数据分析建模简介

观察和实验是科学家探究自然的主要方法，但如果你有数据，那么如何让这些数据开口说话呢？数据用现代人的话说即信息，信息的挖掘与分析也是建模的一个重要方法。

科学史上最有名的数据分析例子

- 开普勒三定律
- 数据来源：第谷·布拉赫（1546-1601,丹麦人），观察力极强的天文学家，一辈子（20年）观察记录了750颗行星资料，位置误差不超过0.67°。
- 观测数据可以视为实验模型。
- 数据处理：开普勒（1571-1630，德国人），身体瘦弱、近视又散光，不适合观天，但有一个非常聪明的数学头脑、坚韧的性格（甚至有些固执）和坚强的信念（宇宙是一个和谐的整体），花了16年（1596-1612）研究第谷的观测数据，得到了开普勒三定律。

数据分析法

- 思想：
- 采用数理统计方法（如回归分析、聚类分析等）或插值方法或曲线拟合方法，对已知离散数据建模。
- 适用范围：系统的结构性质不大清楚，无法从理论分析中得到系统的规律，也不便于类比，但有若干能表征系统规律、描述系统状态的数据可利用。

基 础 知 识

- 数据分析
  - 数据描述性分析：粗线条的描述方式
  - 回归分析：变量间的关系
  - 主成分分析：谁更重要？怎样降低维数？
  - 判别分析：该归入哪一类？
  - 聚类分析：怎样归类？
  - 时间序列分析：现象随时间怎样变化？

1数据的描述性统计

针对一组杂乱无章的数据（即样本），描述性统计的步骤为：

- **初步整理和直观描述**----作出频数表和直方图

- 进一步加工，提取有用信息----**计算统计量**

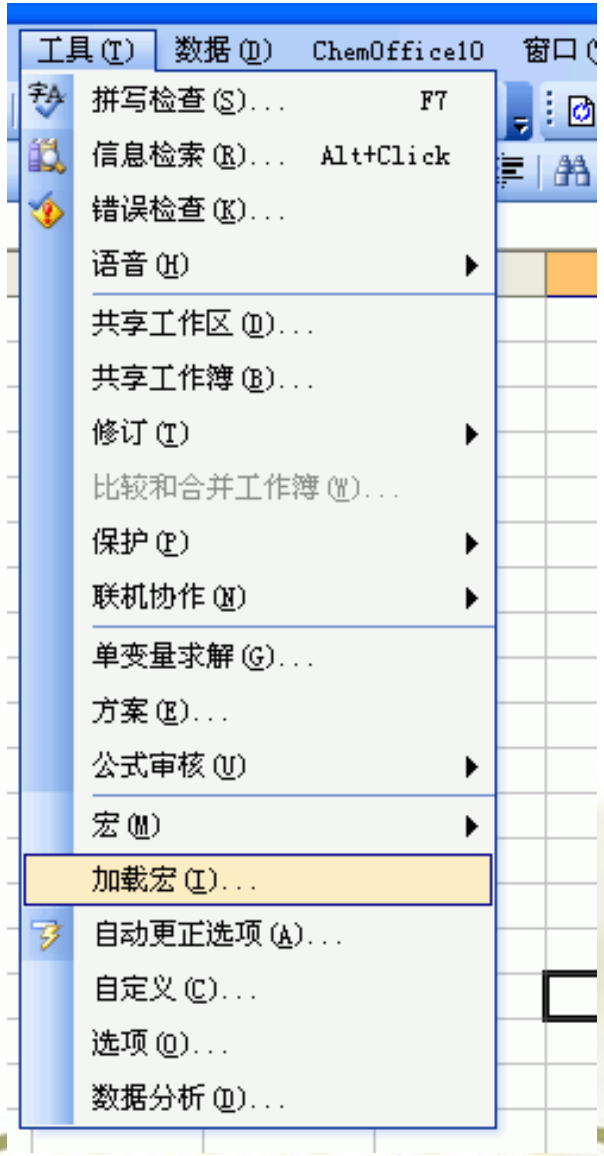
  统计量：用来进一步反映数据特征，它是加工出来的，反映样本数据特征的函数，它不含任何未知量，大致可以分三类：

  (1)表示位置的统计量----样本均值、中位数、上下1、4分位点

  (2)表示变异程度的统计量----标准差、方差、极差

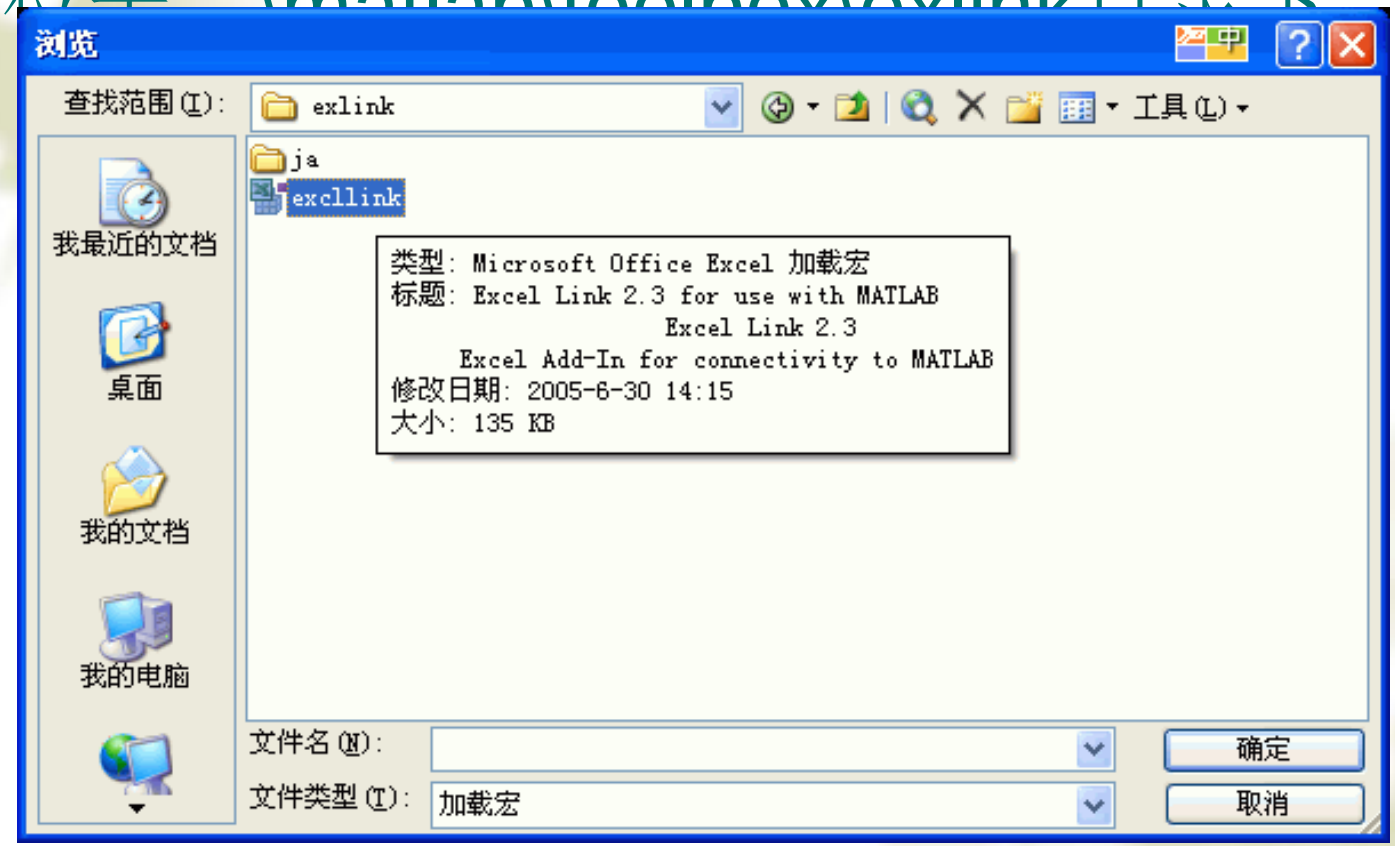
  (3)表示分布形状的统计量----偏度、峰度

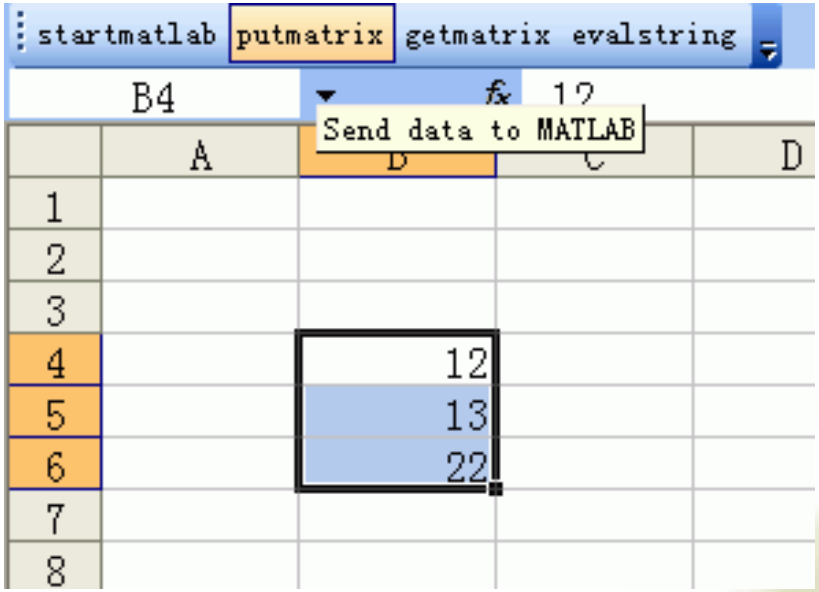
Excel与Matlab数据传递

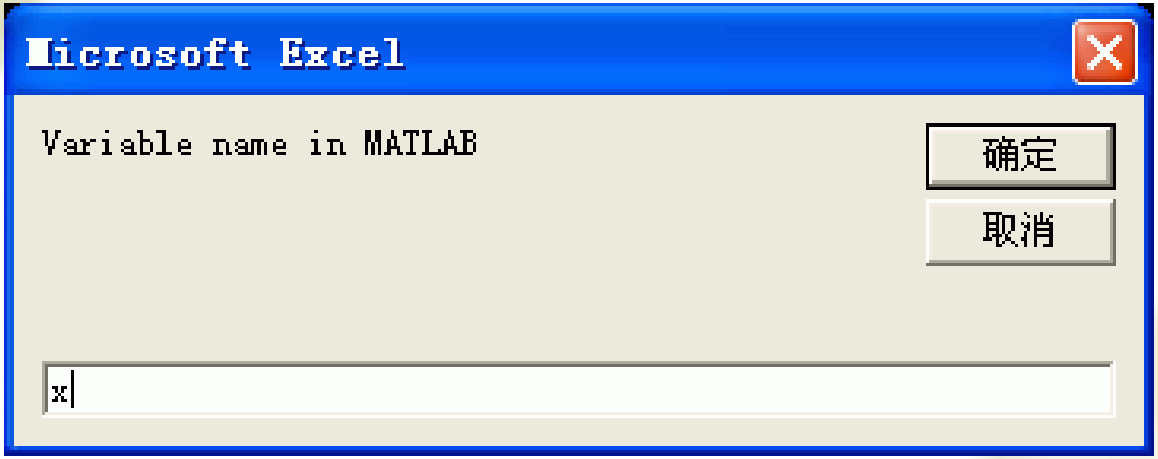


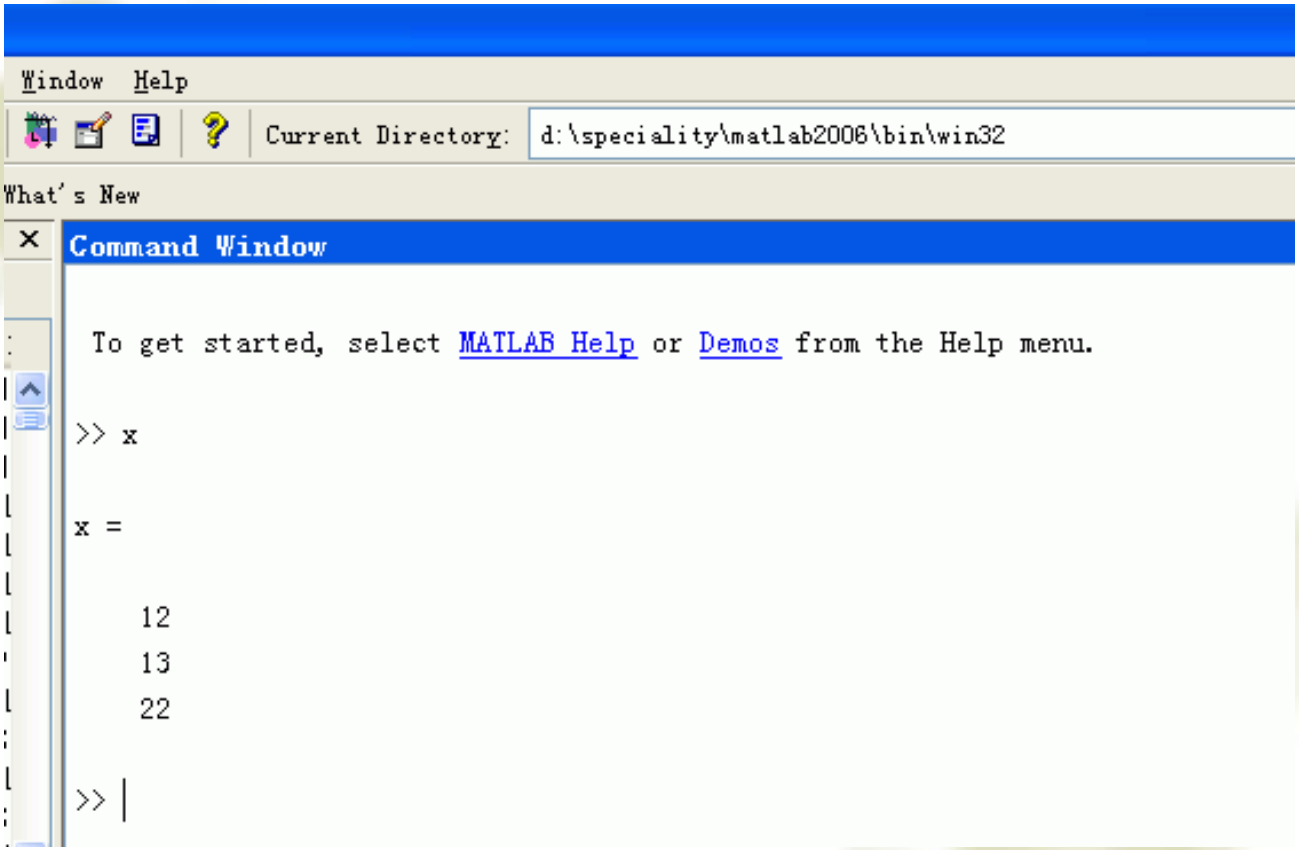

加载宏
可用加载宏(A):
ChemDraw/Excel 10
CombiChem/Excel 10
Excel Link 2.3 for use with MATLAB
Internet Assistant VBA
S-PLUS Add-In
查阅向导
分析工具库
分析工具库 - VBA 函数
规划求解
欧元工具
条件求和向导
确定
取消
浏览(B)...
自动化(U)...
ChemDraw/Excel 10
Adds chemical data handling to Excel.
(ChemOffice 10)

位于...\matlab\toolbox\exlink目录下




startmatlab
putmatrix
getmatrix
evalstring
B4
12
Send data to MATLAB
A
D
1
2
3
4
5
6
7
8
12
13
22


Microsoft Excel
Variable name in MATLAB
确定
取消
x


Window Help
Current Directory: d:\speciality\matlab2006\bin\win32
What's New
Command Window
To get started, select MATLAB Help or Demos from the Help menu.
>> x
x =
12
13
22
>>

2 统计推断

- 统计推断主要有参数估计和假设检验。
- 参数估计：点估计、均值的区间估计和方差的区间估计
- 假设检验：均值检验、方差检验

(1) 参数的点估计

- 设有如下20个服从正态分布的随机数：

  1.2, 0, 2.3, 1.3, 2.5, 2.1, 0.3,-0.3, 0.9, 0.7 ,

  0.2, 1.5, 2.5, 0.5, 0.2, 0.8, 1.7, 0.1,-0.2, 0.9。

- 问该分布的数学期望和标准差是多少？

设总体 $X \sim N(\mu, \sigma^2)$

样本 $X_1, \cdots, X_n$ $X_i \sim N(\mu, \sigma^2),\ i = 1, \cdots, n$

- 则X的数学期望和方差的点估计为：

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

  - 对如上数据可计算得到数学期望的估计为0.96，标准差的估计为0.8988。

(2) 参数的区间估计(假设检验)

- 设有如下20个服从正态分布的随机数：

  1.2,　0, 2.3, 1.3, 2.5, 2.1, 0.3,-0.3, 0.9, 0.7,

  0.2, 1.5, 2.5, 0.5, 0.2, 0.8, 1.7, 0.1,-0.2, 0.9。问：

A. 若标准差已知(等于1)，问此随机数可能是由均值为1的正态分布产生的吗？

B. 若均值已知(等于1)，问此随机数可能是由标准差为1.1的正态分布产生的吗？

C. 若标准差和均值未知，问此随机数可能是由均值为1.1,标准差为0.9的正态分布产生的吗？

正态总体数学期望的区间估计1-方差已知

- 构造统计量：

$$Z=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad \text{因} \bar{X}\sim N(\mu,\frac{\sigma^2}{n}) \quad \text{故} Z\sim N(0,1)$$

- 给定置信水平1-a，取 $z_{1-\alpha/2}$ 满足：

norminv(0.975)

$$P(-z_{1-\alpha/2}<Z<z_{1-\alpha/2})=1-\alpha \qquad z_{1-0.05/2}=1.96$$

$$P(\bar{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\bar{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})=1-\alpha$$

- 故数学期望置信水平1-a的区间估计为：

$$\left(\bar{X}-z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$\bar{X}=0.96$

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}=1.96\times\frac{1}{\sqrt{20}}\approx 0.47$$

- 因 0.96-0.47<1<0.96+0.47，故u=1很可能成立。

(3) matlab参数估计与假设检验

参数估计Matlab命令

均值和标准差未知

[muhat,sigmahat,muci,sigmaci] = normfit(X,alpha)

- 此命令在显著性水平alpha下估计数据X的参数（alpha缺省时设定为0.05），返回值muhat是X的均值的点估计值，sigmahat是标准差的点估计值, muci是均值的区间估计,sigmaci是标准差的区间估计.
- 例： r =[0.2, 1.5, 2.5, 0.5, 0.2, 0.8, 1.7, 0.1,-0.2, 0.9];

  [mu,sigma,muci,sigmaci] = normfit(r);

假设检验Matlab命令

- 在总体服从正态分布的情况下，可用以下命令进行假设检验.

1、总体方差sigma2已知时，总体均值的检验使用 z-检验

[h,sig,ci] = ztest(x,m,sigma,alpha,tail)

- 检验数据 x 的关于均值的某一假设是否成立，其中sigma为已知方差， alpha 为显著性水平，究竟检验什么假设取决于 tail 的取值：

tail = 0，检验假设“x 的均值等于 m ”

tail = 1，检验假设“x 的均值大于 m ”

tail =-1，检验假设“x 的均值小于 m ”

tail的缺省值为 0， alpha的缺省值为 0.05.

- 返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为均值的 1-alpha 置信区间.

3、两总体均值的假设检验使用 t-检验

[h,sig,ci] = ttest2(x,y,alpha,tail)

检验数据 x ，y 的关于均值的某一假设是否成立，其中 alpha 为显著性水平，究竟检验什么假设取决于 tail 的取值：

tail = 0，检验假设“x 的均值等于 y 的均值 ”

tail = 1，检验假设“x 的均值大于 y 的均值 ”

tail =-1，检验假设“x 的均值小于 y 的均值 ”

tail的缺省值为 0， alpha的缺省值为 0.05.

返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为与x与y均值差的的 1-alpha 置信区间.

(4) 例：自动化车床管理（CUMCM96A）

一道工序用自动化车床连续加工某种零件，由于刀具损坏等会出现故障.故障是完全随机的，并假定生产任一零件时出现故障机会均相同.工作人员是通过检查零件来确定工序是否出现故障的.现积累有100次故障纪录，故障出现时该刀具完成的零件数如下：

459 362 624 542 509 584 433 748 815 505
612 452 434 982 640 742 565 706 593 680
926 653 164 487 734 608 428 1153 593 844
527 552 513 781 474 388 824 538 862 659
775 859 755 49 697 515 628 954 771 609
402 960 885 610 292 837 473 677 358 638
699 634 555 570 84 416 606 1062 484 120
447 654 564 339 280 246 687 539 790 581
621 724 531 512 577 496 468 499 544 645
764 558 378 765 666 763 217 715 310 851

试观察该刀具出现故障时完成的零件数属于哪种分布.

解 1、数据输入

2、作频数直方图

```
hist(x,10)
```

（看起来刀具寿命服从正态分布）

3、分布的正态性检验

normplot(x)

（刀具寿命近似服从正态分布）

4、参数估计：

[muhat,sigmahat,muci,sigmaci] = normfit(x)

估计出该刀具的均值为594，方差204，均值的0.95置信区间为[553.4962，634.5038]，方差的0.95置信区间为[179.2276，237.1329].

5、假设检验

已知刀具的寿命服从正态分布，现在方差未知的情况下，检验其均值 m 是否等于594.

[h,sig,ci] = ttest(x ,594)

结果：h = 0，sig = 1，ci =[553.4962，634.5038].

检验结果:

1. 布尔变量h=0, 表示不拒绝零假设. 说明提出的假设寿命均值594是合理的.
2. 95%的置信区间为[553.5，634.5], 它完全包括594, 且精度很高.
3. sig-值为1, 远超过0.5, 不能拒绝零假设.

3 相关性分析

- 设随机变量X与Y进行了n次随机试验，得到观察值为(Xi,Yi)，且：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

$$\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$$

则定义相关系数：

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

- 这是衡量x、y之间关系密切程度的指标。

例：从某大学中随机选择10名男生，的观测值如表3。

表3 身高与体重观测值

身高(x)	1.71	1.63	1.84	1.90	1.58	1.60	1.75	1.78	1.80	1.64
体重(y)	65	63	70	75	60	55	64	69	65	58

- 计算相关系数：$\overline{\rho_{xy}} = 0.91$ ；
- 设（x,y）服从正态分布，考虑假设检验问题：
  H0：x与y线性不相关，取显著水平$\alpha = 0.01$，查表得

$$t_{n-2}(\tfrac{\alpha}{2}) = t_8(0.005) = 3.355\,,\ c = \frac{t_{n-2}(\frac{\alpha}{2})}{\sqrt{n-2+t_{n-2}^2(\frac{\alpha}{2})}} = 0.765 < 0.91$$

- 所以拒绝H0，即x与y的线性相关性高度显著，即格子高的人一般体重也要大些。

4 方差分析

- 在试验和实践中，影响试验或生产的因素往往很多，我们通常需要分析哪种因素对事情有显著影响，并希望知道起决定影响的因素在什么时候有着最有利的影响。
- 可以分为：单因素多水平方差分析和多因素方差分析。

- 在实际中，许多问题都涉及多因素的影响，一般要先在众多的因素中选出影响大的，以进一步做更细致的研究。用来判断一个因素的影响“是否大”的主要方法就是方差分析法。

- 在方差分析中，把试验数据的总波动（总变差或总方差）分解为由所考虑因素引起的波动（各因素的变差）和随机因素引起的波动（误差的变差），然后通过分析比较这些变差来推断哪些因素对所考察指标的影响是显著的，哪些是不显著的。

- 单因子方差分析
- 在其它所有可控制因素都保持不变的情况下，只让因素A变化，并观测其结果的变化，这种试验称为“单因素试验”

- 双因素方差分析
- 在两个因素的试验中，往往两个因素的不同水平组合还会产生一定的合作效应，在方差分析中称为交互效应。这种试验称为“双因素试验”

例1（单因素方差分析）一位教师想要检查3种不同的教学方法的效果，为此随机地选取水平相当的15位学生。把他们分为3组，每组5人，每一组用一种方法教学，一段时间以后，这位教师给15位学生进行统考，成绩见下表1。问这3种教学方法的效果有没有显著差异。

方法	成绩				
甲	75	62	71	58	73
乙	71	85	68	92	90
丙	73	79	60	75	81

Matlab中可用函数anova1(…)函数进行单因子方差分析。

调用格式：p=anova1(X)

含义：比较样本 m×n的矩阵X中两列或多列数据的均值。其中，每一列表示一个具有m个相互独立测量的独立样本。

返回：它返回X中所有样本取自同一总体（或者取自均值相等的不同总体）的零假设成立的概率p。

解释：若p值接近0，则认为零假设可疑并认为至少有一个样本均值与其它样本均值存在显著差异。

- Matlab程序：
- Score=[75 62 71 58 73;81 85 68 92 90;73 79 60 75 81]';
- P=anova1(Score)
- 输出结果：方差分析表和箱形图

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	604.9333	2	302.4667	4.2561	0.040088
Error	852.8	12	71.0667		
Total	1457.7333	14			

由于p值小于0.05，拒绝零假设，认为3种教学方法存在显著差异。

例2（双因素方差分析）为了考察4种不同燃料与3种不同型号的推进器对火箭射程（单位：海里）的影响，做了12次试验，得数据如表2所示。

表2 燃料-推进器-射程数据表

	推进器1	推进器2	推进器3
燃料1	58.2	56.2	65.3
燃料2	49.1	54.1	51.6
燃料3	60.1	70.9	39.2
燃料4	75.8	58.2	48.7

在Matlab中利用函数 anova2函数进行双因素方差分析。

调用格式：p=anova2(X,reps)

含义：比较样本X中两列或两列以上和两行或两行以上数据的均值。不同列的数据代表因素A的变化，不同行的数据代表因素B的变化。若在每个行-列匹配点上有一个以上的观测量，则参数reps指示每个单元中观测量的个数。

返回：

- 当 reps=1（默认值）时，anova2将两个p值返回到向量p中。
  - H0A：因素A的所有样本（X中的所有列样本）取自相同的总体；
  - H0B：因素B的所有样本（X中的所有行样本）取自相同的总体。
- 当reps>1时，anova2还返回第三个p值：
  - H0AB：因素A与因素B没有交互效应。
- **解释**：如果任意一个p值接近于0，则认为相关的零假设不成立。

Matlab程序：

```
disp1=[58.2 56.2 65.3;49.1 54.1 51.6;60.1 70.9 39.2;75.8 58.2 48.7]';
p=anova2(disp1,1)
```

输出结果： 方差分析表

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	157.59	3	52.53	0.43059	0.73875
Rows	223.8467	2	111.9233	0.91743	0.44912
Error	731.98	6	12 1.9967		
Total	1113.4167	11			

由于燃料和推进器对应的**p**值均大于**0.05**，所以可以接受零假设**H0A**和**H0B**，认为燃料和推进器对火箭的射程没有显著影响。