统计信息分析

——管理·经济·生产中的信息处理——

宋俊杰编著

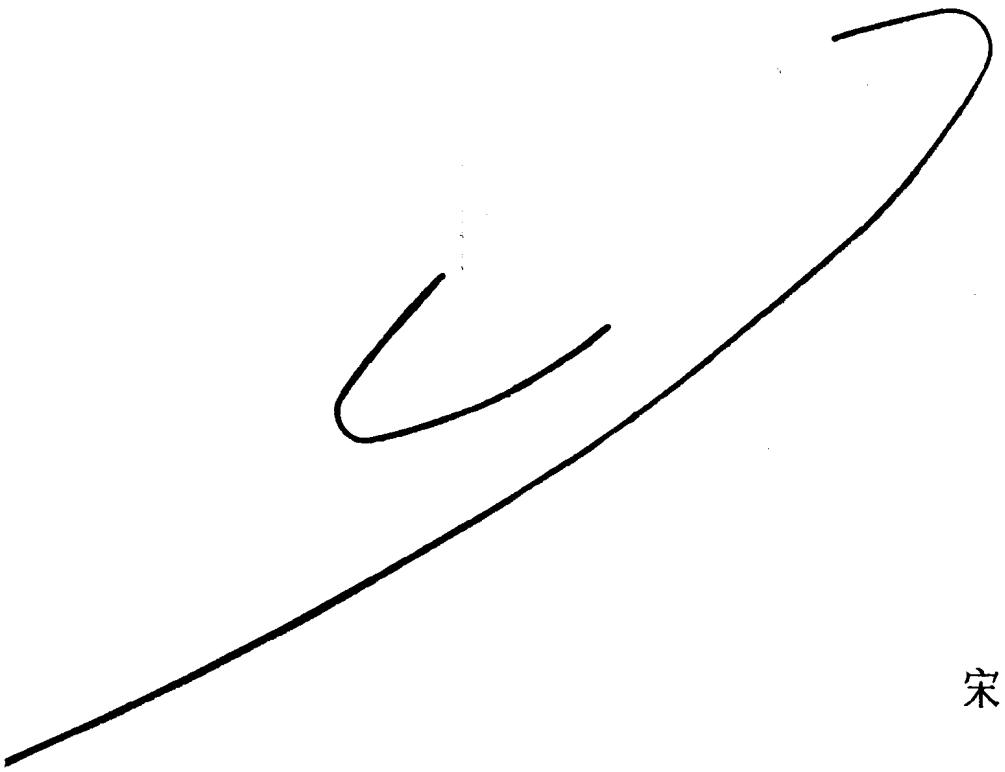
下

南开大学出版社

统计信息分析

——管理·经济·生产中的信息处理——

下



宋俊杰编著

内 容 简 介

本书从实用角度出发，比较系统地介绍了各种重要的统计信息分析的方法，包括一些近代的新成就。全书共分上、下两册，上册包括九章，其主要内容有：概率分析及其应用、信息量及其应用、随机抽样与试验设计、统计推断、各种常用的统计检验方法。下册包括八章，其主要内容有：方差分析、多样性分析、相关分析、列联表的信息分析、回归预测与控制、时间序列的预测与滤波、主分量分析与产品预测、响应面方法学。可供工程技术人员、管理人员、科研和设计人员、大专院校学生、研究生、教师参考使用。

统计信息分析（下）

宋俊杰　编著

南开大学出版社出版

（天津八里台南开大学校内）

新华书店天津发行所发行

河北省滦县印刷厂印刷

1986年9月第1版　　1986年9月第1次印刷

开本：787×1092 1／16 印张：22.5　插页2

字数：560千　　印数：1—10,000

统一书号：13301·34　　定价：4.30元

目　　录

下　册

10　方差分析 …… 323
10·1　一元方差分析 …… 323
10·2　二元方差分析（有重复试验的情况） …… 333
10·3　交互作用与自由度 …… 342
10·3·1　交互作用 …… 342
10·3·2　自由度 …… 344
10·3·3　交互作用因子的自由度 …… 346
10·4　无交互作用的二元方差分析 …… 347
10·5　观测数据个数不等的二元方差分析 …… 350
10·5·1　一般计算方法 …… 350
10·5·2　缺失数据估计 …… 353
10·6　随机效应分析及其显著性检验 …… 353
10·7　系统分类试验的方差分析 …… 357
10·8　正交设计的方差分析 …… 360
10·8·1　试验误差估计 …… 360
10·8·2　利用正交表进行偏差平方和分解 …… 362
10·8·3　正交表的方差分析 …… 365
10·9　方差分析对数据要求的条件 …… 380
10·10　最小信息准则在方差分析中的应用 …… 380

11　多样性分析 …… 383
11·1　多样性 …… 383
11·2　多样性的计量——多样度 …… 383
11·2·1　多样度的定义 …… 383
11·2·2　常用的几个多样性指数 …… 384
11·2·3　多样性指数族△β …… 385
11·3　多样度数学模型的构造 …… 387
11·3·1　多样度的基本特性 …… 387
11·3·2　多样度的几种数学模型 …… 389
11·3·3　多样性的排列次序 …… 391
11·4　总体多样性的分解 …… 393
11·5　混合体的多样性分配 …… 395

11·6 多样性分析…… 402
11·6·1 两种方式分类的多样性分析…… 402
11·6·2 三种方式分类的多样性分析…… 407
12 相关分析…… 411
12·1 相关关系与相关分析…… 411
12·2 相关图上的简易相关分析…… 411
12·3 一元相关分析…… 414
12·3·1 样本相关系数r…… 414
12·3·2 r的概率分布…… 418
12·3·3 检验$\rho=0$是否可信…… 419
12·3·4 检验$\rho=\rho_0\neq 0$是否可信…… 420
12·3·5 检验$\rho_1=\rho_2$是否可信…… 421
12·3·6 相关系数ρ的区间估计…… 422
12·4 等级相关分析…… 423
12·5 多元相关分析…… 432
12·5·1 关于$\{Y^{(1)},\cdots,Y^{(k)}\}$相互独立的检验…… 432
12·5·2 全相关系数检验…… 434
12·5·3 偏相关系数检验…… 434
13 定性变量的相关分析…… 439
13·1 数据分类与列联表…… 439
13·1·1 列联表…… 439
13·1·2 独立分类与连带…… 442
13·1·3 有序表…… 442
13·2 二维定性变量的相关分析…… 442
13·3 2×2列联表的*Fisher*检验…… 450
13·4 相关(连带)测度…… 453
13·4·1 *Goodman*(古德曼)的λ测度…… 453
13·4·2 *Goodman*的r测度…… 455
13·5 多维定性变量的相关分析…… 455
13·5·1 谱系模型…… 456
13·5·2 频数期望估值的计算方法…… 461
13·5·3 最小判别信息统计量…… 464
13·5·4 信息分析…… 465
13·5·5 2×2列联表的信息分析…… 466
13·5·6 应用实例…… 471
14 回归分析…… 483

14·1 回归模式与预测功能……483
14·2 一元线性回归……484
14·2·1 线性模型……485
14·2·2 回归系数a、b的最小二乘估计……485
14·2·3 估计量a、b的性质……486
14·2·4 线性假设的显著性检验……488
14·2·5 回归直线拟合精度检验……489
14·2·6 预测和控制……491
14·2·7 计算步骤……493
14·3 多元线性回归……497
14·3·1 线性模型……497
14·3·2 回归系数估计……497
14·3·3 回归方程的显著性检验……502
14·3·4 回归系数的显著性检验……504
14·3·5 预测和控制……505
14·3·5 应用实例……507
14·4 一元非线性回归（函数形式已知）……510
14·4·1 曲线回归化为直线回归……510
14·4·2 相关指数……512
14·4·3 选配曲线的常见类型……513
14·5 多项式回归（函数形式未知）……515
14·5·1 一元多项式回归……515
14·5·2 多元多项式回归……515
14·6 选择回归模型的最小信息准则……516
14·7 一元正交多项式回归……519
14·7·1 正交多项式……519
14·7·2 一元正交多项式回归……521
14·8 二元正交多项式回归……527
14·8·1 二元正交多项式……527
14·8·2 二元正交多项式回归……528
14·8·3 显著性检验……532
14·8·4 计算步骤……533

15 主分量分析（矩阵数据分析法）……540

15·1 预备知识……540
15·2 主分量……542
15·3 样本主分量……547
15·4 原指标变量对主分量的回归……550
15·5 主分量分析的一般方法步骤……552

15·5·1 选择主分量……552
15·5·2 指标分类……553
15·5·3 样品分类……554
15·6 应用：产品预测……555

16 时间序列的预测与滤波……563

16·1 时间序列的基本概念……563
16·1·1 时间序列及其分类……563
16·1·2 随机序列的数学表征……563
16·2 平稳时间序列……564
16·2·1 广义平稳时间序列……565
16·2·2 自协方差函数与自相关函数……565
16·2·3 功率谱……566
16·2·4 白噪声序列……566
16·2·5 遍历性……567
16·3 线性随机模型……567
16·3·1 ARMA(p,q)模型……567
16·3·2 ARMA(p,q)模型的性质……569
16·3·3 ARMA(p,q)模型的实用性……573
16·4 确定型时间序列预测技术……573
16·4·1 预测模型……573
16·4·2 移动平均法……573
16·4·3 指数平滑法……574
16·4·4 时间回归法……576
16·4·5 季节周期预测法……576
16·5 随机型时间序列预测技术……581
16·5·1 预测模型……581
16·5·2 样本自相关与样本偏相关函数……583
16·5·3 模型识别……585
16·5·4 模型参数估计……586
16·5·5 模型检验……588
16·5·6 最小方差预报……590
16·5·7 应用实例……596
16·6 用最大熵谱提取隐含周期信息……600
16·6·1 最大熵谱估计……600
16·6·2 AR(p,0)模型参数的Burg估计……603
16·6·3 确定显著周期……605
16·6·4 估计周期波形……606
16·7 时间序列的滤波……606

16·7·1 方差分析滤波……606
16·7·2 Fourier周期分析滤波……611

17 响应面方法学……615

17·1 一次回归的正交设计……615
17·1·1 对因子水平进行编码……615
17·1·2 编码空间上的线性回归模型……616
17·1·3 回归方程的拟合检验……618
17·1·4 一次回归正交设计的旋转性……620
17·1·5 包含交互效应的非线性模型……621
17·2 二次回归的正交设计……621
17·2·1 组合设计……622
17·2·2 组合设计的统计分析……627
17·3 响应面方法……630
17·3·1 一阶策略……630
17·3·2 二阶策略……632

主要参考文献 ……638

附 录 常用数理统计
表1 二项分布表…… 1
表2 泊松(Poisson)分布表…… 3
表3 正态分布…… 6
表4 χ^2分布的上侧分位数(x_α^2)表…… 7
表5 t 分布的双侧分位数(t_α)表 …… 8
表6 F检验的临界值(F_α)表…… 9
表7 极差系数d_n和极差分布的分位数表…… 28
表8 随机数表…… 14
表9 χ^2分布 β检验…… 18
表10 t 分布 β 检验表…… 16
表11 t 分布 β 检验表…… 17
表12 F分布 β 检验表…… 18
表13 游程总数检验表 …… 22
表14 $X=\mathrm{Sin}^{-1}\sqrt{P}$变换…… 24
表15 符号检验表…… 26
表16 检验相关系数$\rho=0$的临界值(r_α)表…… 27
表17 r与Z的换算表…… 27
表18 正交多项式表…… 28
表19 正交表…… 31
表20 柯尔莫哥洛(Колмогоров)检验的临界值($D_{n\alpha}$)表…… 23

10 方差分析

在生产过程中，影响产品质量或产量的因素往往是很多的。例如在化工生产中有原料成分、原料剂量、催化剂种类、反应温度、压力、溶液浓度、反应时间、机器设备以及操作技术等因素。每一个因素的变化都会对产品的质量和产量产生影响，其中有些因素影响大一些，有些因素影响小一些。为了能够达到高产、优质、低消耗，就需要搞清哪些因素影响显著，哪些因素影响不显著，哪些因素之间存在着有利的或不利的交互作用，以便将显著因素控制到使产量和产品质量达到最高的水平上，将那些不显著因素控制到消耗、人力、物力、时间最少的水平上。方差分析就是解决这些问题的一种有效方法。方差分析与试验设计有关。下面我们介绍一下单因子和双因子的方差分析。

10.1 一元方差分析

在第2章中曾介绍用t检验法来检验两个等方差(未知)正态总体的均值是否相等的问题，假如有n个等方差（未知）正态总体，如何检验它们的均值是否相等？用上述方法就十分繁琐，而一元方差分析是一种简单有效的方法。

10·1·1 试验设计与试验数据的结构

为了提高某化工产品的产量，从理论分析和实践经验可知，反应温度可能对化工产品的得率有较大影响，为此根据生产经验选择5个温度水平：60℃，65℃，70℃，75℃，80℃做对比试验，每个水平下重复试验3次，以考察温度对产品得率的影响特性。在试验方案确定之后，我们通过试验得到结果（得率）如下：

温　度（℃）	60	65	70	85	80
得　率（%）	90	97	96	84	84
	92	93	96	83	86
	88	92	93	88	82
平均得率（%）	90	94	95	85	84

在上述试验中，我们要求除了温度按方案改变之外，其它条件尽力保持一样，以保证试验数据的质量。

一般地，将考察的因素用字母A表示，试验方案所选择的水平用A_1、A_2、A_3、A_4、A_5表示试验数据用xij（i＝1，…，5），（j＝1，2，3）来表示。为了分析的方便我们将试验结果按试验方案列出表格：

试验号 \ 因子	A_1	A_2	A_3	A_4	A_5
1	x_{11}	x_{21}	x_{31}	x_{41}	x_{51}
2	x_{12}	x_{22}	x_{32}	x_{42}	x_{52}
3	x_{13}	x_{23}	x_{33}	x_{43}	x_{63}

x_{11}，x_{12}，x_{13}即98，92，88之间有差异应该是不可控的那些因子影响的结果，表现为随机误差，因此我们可以将它们表示为：

$$x_{1j}=\mu_1+\varepsilon_{1j}\quad(j=1,\ 2,\ 3),\ \varepsilon_{1j}\in N(0,\ \sigma^2).$$

其中μ_1表示A_1-60°和其它的稳定的一组试验条件所决定的得率，ε_{1j}表示第j次试验中一些随机因子的总效应．由中心极限定理我们可以认为它近似地服从正态分布$N(0,\ \sigma^2)$。

同样，其它4列数据我们也同样分析，只是在一组基本试验条件中温度水平不同，它只影响均值不会影响方差σ^2（若温度水平在试验中能够控制不变化时），因此：第i列即相应A_i水平下的数据可表示为：

$$x_{ij}=\mu_i+\varepsilon_{ij},\ (i=1,\ \cdots,\ 5;\ j=1,\ 2,\ 3),\ \varepsilon_{ij}\in N(0,\ \sigma^2)$$

我们要考察的是不同温度水平下得率的理论指标值μ_1，…，μ_5是否有显著差异．若都相等$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$，则说明温度对得率无影响，若有微小差异说明温度对得率影响不大，若有显著差异（即至少有一μ_i与其它μ_j差异显著）说明温度对得率的影响是显著的．但是我们只知道x_{ij}而不知道μ_i（$i=1$，…，5）故此，问题归结为根据试验数据检验$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5$是否成立。

为了便于分析，我们进一步地把μ_i分解为：

$$\mu_i=\mu+\alpha_i,\ (i=1,\ \cdots,\ 5)$$

其中 $\mu=\frac{1}{5}\sum_{i=1}^{5}\mu_i$，　$\alpha_i=\mu_i-\mu$

μ表示5组不同试验条件对得率贡献的平均值，作为度量A_i效应的一个最为方便的起始点，而 α_i表示A_i水平效应的度量．由于μ的选择的特殊性，必然有$\sum_{i=1}^{5}\alpha_i=0$；当$\mu_1=\mu_2=\mu_3=\mu_4=\mu_5=\mu$时，显然有 $\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha_5=0$．于是最后归结为：根据样本数据检验 $\alpha_i=0$（$i=1$，2，3，4，5）是否成立。

12·1·2 参数估计

从数据模型$x_{ij}=\mu_i+\varepsilon_j=\mu+\alpha_i+\varepsilon_{ij}$（$i=1$，…，5；$j=1$，2，3）

可以看出，每个数据x_{ij}都携带着温度A的影响 α_i和随机误差ε_{ij}的信息，对我们有用的信息是 α_i（$i=1$，…，5）．由参数的估计理论可知：μ_i，μ、α_i的无偏最优估计量分别为：

$$\hat{\mu_i}=\bar{X}_i=\frac{1}{3}\sum_{j=1}^{3}X_{ij},\quad \hat{\mu}=\bar{\bar{X}}=\frac{1}{5}\sum_{i=1}^{5}\bar{X}_i=\frac{1}{15}\sum_{i=1}^{5}\sum_{j=1}^{3}X_{ij},$$

$$\hat{\alpha}_i=\bar{X}_i-\bar{X},\quad (i=1,\cdots,5)$$

$\because$ $x_{ij}=\mu+\alpha_i+\varepsilon_{ij}$ $\quad (i=1,\cdots,5,\ j=1,2,3)$

$\therefore$ $\bar{X}_i=\mu+\alpha_i+\bar{\varepsilon}_i\ (i=1,\cdots,5)$ $\qquad \bar{\varepsilon}_i=\frac{1}{3}\sum_{j=1}^{3}\varepsilon_{ij}$

$$\bar{X}=\mu+\bar{\varepsilon}。\qquad \bar{\varepsilon}=\frac{1}{5}\sum_{i=1}^{5}\bar{\varepsilon}_i=\frac{1}{15}\sum_{i=1}^{5}\sum_{j=1}^{3}\varepsilon_{ij}。$$

$\because$ $\varepsilon_{ij}\sim N(0,\ \sigma^2)$, $\quad\therefore$ $\bar{\varepsilon}_i\sim N(0,\ \frac{\sigma^2}{3})$, $\bar{\varepsilon}\sim N(0,\ \frac{\sigma^2}{15})$

$\hat{\alpha}_i=\alpha_i+(\bar{\varepsilon}_i-\bar{\varepsilon})$, $\quad$ 即 $\hat{\alpha}_i$中包含了 α_i的信息。

10·1·3 显著性检验

H_0: $\alpha_1=\alpha_2=\alpha_3=\alpha_4=\alpha_5=0$

选择统计量，要找一个量能同时包含5个信息 α_i是否同时为0，这样对处理问题是方便的，显然 $3\sum_{i=1}^{5}\hat{\alpha}_i^2$ 是合适的，因为 $\hat{\alpha}_i$包 α_i的信息，而 $\sum_{i=1}^{5}\alpha_i^2=0$ 等价于 H_0成立。（3倍是为了计算的方便）。

但是 $3\sum_{i=1}^{5}\hat{\alpha}_i^2=3\sum_{i=1}^{5}(\bar{X}_i-\bar{X})^2=3\sum_{i=1}^{5}\left[\alpha_i+(\bar{\varepsilon}_i-\bar{\varepsilon})\right]^2$ 之中包含有A的效应和随机效应两部分，可进一步分解为

令 $$S_A=3\sum_{i=1}^{5}\hat{\alpha}_i^2=3\sum_{i=1}^{5}\alpha_i^2+3\sum_{i=1}^{5}(\bar{\varepsilon}_i-\bar{\varepsilon})^2+6\sum_{i=1}^{5}\alpha_i(\bar{\varepsilon}_i-\bar{\varepsilon}),$$

$$E(S_A)=3\sum_{i=1}^{5}\alpha_i^2+(5-1)\sigma^2$$

S_A之中包含A的效应 $3\sum_{i=1}^{5}\alpha_i^2$ 和4 σ^2的一无偏估计量，当H_0成立时，即 $\sum_{i=1}^{5}\hat{\alpha}_i^2=0$，

S_A应是4 σ^2的估计量，但 σ^2是未知的，若能通过其它途径估计出 σ^2，就可以来判别 $\sum_{i=1}^{5}\alpha_i^2$ 是否显著的异于0。

由于每个水平下有重复试验，因而估计 σ^2是不困难的，我们知道：X_{i1}，X_{i2}，X_{i3}是 $N(\mu_i,\ \sigma^2)$的随机样本，$S_i^2=\frac{1}{2}\sum_{j=1}^{3}(\bar{X}_{ij}-\bar{X}_i)^2=\hat{\alpha}^2$。$(i=1,\cdots,5)$

为了提高 σ^2的估计精度，采用全部数据的信息，

$$S_e=\sum_{i=1}^{5}\sum_{j=1}^{3}(X_{ij}-\bar{X}_i)^2=\sum_{i=1}^{5}\sum_{j=1}^{3}(\varepsilon_{ij}-\bar{\varepsilon}_i)^2$$

$$E(S_e)=E\left[\sum_{i=1}^{5}\sum_{j=1}^{3}(\varepsilon_{ij}-\bar{\varepsilon}_j)^2\right]=\sum_{i=1}^{5}E\left[\sum_{j=1}^{3}(\varepsilon_j-\bar{\varepsilon}_i)^2\right]=\sum_{i=1}^{5}2\sigma^2=10\sigma^2$$

通常我们称SA的自由度为5－1＝4是指，SA是用含有随机误差的数据进行运算，其结果也必含有随机误差，含有一份总体 σ^2 自由度就是1。该例含有4倍 σ^2，则自由度为4（即水平数减1）而S_e的自由度为10，〔一般地为：全部试验次数减去A的水平数〕

最后我们确定检验H_0是否成立的统计量为：

$$F=\frac{\bar{S}_A}{S_e},\qquad \bar{S}_A=\frac{S_A}{4},\qquad \bar{S}_e=\frac{S_e}{10}.$$

当H_0成立时：$F\sim F〔4，10〕$。当H_0不成立时，SA值增大，∴F不再服从F〔4，10〕，应该服从峰值（F的大概率密集区间）向右偏移的概率分布。即应向右偏离F〔4，10〕。要用F的一次实现值来判定这种分布的变异性是否显著（即$\sum_{i=1}^{5}\alpha_i^2$是否显著大于0）。就是第2章中介绍的F－检验法。

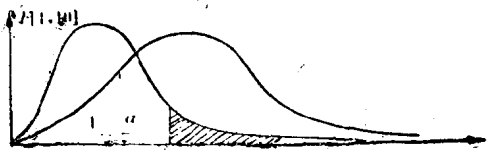


图10—1

由统计量 $F=\dfrac{\left(\frac{3}{4}\sum_{i=1}^{5}\alpha_i^2\right)+\left(\hat{\sigma}^2\right)_1}{\left(\hat{\sigma}^2\right)_2}$（$(\hat{\sigma}^2)_1$，$(\hat{\sigma}^2)_2$为$\sigma^2$的两个估计量）

不难看出，当F取值落在F〔3，10〕右端时应否定H_0，当给定显著性水平α时，可查表F〔4，10〕得到$F_\alpha〔4，10〕$，否定域为（$F_\alpha〔4，10〕$，$+\infty$）α为否定错的概率。当F值落在（0，$F_\alpha〔4，10〕$）之内，只能做出不能否定H_0的结论，结论可靠性难以给出。

结合本例进行计算如下：

$$S_A=3\times〔(90-89.6)^2+(94-89.6)^2+(95-89.6)^2+(85-89.6)^2+(84-89.6)^2〕=303.6.$$

$$\begin{aligned}S_e=&〔(90-90)^2+(92-90)^2+(88-90)^2〕+〔(97-94)^2+(93-94)^2+(92-94)^2〕\\&+〔(96-95)^2+(96-95)^2+(93-95)^2〕+〔(84-85)^2+(83-85)^2+(88-85)^2〕\\&+〔(84-84)^2+(86-84)^2+(82-84)^2〕=8+14+6+14+8=50\end{aligned}$$

$$\bar{S}_A=\frac{1}{4}S_A=75.9.\quad \bar{S}_e=\frac{1}{10}S_e=5.\quad F=\frac{\bar{S}_A}{S_e}=15.18.\quad F_{0.01}〔4，10〕=6.$$

由于F＞6，∴应否定H_0，即有99%的把握断定温度对产品得率有高度显著的影响。

10·1·4 计算方法和步骤

考察因子A对指标X的影响，取m个水平，每个水平下重复试验n_i次（$i=1，\cdots，m$），试验结果为：

A_1	A_2	A_3	…	A_m
x_{11}	x_{21}	x_{31}		m_1
⋮	⋮	⋮		⋮
x_{1n_1}	x_{n2}	x_{3n_3}		x_{mr_m}

$n_1+n_2+\cdots+n_m=N$

（1）数据检验

方差分析要求数据满足的条件是：

服从正态分布：$x_{ij}\sim N(\mu_i,\ \sigma_i^2)$，$(i=1\cdots m)$；（记$\mu=\frac{1}{m}\sum_{i=1}^{m}\mu_i$）

方差齐性：$\sigma_1^2=\sigma_2^2=\cdots=\sigma_m^2=\sigma^2$。

独立性：所有数据都是独立抽取的。

如果试验数据明显地不满足以上条件，显著性检验的结论是不可靠的。

（2）信息提取——因子效应和随机误差估计

全部数据总的变动平方和（记为ST）的分解：

$$ST=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2=\sum_{i=1}^{m}\sum_{i=1}^{n_i}[(X_{ij}-\bar{X}_i)+(\bar{X}_i-\bar{X})]^2$$

$$=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{X}_i)^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{X}_i-\bar{X})^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{X}_i)^2+\sum_{i=1}^{m}n_i(\bar{X}_i-\bar{X})^2$$

其中：$\bar{X}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij}\ (i=1,\ \cdots,\ m)$，$\bar{X}=\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}X_{ij}=\frac{1}{N}\sum_{i=1}^{m}n_i\cdot\bar{X}_i$。

记$S_A=\sum_{i=1}^{m}n_i(\bar{X}_i-\bar{X})^2$，$E(S_A)=(m-1)\sigma^2+\sum_{i=1}^{m}n_i\alpha_i^2\ (\mu_i-\mu=\alpha_i)$。

$$S_e=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij}-\bar{X}_i),\quad E(S_e)=(N-m)\sigma^2$$

则：$S_T=S_A+S_e$。

S_A中含有A的效应信息，S_e含有纯随机效应信息，

记S_A的自由度为f_A，有$f_A=m-1$。

S_e的自由度为f_e，有$f_e=N-m$。

S_T的自由度为f_T，有$f_T=f_A+f_e=N-1$。

由此可得到方差估计：

$$\bar{S}_e=S_e/f_e=\hat{\sigma}^2.\qquad \bar{S}_A=S_A/f_A=\hat{\sigma}^2+\frac{1}{m-1}\sum_{i=1}^{m}n_i\alpha_i^2.$$

（3）根据样本数据提取的信息进行显著性检验，

H_0：$\alpha_1=\cdots=\alpha m=0$

选用统计量$F=\frac{\bar{S}_A}{S_e}$，当H_0成立时，$F\sim F[m-1,\ N-m]$。

给定显著性水平α。（$0<\alpha<1$）查表求$F\alpha[m-1,\ N-m]$。

确定H_0的否定域：〔$F\alpha(m-1, N-m)$，$+\infty$〕。

记F的实现值为$F_0=\bar{S}_A/\bar{S}_e$。若$F_0>F\alpha〔m-1, N-m〕$，则拒绝H_0，有$1-\alpha$的把握断定因子A对指标影响显著，反之，认为A对指标影响不显著。

取$\alpha=1\%$为高度显著水平，$\alpha=5\%$为显著水平，$\alpha=10\%$为不太显著水平。

（4）为了实际工作者分析方便将上述方去步骤表格化：

	A_1	A_2	…	A_i	…	A_m	Σ
	x_{11} x_{12} ⋮ x_{1n_1}	x_{21} x_{22} ⋮ x_{2n_2}	…	x_{i1} x_{i2} ⋮ x_{ini}	…	x_{m1} x_{m2} ⋮ x_{mn_m}	
Σ	$\sum_{j=1}^{n_1} x_{1j}$	$\sum_{j=1}^{n_2} x_{2j}$	…	$\sum_{j=1}^{n_i} x_{ij}$	…	$\sum_{j=1}^{n_m} x_{mj}$	$\sum_{i=1}^{m}\sum_{t=1}^{n_i} x_{ij}$
$\frac{1}{n_i}(\Sigma)^2$	$\frac{\left(\sum_{j=1}^{n_1} x_{1j}\right)^2}{n_1}$	$\frac{\left(\sum_{j=1}^{n_2} x_{2j}\right)^2}{n_2}$	…	$\frac{\left(\sum_{j=1}^{n_i} x_{ij}\right)^2}{n_i}$	…	$\frac{\left(\sum_{t=1}^{n_m} x_{mj}\right)^2}{n_m}$	$\sum_{i=1}^{m}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} x_{ij}\right)^2$
Σ^2	$\sum_{j=1}^{n_1} x_{1j}^2$	$\sum_{j=1}^{n_2} x_{2j}^2$	…	$\sum_{j=1}^{n_i} x_{ij}^2$	…	$\sum_{j=1}^{n_m} x_{mj}^2$	$\sum_{i=1}^{m}\sum_{j=1}^{n_m} x_{mj}^2$

令$CP=\frac{1}{N}\left(\sum_{i=1}^{m}\sum_{i=1}^{n_i} x_{ij}\right)^2$，　　$N=\sum_{i=1}^{m} n_i$。

$$Q=\sum_{i=1}^{m}\frac{1}{n^i}\left(\sum_{!=1}^{n_i} x_{ij}\right)^2$$

$$R=\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}{}^2$$

$S_A=Q-CT$；　　$S_e=R-Q$，　　$S_T=R-CT$；

它们的自由度：$f_A=m-1$，$f_e=N-m$，$f_T=N-1$；

方差分析表

因　子	偏差平方和	自 由 度	均方估计	检验统计量值	显著性
A	S_A	$f_A=m-1$	$\bar{S}_A=S_A/f_A$	$F=\frac{\bar{S}A}{Se}$	
E	S_e	$f_e=N-m$	$\bar{S}_e=S_e/f_e$		
总	S_T	$f_T=N-1$			
临 界 值	$F\alpha〔m-1,N-m〕$；		（**）高度显著，（*）显著。		

（5）预测：

（a）显著性检验的结论：A不显著.

此时 $\alpha_i=0$，（$i=1$，…，m）；$x_{ij}=\mu+\varepsilon_{ij}$，（$i=1$，…，$m$；$j=1$，…$n_i$）

点估计：$\hat{\mu}=\bar{x}=\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{ni}x_{ij}$.

μ的 $1-\alpha$的置信区间：$\left[\bar{x}-t_{\alpha/2}^{(n-1)}\sqrt{\frac{S}{N(N-1)}},\ \bar{x}+t_{\alpha/2}^{(n-1)}\sqrt{\frac{S}{N(N-1)}}\right]$

（b）显著性检验的结论：A显著。

点估计：$\hat{\mu}_i=\hat{\mu}+\hat{\alpha}_i=\overline{X}_i$　　（$i=1$，…，m）．$\overline{X}_i=\frac{1}{n_i}\sum_{j=1}^{ni}x_{ij}$.

μ_i的 $1-\alpha$的置信区间：

$$\left[\overline{X}_i-\sqrt{\frac{\overline{S_e}F\alpha(1,N-m)}{n_i}},\ \overline{X}+\sqrt{\frac{\overline{S_e}F\alpha\langle 1,N-m\rangle}{n_i}}\right]$$

〔注〕

可以证明$\hat{\mu}_i$（$i=1$，…，m）与S_e相互独立．又知，$\overline{X}_i-\mu_i\sim N(0,\frac{\sigma^2}{n_i})$，

$$\frac{\sqrt{n_i}(\bar{x}_i-\mu_i)}{\sigma}\sim N(0,1).\quad \frac{n_i(\bar{x}_i-\mu_i)^2}{\sigma^2}\sim\chi^2(1),\quad \frac{S_e}{\sigma^2}\sim\chi^2(N-m);$$

$$\therefore\frac{n_i(\bar{x}_i-\mu_i)^2}{\sigma^2}\Big/\frac{S_e}{\sigma^2(N-m)}=\frac{n_i(\bar{x}_i-\mu_i)^2}{s_e/N-m}\sim F_i\langle 1,N-m\rangle.$$

给定置信水平 α，就有

$$P\left[\frac{n_i(\bar{x}_i-\mu_i)^2}{s_e/N-m}<F\alpha\langle 1,N-m\rangle\right]=1-\alpha,\quad\left(\bar{se}=\frac{se}{N-m}\right)$$

亦即：$P\left[|\bar{x}_i-\mu_i|<\sqrt{\frac{\overline{S_e}\cdot F\alpha\langle 1,N-m\rangle}{n_i}}\right]=1-\alpha$.

故：$P\left[\bar{x}_i-\sqrt{\frac{\overline{S_e}F\alpha\langle 1,N-m\rangle}{ni}}<\mu_i<\bar{x}_i+\sqrt{\frac{S_e\cdot F\alpha\langle 1,N-m\rangle}{n_i}}\right]=1-\alpha$.

10·1·5　应用举例：

例 1　某工厂的锻造零件是由三个地区制造的。该零件的重要特性是强度，质量差异大就成问题。假若不同地区制造的零件强度有较大差异，就应该从中选择一个合适的地区制造的零件。为此三个地区的零件中各随机抽取 4 个零件，由同一个操作者同一试验按随机的顺序进行强度试验（破坏性试验）。试验结果如下（单位：100kg）

地区 \ 序号	1	2	3	4	Σ	$\overline{X}$(平均)	总　体
A_1	115	116	98	83	412	103	$N(\mu_1,\sigma^2)$
A_2	103	107	118	116	444	111	$N(\mu_2,\sigma^2)$
A_3	73	89	85	97	344	86	$N(\mu_3,\sigma^2)$

方差分析表

因　子	偏差平方和	自 由 度	均方估计	F	显 著 性
A	$S_A=1304$	$f_A=2$	$\bar{S}_A=652$	$F=4.92$	*
E	$S_e=1192$	$f_e=9$	$S_e=132.4$		
T	$S_T=2496$	$f_T=11$			

$F_{0.005}(2,9)=4.26$

结论：有95%的把握断定不同地区的零件强度是有显著差异的。

点估计：$\hat{\mu}_1=103$，$\hat{\mu}_2=111$，$\hat{\mu}_3=86$；

区间估计：$\alpha=0.05$，$F_{0.05}(1,9)=5.12$　$S_e=132.4$，$n=4$。

$$R=\sqrt{\frac{\bar{S}e\cdot F_{0.05}}{n}}=\sqrt{\frac{132.4\times5.12}{4}}\approx13.$$

$P\{90<\mu<116\}=95\%$；$P\{98<\mu_2<124\}=95\%$，$P\{73<\mu_3<99\}=95\%$。

例 2　研究某种大型机械的新闻广告的效果。

广告A_1：强调运输的方便性。

广告A_2：强调燃料节省的经济性。

广告A_3：强调噪音低的优良性。

在广告广泛宣传后，按寄回的广告上的订购数计算，一年4个季度的销售量为：

季度＼广告	A_1	A_2	A_3	Σ
1	163	184	206	
2	176	198	191	
3	170	179	218	
4	185	190	224	
Σ	694	751	839	2284
$(\Sigma)^2$	481636	564001	703921	1749558=4Q
Σ^2	120670	141201	176617	438488=R

$$C_T=\frac{2284^2}{12}=434721.33.\qquad S_T=R-CT=3766.67.$$

$$S_A=Q-C_T=\frac{1749558}{4}-C_T=2668.17.\quad S_e=S_T-S_A=1098.5$$

方差分析表

因　子	S	f	$\bar{S}$	F	$F^{\alpha}(2,9)$
A（广告形式）	2668.17	2	1334.09	10.93 * *	$F_{0.05}=4.16$
E（随机误差）	1098.5	9	122.06		$F_{0.01}=8.02$
T（合计）	3766.67	11			

结论：由于广告内容不同对销售量影响很大。

不同广告引起的销售量的估计：

$$\hat{\mu}_{A_1}=\frac{694}{4}=173.5.\quad \hat{\mu}_{A_2}=\frac{751}{4}=187.75.\quad \hat{\mu}_{A_3}=\frac{839}{4}=209.75.$$

95%的置信区间分别为：（略去小数）

μ_{A_1}：〔161，186〕；μ_{A_2}：〔175，200〕，μ_{A_3}：〔197，222〕。

由此可以看出广告A_3引起的销售量最多，今后应多宣传噪音低的优良性，对多销售有利，同时今后应进一步进行工艺改革降低噪音。

例3 为了延长切削刀具的寿命，考虑到后角有影响，为此进行试验，通过分析找出最适宜的角度。现行后角2°，再选择3°、4°，各做4次试验，磨损量的记录如下：

	$A_1(2^0)$	$A_2(3^0)$	$A_3(4^0)$	Σ
1	0.52	0.45	0.42	
2	0.55	0.41	0.41	
3	0.53	0.43	0.43	
4	0.55	0.43	0.42	
Σ	2.15	1.72	1.68	5.55
$(\Sigma)^2$	4.6225	2.9584	2.8224	$10.4033=4Q$
Σ^2	0.1563	0.7404	0.7058	$2.6025=R$

$$C_T=\frac{5.55^2}{12}=2.5669.\qquad S_T=R-P=2.6025-2.5669=0.0356.$$

$$S_A=Q-P=\frac{10.4033}{4}-2.5669=0.0339,\ S_e=S_T-S_A=0.0017.$$

方差分析表

因　子	S	f	$\bar{S}$	F	$F\alpha〔2,9〕$
A（后角）	0.0339	2	0.01695	89.21^{**}	$F_{0.05}=4.26$
E（误差）	0.0017	9	0.00019		$F_{0.01}=8.02$
T（合计）	0.0356	11			

结论：后角不同对刀具寿命有很大的影响。不同角度损耗量的估计：只估μ_{A_2}，μ_{A_3}，（因μ_{A_1}较大）

$$\hat{\mu}_{A_2}=\frac{1.72}{4}=0.43.\ \mu_{A_2}95\%\text{的量信区间：}〔0.414,\ 0.446〕.$$

$$\hat{\mu}_{A_3}=\frac{1.68}{4}=0.42,\ \mu_{A_3}95\%\text{的量信区间：}〔0.404,\ 0.436〕.$$

由此可知：切削刀具的磨损量最少的为后角4°。

例4 某印刷厂考察染整工艺对缩水率的影响。在六种不同染整工艺下各进行了四次试验（布样相同）。测得数据如下：（%）