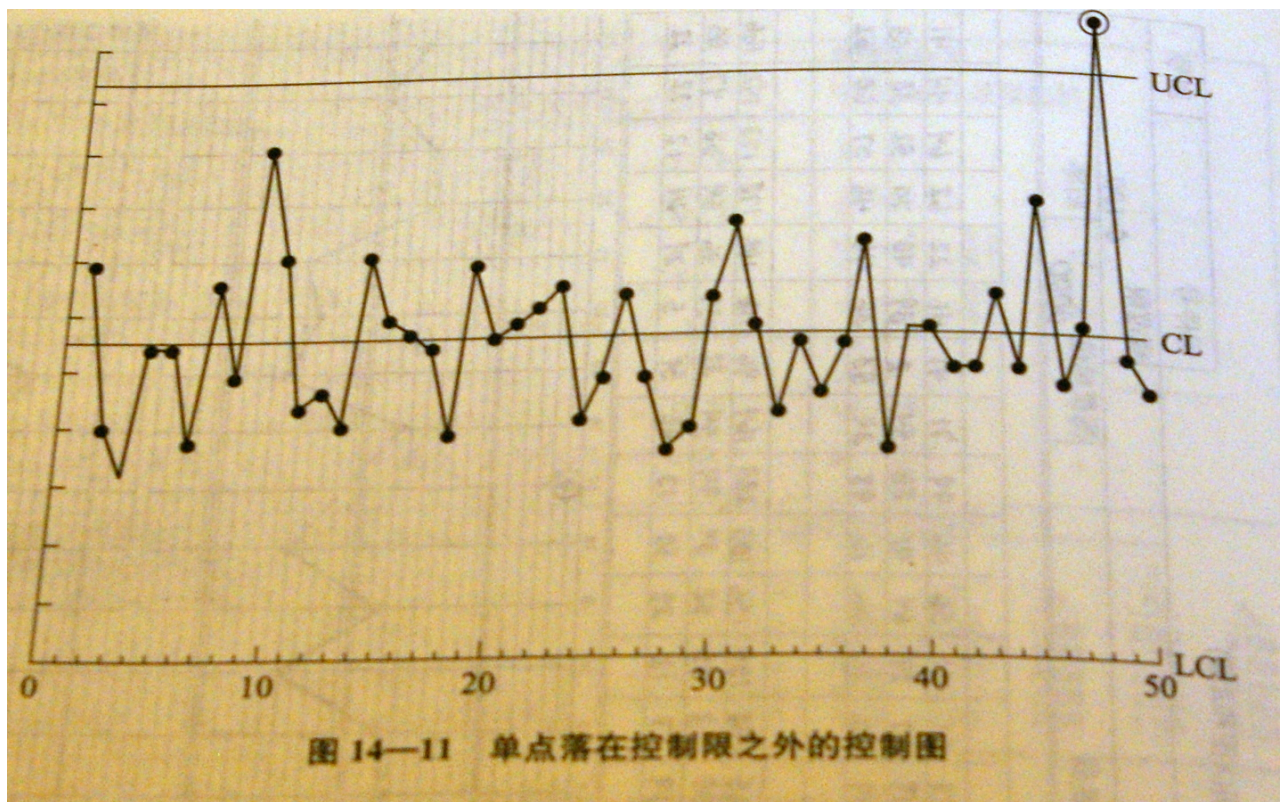
- 用前25次观测得到控制限监测后15次观测，做标准差控制图:

```
obj4 <- qcc(diameter[1:25,], type="S",
            newdata=diameter[26:40,])
```

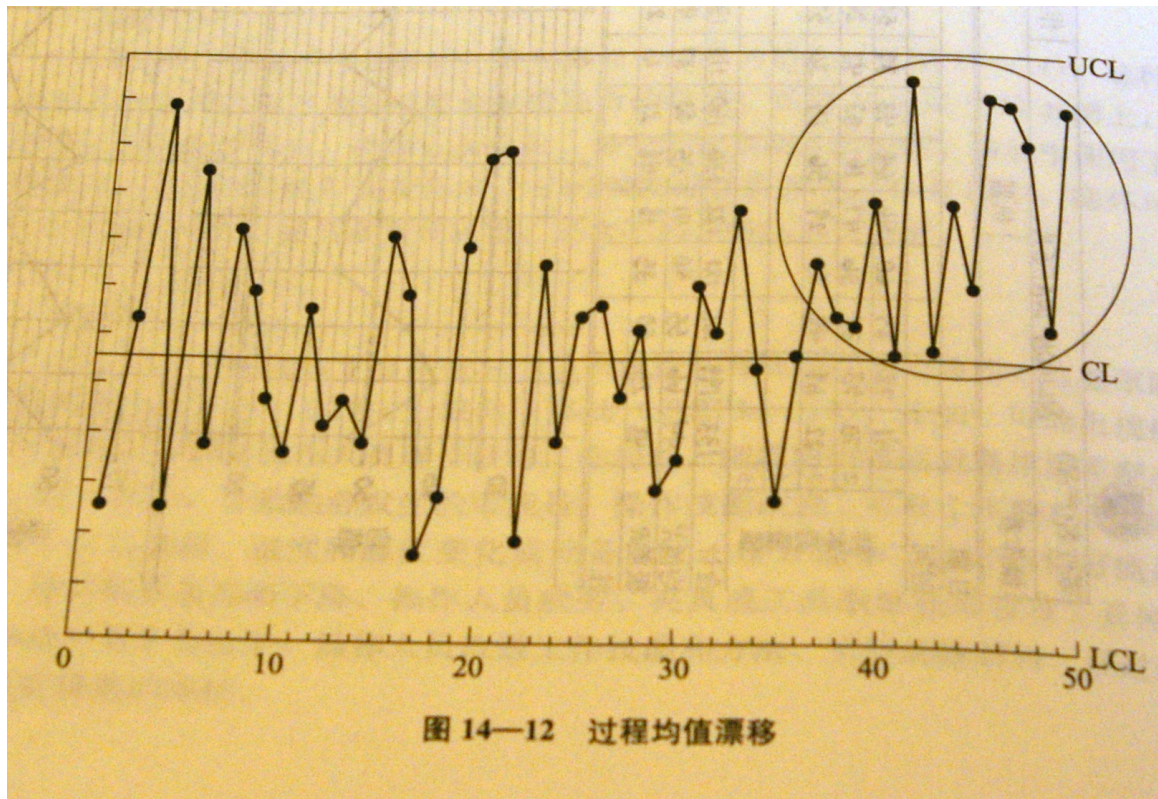


控制图的一些失控模式

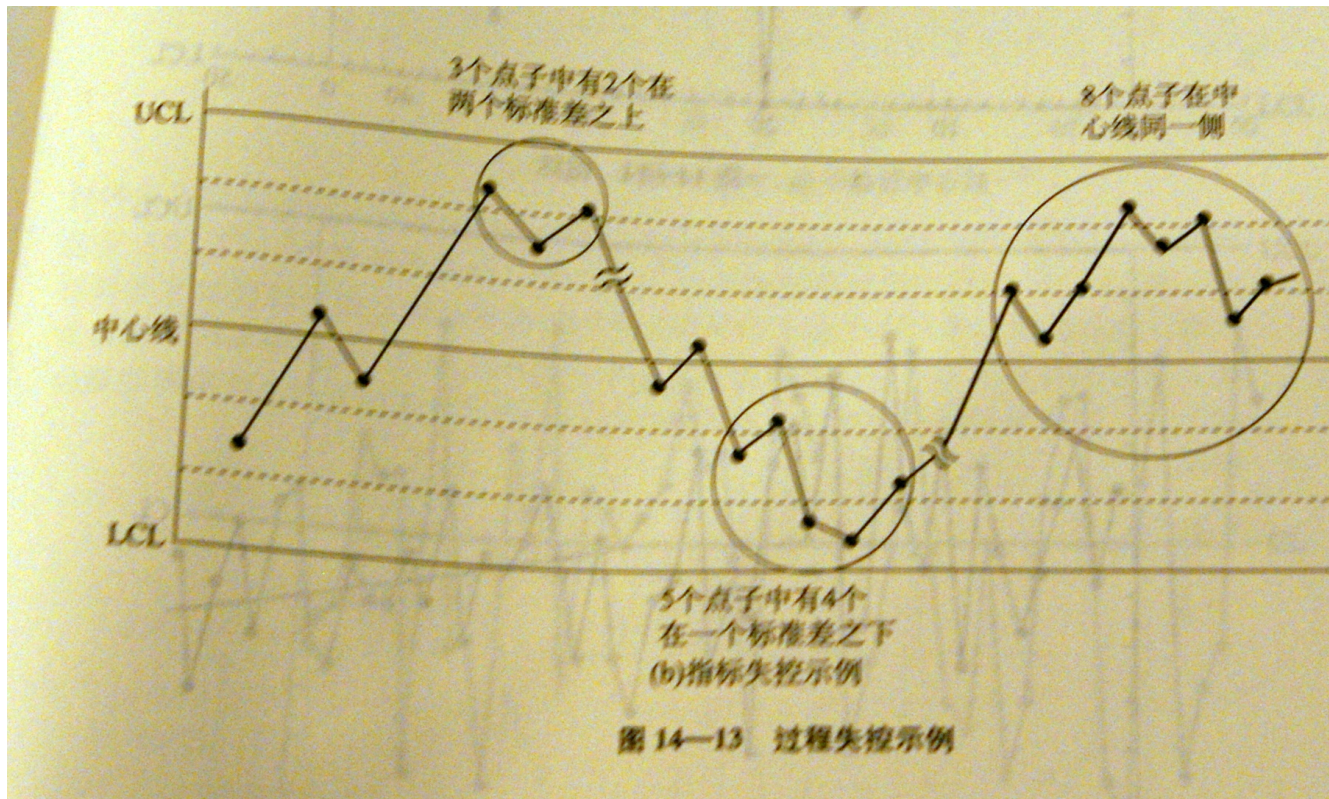
- 单点落在控制限外:



- 过程均值漂移:

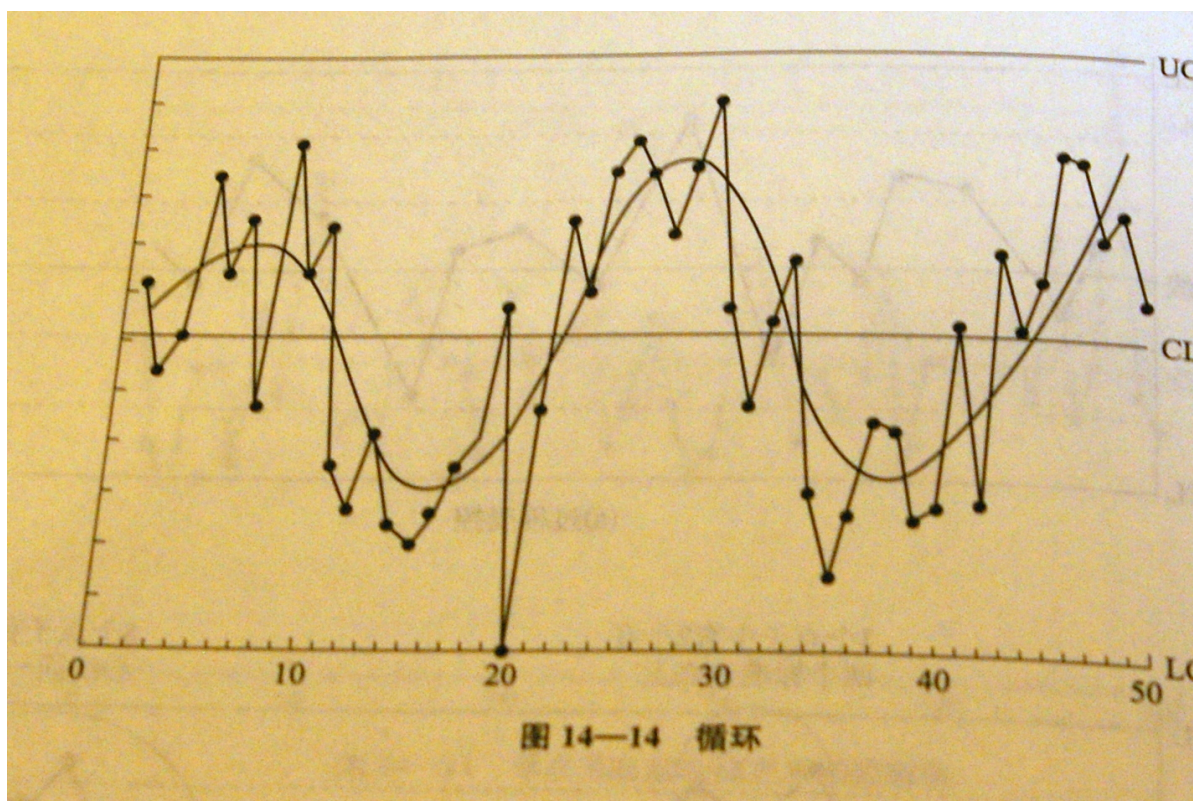


- 过程均值漂移另外一例。

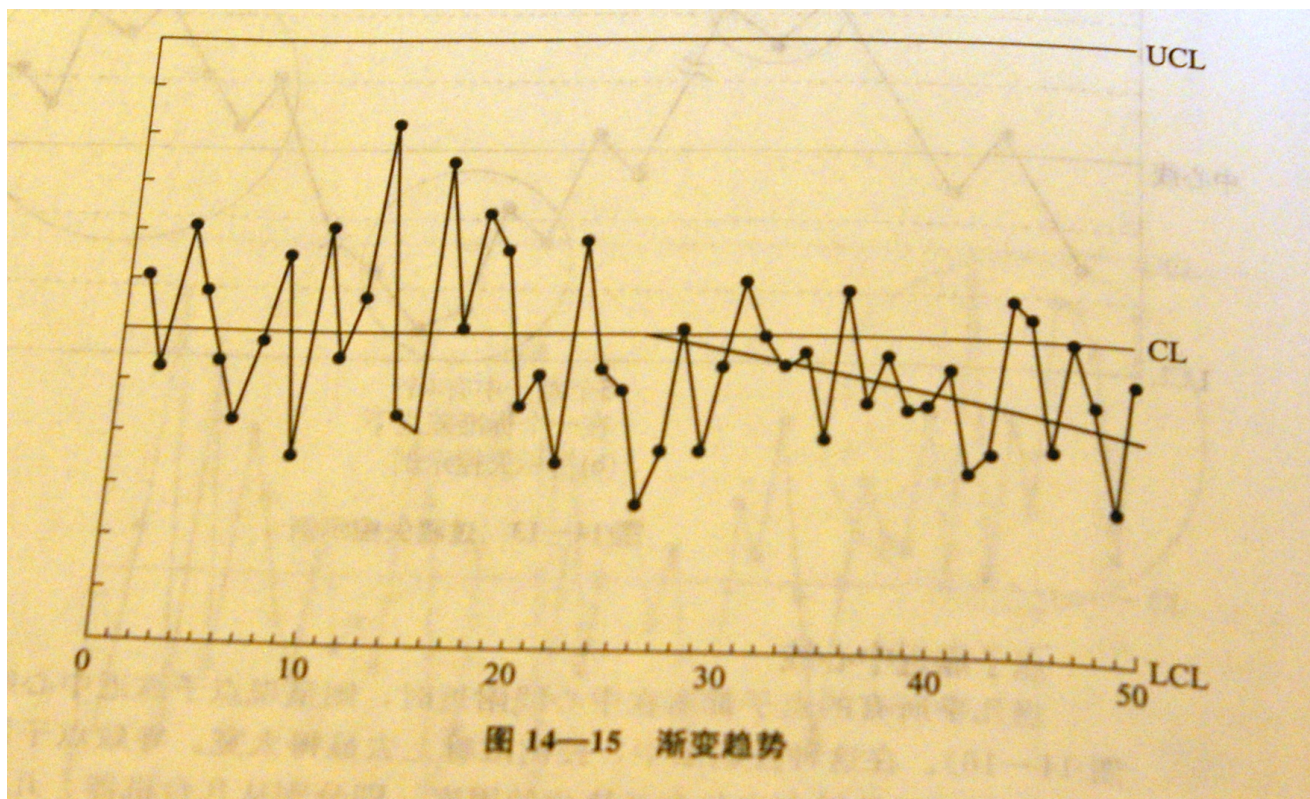


过程均值漂移的简单判断

- 如果有8个连续点落在中心线的一边；
- 将中心线与控制限中间的区域等分为三部分，则(1) 如果在3个连续的点中，有2个落在中心线与控制限之间外侧1/3的区域；(2) 如果在5个连续的点中，有4个落在外侧的2/3的区域。也可以判断过程失控。
- 周期性波动：



- 趋势:



§5.9.5 计数值的数据控制图

计数值的数据控制图

- 计数值数据只有两个值，好或坏，及格或不及格等。
- 计数值的控制图最常用的是 p 图。
- 需要区分缺陷和缺陷品。缺陷(defect)是某一种不合格的质量特性，一个样品可以多个缺陷。缺陷品(defective)是指由一个或多个缺陷的样品。要了解计数值控制图到底是针对缺陷还是缺陷品。

R中的计数值控制图

- 数据框orangejuice包含了54次计数结果，其中变量size是总数，D是成功数。用前30次结果建立控制限。

```
data(orangejuice)
head(orangejuice)
attach(orangejuice)
obj <- qcc(D[trial], sizes=size[trial],
```

```
type="p")
summary(obj)
```

Call:
qcc(data = D[trial], type = "p",
 sizes = size[trial])

p chart for D[trial]

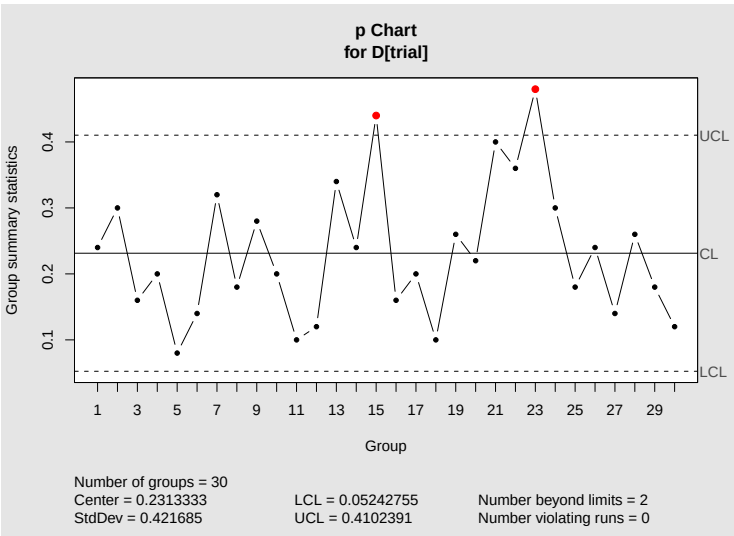
Summary of group statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0800	0.1600	0.2100	0.2313	0.2950	0.4800

Group sample size: 50
Number of groups: 30
Center of group statistics: 0.2313333
Standard deviation: 0.421685

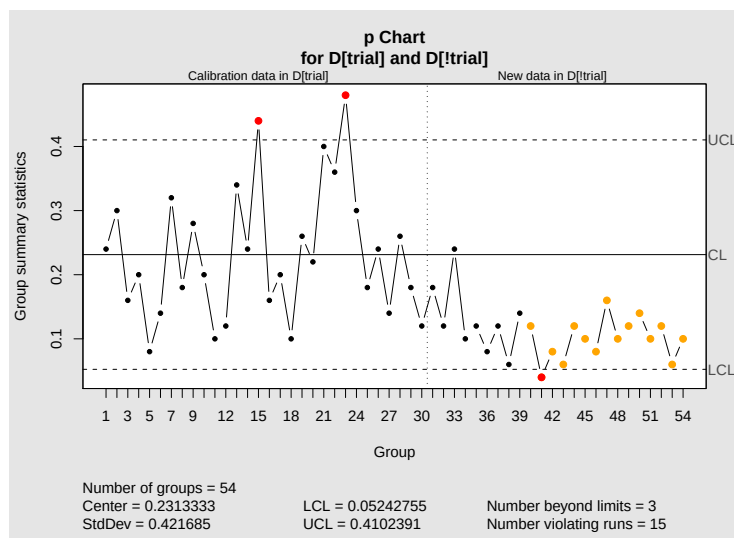
Control limits:

LCL	UCL
0.05242755	0.4102391



- 用前30次得到的控制限监测后24次:

```
obj2 <- qcc(D[trial], sizes=size[trial],
            type="p", newdata=D[!trial],
            newsizes=size[!trial])
```



计数值控制图的种类

- 除了 p 图外, 还可以做不合格数的 np 图, 缺陷数的 c 图或 u 图。 c 图假定单位个数恒定, u 图不需要假定单位个数恒定。