统计基础

源头问题与当今应用

- 统计基础知识

1. 位置统计量
   - 均值
   - 顺序统计量
   - 中位数
   - 百分位数

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

$$m_e = \begin{cases} x_{(\frac{n+1}{2})}, & \text{当}n\text{为奇数时,} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{当}n\text{为偶数时.} \end{cases}$$

$$m_p = \begin{cases} x_{(\lfloor np \rfloor+1)}, & \text{当}np\text{不是整数时} \\ \frac{1}{2}(x_{(np)} + x_{(np+1)}), & \text{当}np\text{是整数时.} \end{cases}$$

2.分散程度度量

样本方差 $s^{*2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

样本标准差 $s^*=\sqrt{s^{*2}}$, $s=\sqrt{s^2}$

变异系数 $CV=\frac{s}{\bar{x}}\times 100\%.$

极 差 $R=x_{(n)}-x_{(1)}=max(x)-min(x)$

样本标准误差 $s_m=\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{s}{\sqrt{n}}$

3.分布形状度量

K阶原点矩 $a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$

K阶中心矩 $b_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$

样本偏度 $\gamma_1 = b_3/b_2^{3/2}$

刻画数据的对称性指标，关于均值对称则偏度为0.

样本峰度 $\gamma_2 = \frac{b_4}{b_2^2} - 3$

反映总体分布密度曲线在其峰值附近的陡峭程度.

4.统计中几个重要的概率分布

正态分布 $N(\mu, \sigma^2)$，$N(0,1)$ 为标准正态分布.

- 若 $X_1, X_2, \cdots, X_n$ 为标准正态分布，那么 $\sum_{i=1}^{n} x_i^2$ 服从卡方分布 $\chi^2(n)$

$\frac{X_1}{\sqrt{\chi^2/n}}$ 服从 t 分布 $t(n)$

若 $S_1^2 \sim \chi^2(n), S_2^2 \sim \chi^2(m)$，则 $\frac{S_1^2/n}{S_2^2/m}$ 服从 $F(m,n)$

5.统计量分布

单个正态总体情形 $X \sim N(\mu, \sigma^2)$

则：$\overline{x} \sim N(\mu, \frac{\sigma^2}{n})$

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

$\frac{\overline{x}-\mu}{S/\sqrt{n}} \sim t(n-1)$

两个正态总体情形 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$

则：$\frac{(\overline{x}-\mu_1)-(\overline{y}-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$

$\frac{(\overline{x}-\mu_1)-(\overline{y}-\mu_2)}{S/\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2)$ 其中 $S = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$

- 6.1.2 参数估计

点估计定义：设 $x_1, x_2, \ldots, x_n$ 是来自于同一个样本，用于估计未知参数θ的统计量$\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$称为$\theta$的估计量，或称为$\theta$的点估计。

- 点估计的常用两种方法：

矩估计 $EX^k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$

极大似然估计 $\hat{\theta} : L(\theta) = \sup_{\theta\in\Theta} \prod_{i=1}^{n} f(x_i;\theta)$

评价估计优劣的标准有无偏性、有效性、一致性（相合性）等

无偏性 $E\hat{\theta} = \theta$

有效性 对无偏估计 $\hat{\theta}_1, \hat{\theta}_2$，若 $D\hat{\theta}_1 \le D\hat{\theta}_2$ 则称 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 有效.

一致性 $\forall \varepsilon > 0, \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$

$\Leftrightarrow \lim_{n\to\infty} E(\hat{\theta}_n) = \theta, \lim_{n\to\infty} D(\hat{\theta}_n) = 0$

•区间估计

当 $P(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) = 1-\alpha$ 时，称 $[\widehat{\theta}_1, \widehat{\theta}_2]$ 为参数 θ 水平为 $1-\alpha$ 置信区间.

单个正态总体情形 $X \sim N(\mu, \sigma^2)$,μ 的双侧置信区间：

方差 σ^2 已知　$[\overline{x} - \mu_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + \mu_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$

方差 σ^2 未知　$[\overline{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1), \overline{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}(n-1)]$

单个正态总体情形 $X \sim N(\mu, \sigma^2)$,σ^2 的双侧置信区间：

μ 已知　$[\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\chi^2_{1-\alpha/2}(n)}]$

μ 未知　$[\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}]$

两个正态总体情形 $X \sim N(\mu_1, \sigma_1^2), Y \sim N(\mu_2, \sigma_2^2)$ 方差均已知时 $\mu 1 - \mu 2$ 的置信区间

$$\left[(\bar{x}-\mu_1)-(\bar{y}-\mu_2)-u_{1-\frac{\alpha}{2}}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2},(\bar{x}-\mu_1)-(\bar{y}-\mu_2)+u_{1-\frac{\alpha}{2}}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}\right]$$

方差均未知但是相等时，$\mu 1 - \mu 2$ 的置信区间

$$[(\bar{x}-\mu_1)-(\bar{y}-\mu_2)-t_{1-\frac{\alpha}{2}}(n_1+n_2-2)S_w\sqrt{1/n_1+1/n_2},(\bar{x}-\mu_1)-(\bar{y}-\mu_2)+t_{1-\frac{\alpha}{2}}(n_1+n_2-2)S_w\sqrt{1/n_1+1/n_2}]$$

其中 $S_w^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$

μ_1, μ_2 均已知时，方差比 σ_1^2/σ_2^2 的置信区间

$$[\frac{\sum_{i=1}^{n_1}(x_i-\mu_1)^2/n_1}{\sum_{i=1}^{n_2}(y_i-\mu_2)^2/n_2}\cdot\frac{1}{F_{\alpha/2}(n_1,n_2)},\frac{\sum_{i=1}^{n_1}(x_i-\mu_1)^2/n_1}{\sum_{i=1}^{n_2}(y_i-\mu_2)^2/n_2}\cdot\frac{1}{F_{1-\alpha/2}(n_1,n_2)}]$$

μ_1, μ_2 均未知时，方差比 σ_1^2/σ_2^2 的置信区间

$$[\frac{s_1^2}{s_2^2}F_{\alpha/2}(n_1-1,n_2-1),\frac{s_1^2}{s_2^2}F_{1-\alpha/2}(n_1-1,n_2-1)]$$

假设检验

- 6.1.3 假设检验

1.单个正态总体均值检验

$$H_0:\mu \le \mu_0, H_1:\mu > \mu_0$$

$$H_0:\mu \ge \mu_0, H_1:\mu < \mu_0$$

$$H_0:\mu=\mu_0, H_1:\mu \ne \mu_0$$

u检验(σ已知)

检验统计量:

$$u = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

情形一: 拒绝域 $\{u \geq u_{1-\alpha}\}$, p值 $1-\Phi(u_0)$

情形二: 拒绝域 $\{u \leq u_{\alpha}\}$, p值 $\Phi(u_0)$

情形三: 拒绝域 $\{|u| \geq u_{1-\alpha/2}\}$, p值 $2\left(1-\Phi(|u_0|)\right)$

t 检验(σ 未知)

检验统计量:

$$t=\frac{\overline{x}-\mu_0}{s/\sqrt{n}}$$

情形一: 拒绝域 $\{t\geq t_{1-\alpha}(n-1)\}$,p 值 $P(t\geq t_0)$

情形二: 拒绝域 $\{t\leq t_{\alpha}(n-1)\}$,p 值 $P(t\leq t_0)$

情形三: 拒绝域 $\{|t|\geq t_{1-\alpha/2}(n-1)\}$,p 值 $P(|t|\geq|t_0|)$

[h,p,ci]=ztest(x,mu,sigma,alpha,tail)

tail=0(缺省)，1，-1 分别对应于上述三种备选假设之一输出参数 $h=0$ 表示接受 H_0，$h=1$ 表示拒绝 H_0，p 表示拒绝原假设 H_0 的最小显著性水平. p 越小 H_0 越值得怀疑，ci 是 μ_0 的置信区间.

σ^2 未知，均值 μ 的检验

[h,p,ci]=ttest(x,mu,alpha,tail)

- 例 某车间用一台包装机包装糖果.包的袋装糖重是一个随机变量，服从正态分布.当机器正常时，其均值为0.5公斤，标准差为0.015公斤.某日开工后为检验包装机是否正常，随机地抽取它所包装的糖9袋，称得净重为(公斤):
  0.497,0.506,0.518,0.524,0.498,0.511,0.520,0.515,0.512.
  问机器是否正常?

$$x \sim N(\mu, 0.015^2), H_0 : \mu = \mu_0 = 0.5 \quad H_1 : \mu \neq 0.5$$

解：Matlab 实现如下：

x=[0.497 0.506 0.518 0.524 0.498 0.511 0.520 0.515 0.512];

[h,p,ci]=ztest(x,0.5,0.015)

求得 h=1，p=0.0248，ci =[0.5014 0.5210]

在0.05水平下可拒绝原假设，即认为这天包装机工作不正常.σ2未知，均值μ的检验

例：据调查，一组元件的寿命为：159 280 101 212 224 379 179 264 222 362 168 250 149 260 485 170，已知该元件寿命复合正态分布，请问该元件的平均寿命是否大于225小时？

我们设原假设$H_0: \mu \leq 225$

则备择假设为$H_1: \mu > 225$

x=[159 280 101 212 224 379 179 264 222 362 168 250 149 260 485 170];

[h,p,ci]=ttest(x,225,0.05,1)

求得 h=0，p=0.2570，ci =[198.2321 Inf]

说明在显著水平为 0.05 的情况下，不能拒绝原假设，认为元件的平均寿命不大于225小时

2.两个正态总体均值检验

（显著性水平为 α）

检验法	条件	H_0	H_1	检验统计量	拒绝域
U 检验	σ_1,σ_2 已知	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$U=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n}+\dfrac{\sigma_2^2}{m}}}$	$\{u\geqslant u_{1-\alpha}\}$ $\{u\leqslant u_\alpha\}$ $\{\lvert u\rvert\geqslant u_{1-\frac{\alpha}{2}}\}$
t 检验	$\sigma_1=\sigma_2=\sigma$ 未知	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$t=\dfrac{\overline{X}-\overline{Y}}{S_W\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$	$\{t\geqslant t_{1-\alpha}(n+m-2)\}$ $\{t\leqslant t_\alpha(n+m-2)\}$ $\{\lvert t\rvert\geqslant t_{1-\frac{\alpha}{2}}(n+m-2)\}$
近似 U 检验	σ_1,σ_2 未知 m,n 充分大	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$U=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}}}$	$\{u\geqslant u_{1-\alpha}\}$ $\{u\leqslant u_\alpha\}$ $\{\lvert u\rvert\geqslant u_{1-\frac{\alpha}{2}}\}$
近似 t 检验	σ_1,σ_2 未知 m,n 不太大	$\mu_1\leqslant\mu_2$ $\mu_1\geqslant\mu_2$ $\mu_1=\mu_2$	$\mu_1>\mu_2$ $\mu_1<\mu_2$ $\mu_1\neq\mu_2$	$t^*=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{m}}}$	$\{t^*\geqslant t_{1-\alpha}(l)\}$ $\{t^*\leqslant t_\alpha(l)\}$ $\{\lvert t^*\rvert\geqslant t_{1-\frac{\alpha}{2}}(l)\}$

3、单个正态总体方差的检验

（显著性水平为 α）

检验法	条件	H_0	H_1	检验统计量	拒绝域
χ^2 检验	μ 未知	$\sigma^2 \leqslant \sigma_0^2$ $\sigma^2 \geqslant \sigma_0^2$ $\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	$\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$	$\{\chi^2 \geqslant \chi^2_{1-\alpha}(n-1)\}$ $\{\chi^2 \leqslant \chi^2_{\alpha}(n-1)\}$ $\{\chi^2 \leqslant \chi^2_{\alpha/2}(n-1)\}$ 或 $\chi^2 \geqslant \chi^2_{1-\frac{\alpha}{2}}(n-1)\}$

4、两个正态总体方差的检验

（μ_1,μ_2 未知，显著性水平为 α）

检验法	H_0	H_1	检验统计量	拒绝域
F 检验	$\sigma_1^2\leqslant\sigma_2^2$	$\sigma_1^2>\sigma_2^2$	$F=\dfrac{S_X^2}{S_Y^2}$	$\{F\geqslant F_{1-\alpha}(n-1,m-1)\}$
	$\sigma_1^2\geqslant\sigma_2^2$	$\sigma_1^2<\sigma_2^2$		$\{F\leqslant F_{\alpha}(n-1,m-1)\}$
	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$		$\{F\leqslant F_{\frac{\alpha}{2}}(n-1,m-1)$或 $F\geqslant F_{1-\frac{\alpha}{2}}(n-1,m-1)\}$

例子

为了比较两种谷物种子的优劣,特选取 10 块土质不全相同的土地，并将每块土地分为面积相同的两部分，分别种植两种种子，施肥与天健管理在20 小块土地上都是一样, 下面是各小块上的单位产量:

土地	1	2	3	4	5	6	7	8	9	10
种子一的单位产量	23	35	29	42	39	29	37	34	35	28
种子二的单位产量	30	39	35	40	38	34	36	33	41	31

由于未知两个分布的方差，因此需要进行t检验，统计量与拒绝域分别是：

$$t_1 = \frac{\bar{x} - \bar{y}}{s_w / \sqrt{n/2}}$$

$$W_1 = \left\{ |t_1| > t_{1-\alpha/2}(2n-2) \right\}$$

其中 $s_w^2=\left(s_x^2+s_y^2\right)/2,\alpha$ 是给定的显著性水平. 由给出的数据可算得

$$\bar{x}=33.1,\quad \bar{y}=35.7,\quad s_x^2=33.2110,\quad s_y^2=14.2333,\quad s_w^2=23.7222,$$

从而可算得两样本的 t检验统计呈的值

$$t_1=\frac{33,1-35\cdot 7}{4.8705/\sqrt{10/2}}=-1.1937$$

若给定 $\alpha = 0.05$ 查表得 $t_{0.975}(18) = 2.1009$

由于 $|t_1| < 2.1009$ 故不应拒绝原假设,即认为两种种子的单位产呈平均值没有显著差别.

此处检验的 p值为 0.2467.

对于两个样本均值的检验，同样有函数
[h,p,ci]=ttest2(x,y,alpha,tail)
故而上面的题目，可以通过matlab代码：

- x=[25,35,29,42,39,29,37,34,35,28];
- y=[30,39,35,40,38,34,36,33,41,31];
- [h,p,ci]=ttest2(x,y,0.05,0)

来进行计算

- 得到的结果为：

```
h =

     0


p =

    0.2641
```

分布未知的假设检验——卡方检验

1、总体可以分为有限类，且总体分布不含未知参数

设总体 X可以分成 r 类,记为 $A_1, A_2, \cdots, A_r$ 如今要检验的假设为:

$$H_0 : p(A_i) = p_i \quad , \quad i = 1, 2, \cdots, r$$

其中各 p_i已知, $p_i \geq 0$ 且 $\sum_{i=1}^{r} p_i = 1$， 现对总体作了n次观察,各类出现的频数分别为 $n_1, n_2, \cdots, n_r$ ， 且 $\sum_{i=1}^{r} n_i = n$ 若 H_0 为真, 则各概率 p_i 与频率n_i / n应相差不大， 或各观察频数n_i与理论频数 np_i 应相差不大。因此提出如下统计量：$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$

并指出,当样本容量n充分大且H_0 为真时， χ^2 近似服从自由度为 r-1的 χ^2 分布。

从 χ^2 统计量的结构看,当 H_0 为真时,和式中每一项的分子 $(n_i - np_i)^2$ 都不应太大,从而总和也不会太大,若 χ^2 过大,人们就会认为原假设H_0 不真。基于此想法,检验的拒绝域应有如下形式: $W = \{\chi^2 \geq c\}$

对于给定的显著性水平 α 由分布 $\chi^2(r-1)$ 可定出 $c = \chi^2_{1-\alpha}(r-1)$ 。

2、总体可以分为有限类，但总体分布含k个未知参数

在上面的基础之上，在总体分布中含有k个独立的未知参数时,若这k个 数用极大似然估计代替,则 p_i 用 p_i 代替,当样本容量n充分大时

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(n_i - n\hat{p}_i\right)^2}{n_i}$$

近似服从自由度为r-k-1的 χ^2 分布。

3、 总体为连续分布的情况

设样本 $X_1, X_2, \cdots, X_n$ 为来自总体 X的一个样本,要检验的假设是:

$$H_0 : X \text{ 服从分布} F(x)$$

其中 $F(x)$ 中可以含有k个未知参数,若 k=0,那F(x) 就完全已知。 在这种情况下检验 H_0 的做法如下:

(1)把X的取值范围分成 r个区间, 为确定起见,不妨设为: $-\infty = a_0 < a_1 < a_2 < \cdots < a_{r-1} < a_r = \infty$

设各区间为 $A_1 = (a_0, a_1], A_2 = (a_1, a_2], \cdots, A_{r-1} = (a_{r-2}, a_{r-1}], A_r = (a_{r-1}, a_r)$

(2)统计样本落入这 r 个区间的频数, 分别记为$n_1, n_2, \cdots, n_r.$ 这里要求各 $n_i \geq 5$;

(3)当$k \neq 0$时,对k个未知参数给出其极大似然估计,记 $p_i = P\langle a_{i-1} < X \leq a_i) = F(a_i) - F(a_{i-1})$

从而用未知参数的极大似然估计代替后可算得各 p_i 这样就把检验问题转化为分类数据的检验问题,以后的计算同 前面两个小节视未知参数个数 k=0或 $k \neq 0$ 而定

例子:

为研究混凝土抗压强度的分布,抽取了 200 件混凝土制件测定其抗压强度,经整理得频数分布表如下见表。 试在$\alpha = 0.05$水平上检验抗压强度的分布是否为正态分布。

抗压强度区间 $(a_{i-1}, a_i]$	频数 n_i
$(190, 200]$	10
$(200, 210]$	26
$(210, 220]$	56
$(220, 230]$	64
$(230, 240]$	30
$(240, 250]$	14
合计	200

解:若用 $F(x)$ 表示 $N(\mu,\sigma^2)$ 的分布函数,则本例便要检验假设: H_0 : 抗压强度的分布为$F(x)$

又由于F(x)中含有两个未知参数μ与 σ^2因而需用它们的极大似然估计去替代。这里仅给出了样本的分组数据,因此只能用组中值(即区间中点)去代替原始数据,然后求μ与 σ^2的 MLE。现在 6 个组中值分别为 $x_1=195, x_2=205, x_3=215, x_4=225, x_5=235, x_6=245$

$$\hat{\mu}=\bar{x}=\frac{1}{200}\sum_{i=1}^{6}n_i x_i=221$$

于是

$$\hat{\sigma}^2=s_n^2=\frac{1}{200}\sum_{i=1}^{6}n_i\left(x_i-\bar{x}\right)^2=152,\quad \hat{\sigma}=s_n=12.33$$

在 N(221,152)分布下,求出落在区间 $(a_{i-1}, a_i]$ 内的概率的估计值:

$$p_i = \Phi\left(\frac{a_i - 221}{\sqrt{152}}\right) - \Phi\left(\frac{a_{i-1} - 221}{\sqrt{152}}\right), \quad i = 1, 2, \cdots, r$$

不过常将 a_0 定为 $-\infty$，将 a_r 定为 $+\infty$, 本例中 r=6。采用 $\chi^2 = \sum_{i=1}^{r} \frac{(n_i - n\hat{p}_i)^2}{n_i}$ 作为检验统 计量,在 $\alpha = 0.05$ 时, $\chi^2_{0.95}(6-2-1) = \chi^2_{0.95}(3) = 7.815$ 因而拒绝域为 $W = \{\chi^2 \geq 7.815\}$

由样本计算 χ^2 值的过程列于下表。

区间	n_i	β_i	np_i	$\frac{(n_i-n\beta_i)^2}{n\beta_i}$
$(-\infty, 200]$	10	0.045	9.0	0.111
$(200, 210]$	26	0.142	28.4	0.203
$(210, 220]$	56	0.281	56.2	0.001
$(220, 230]$	64	0.299	59.8	0.295
$(230, 240]$	30	0.171	34.2	0.516
$(240, \infty)$	14	0.062	12.4	0.206
合计	200			1.332

由此可知 $\chi^2 = 1.332 < 7.815$ 这表明样本落入接受域,可接受抗压强度服从正态分布的假定。

卡方检验应用情况

当题目中并不知道所给出的数据的分布，且又明确告诉我们需要对数据分布进行研究是，我们需要使用该方法来对数据进行检验，检验其是否满足我们需要的分布。

一元线性回归

一般形式

一元线性回归模型的数学形式为 $y=\beta_0+\beta_1 x+\varepsilon$

一般情况下, 如果获得 n组样本数据观测值

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$$

则上述模型可以等价的写为:

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i,(i=1,2,\cdots,n)$$

模型假设

(1)假定 ε 是不可观测的随机误差项, 它是一个随机变量, 满足

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases}$$

(2)假设n组数据是独立观测的, 因而 $y_i(i=1,2,\cdots,n)$ 与 $\varepsilon_i(i=1,2,\cdots,n)$ 是相互独立 的随机变量. 并且有

$$\begin{cases} E\left(\varepsilon_i\right)=0 \\ \mathrm{var}\left(\varepsilon_i\right)=\sigma^2 \end{cases}$$

(3)在实际问题中为了方便对参数进行区间估计和假设检验, 我们假定 ε 服从正态分布, 即

$$\varepsilon_i \sim N\left(0,\sigma^2\right)$$

参数估计

1）最小二乘估计（OLSE）

所谓最小二乘估计就是寻找参数 β_0, β_1 的估计值 $\hat{\beta}_0, \beta_1$ ，使得离差平方和最小，即就是

$$Q\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

由于 Q是关于β_0, β_1的非负二次函数, 因而极小值总是存在的. 对离差平方和 $Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$分别关于 β_0, β_1 求导即可得最小二乘估计.

求解即得$\hat{\beta}_0, \hat{\beta}_1$ 的最小二乘估计:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \beta_0 = \bar{y}-\hat{\beta}_1\bar{x} \end{cases}$$

最大似然估计(MLE)

由之前的模型假设条件有 $y_i \sim N\left(\beta_0+\beta_1 x_i, \sigma^2\right)$

相应的 $y_1, y_2, \cdots, y_n$ 似然函数为

$$L\left(\beta_0, \beta_1, \sigma^2\right)=\prod_{i=1}^{n} f_i\left(y_i\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2\right\}$$

对应的对数似然函数为

$$\ln(L)=-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_i\right)^2$$

求上式的最大值等价于求离差平方和的最小值, 从而等价于最小二乘估计.

显著性检验

对于一元线性回归方程 $E(y)=\beta_0+\beta_1 x$ 当 $\beta_1=0$ 时, 不管x如何变化, E(y)都不随x的变化做线性变化, 这是求得的回归方程就没有意义, 也称回归方程不显著. 反之, 若 $\beta_1\neq 0$ 时, E(y)都随 x的变化做线性变化, 称回归方程显著.

综上, 一元线性回归方程的显著性检验如下:

$$H_0:\beta_1=0 \quad \text{vs} \quad H_1:\beta_1\neq 0$$

拒绝原假设 H_0 即表示回归方程显著.

F检验

记

$$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$$ 表示总偏差平方和

$$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$ 表示残差平方和

$$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$ 表示回归平方和

由平方分解式可得: $SST=SSR+SSE$

构造 F统计量 $F = \dfrac{SSR/1}{SSE/(n-2)}$

$$SSR/\sigma^2 \sim \chi^2(1), SSE/\sigma^2 \sim \chi^2(n-2)$$

在正态假设条件下, $F \sim F(1, n-2)$ 其中当显著水平为 α 时, 拒绝域为 $W = \{F \geq F_\alpha(1, n-2)\}$

回归分析表：

方差来源	自由度	平方和	均方	F值	P值
回归	1	SSR	$\frac{SSR}{1}$	$\frac{\frac{SSR}{1}}{\frac{SSE}{n-2}}$	$P(F > F\text{ 值}) = P\text{ 值}$
残差	$n-2$	SSE	$\frac{SSE}{n-2}$		
总和	$n-1$	SST			

t检验

构造 t 统计量

$$l_{xx} = \sum (x_i - \bar{x})^2$$

$$t = \frac{\hat{\beta}_1 \sqrt{L_{xx}}}{\hat{\sigma}}$$

$$\hat{\sigma} = \sqrt{SSE/(\mathrm{n}-2)}$$

在假设条件下有 $t \sim t(n-2)$

当显著水平为 α 时，拒绝域为

$$W = \{|t| \geq t_{1-\alpha/2}(n-2)\}$$

相关系数检验

$$l_{xx}=\sum(x_i-\bar{x})^2$$

$$l_{yy}=\sum(y_i-\bar{y})^2$$

$$l_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})$$

记样本相关系数数为

$$r=\frac{L_{xy}}{\sqrt{L_{xx}}\sqrt{L_{yy}}}$$

检验的拒绝域为

$$W=\{|r|\geq r_{1-\alpha}(n-2)\}$$

其中 $r_\alpha(n-2)$ 查相关系数临界值表可得.

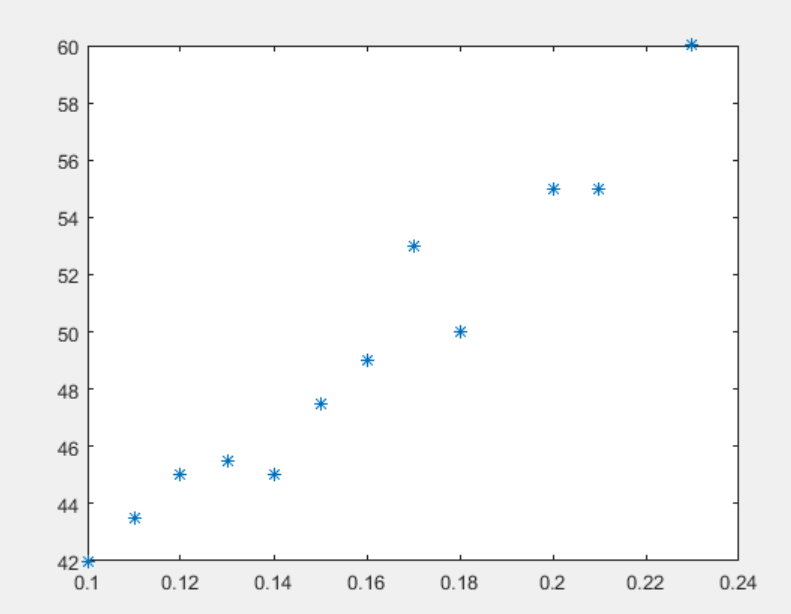
注：上面三个检验，在考察一元线性回归时是等价的，但是在多元线性回归方程场合，经推广的F检验仍然可以使用，但是另外两个检验就无法使用了。

例:

由专业知识知道, 合金强度 $Y\left(N/mm^2\right)$ 与合金中的碳含量 $X(\%)$ 有关.为了了解他们之间的关系, 从生产中收集了一批数据 $\left(x_i, y_i\right)(i=1,2,\cdots,n)$ 具体数据见下表:

序号	碳含量 X	强度 Y	序号	碳含量 X	强度 Y
1	0.10	42.0	7	0.16	49.0
2	0.11	43.5	8	0.17	53.0
3	0.12	45.0	9	0.18	50.0
4	0.13	45.5	10	0.20	55.0
5	0.14	45.0	11	0.21	55.0
6	0.15	47.5	12	0.23	60.0

首先画出上面数据的散点图：



从散点图可以看出这 12 个点基本在一条直线附近, 从而可以近似认为 Y与 X基本是线性的，因此可以用线性回归。在MATLAB中使用regress函数，对上面数据进行回归，并且的到相应的统计分析，程序如图所示：

```
A=xlsread('6.1.xls');
for i=1:12
    x(i)=A(i,1);
    y(i)=A(i,2);
end
plot(x,y,'*')
x=x';
x=[ones(12,1),x];
y=y';
[b,bint,r,rint,stats] = regress(y,x)
```

回归结果为:

```
b =

   28.4928
  130.8348
```

```
stats =

    0.9481  182.5546    0.0000    1.7410
```

因此: $\hat{y} = 28.4928 + 130.8348x$

右边统计数据分别表示相关系数，F值，p值,估计误差方差。我们可以看到相关系数已经到达了0.9以上，因此回归效果很好，且p值较小，说明自变量对因变量的作用是显著的。

多元线性回归

一般形式

多元线性回归模型的一般数学形式为:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

对于一个实际问题, 设 $\left(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i\right)(i = 1, 2, \cdots, n)$ 是我们获得的n次独立观测 值, 则有:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

模型假设

(1)rank(X)=p+1<n, 表示矩阵 X自变量列之间线性无关，并且样本量的个数应大于 解释变量的个数.

(2)随机误差项 ε_i 满足:

$$\begin{cases} E(\varepsilon_i)=0, i=1,2,\cdots,n \\ \operatorname{cov}(\varepsilon_i,\varepsilon_j)=\begin{cases} \sigma^2, i=j \\ 0, i\neq j \end{cases} \quad i,j=1,2,\cdots,n \end{cases}$$

这个条件称为高斯--马尔科夫条件

(3)正态分布假定条件:

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n \text{ 相互独立} \end{cases}$$

同一元回归分析的参数估计原理一样，多元线性回归分析也可以采用最小二乘估计和最大 似然估计，这就不详细说明了．当对离差平方和求偏导后，得到的矩阵形式的正规方程组为

$$X'(Y - X\beta) = 0$$

当 $(X'X)^{-1}$ 存在时，即可得回归参数的最小二乘估计为

$$\boldsymbol{\beta} = (X'X)^{-1} X'Y$$

误差项方差 σ^2 的无偏估计为 $\hat{\sigma}^2 = \frac{1}{n-p-1}(ee')$

其中：$e' = (e_1, e_2, \cdots, e_n) = (y_1 - y_1, y_2 - y_2, \cdots, y_n - y_n)$

此外：

$$E(\boldsymbol{\beta}) = E\left(\left(X'X\right)^{-1}X'Y\right) = \left(X'X\right)^{-1}X'E(Y) = \left(X'X\right)^{-1}X'X\beta = \beta$$

$$D(\hat{\beta}) = \operatorname{cov}(\boldsymbol{\beta}, \boldsymbol{\beta}) = \operatorname{cov}\left(\left(X'X\right)^{-1}X'Y, \left(X'X\right)^{-1}X'Y\right)$$

$$= \left(X'X\right)^{-1}X'\sigma^2 I_n\left(\left(X'X\right)^{-1}X'\right)' = \sigma^2\left(X'X\right)^{-1}$$

显著性检验

F检验

设原假设为 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

构造 F统计量 $F = \dfrac{SSR / p}{SSE / (n - p - 1)}$

则有 $F \sim F(p, n - p - 1)$

方差来源	自由度	平方和	均方	F值	P 值
回归	p	SSR	$\frac{SSR}{p}$	$\frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}}$	$P(F > F\text{ 值}) = P\text{ 值}$
残差	$n-p-1$	SSE	$\frac{SSE}{n-p-1}$		
总和	$n-1$	SST			

t检验

为了检验某个自变量 x_j 对 y的作用显不显著，我们可设原假设为: $H_{0j}:\beta_j=0, j=1,2\cdots,p$

构造 t统计量 $t_j=\dfrac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$ $\hat{\sigma}=\sqrt{\dfrac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}$

其中 $(c_{ij})=(X'X)^{-1}, i, j=0,1,2,\cdots,p$

则有 $t_j\sim t(n-p-1)$

当显著水平为 α 时, 拒绝域为

$$W=\left\{\left|t_j\right|\geq t_{1-\alpha/2}\right\}$$

注：在前面有说在考察一元线性回归时t检验和F检验是等价的，但是在此处，F检验可以检验所有的自变量对y的作用是否都显著，而t检验只能检验某一个自变量是否对y的作用显著。

拟合优度

在多元线性回归中, 我们定义样本决定系数为

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

样本决定系数R^2取值在 [0,1] 区间内, R^2 越接近 1, 表明回归拟合效果越好, 越接近 0 , 表示回归拟合效果越差. 与 F检验相比 , R^2 可以更清楚直观的反应拟合效果，但是并不能作为严格显著性检验.

同时我们可以对 R^2 进行自由度的调整:

$$\bar{R} = \frac{SSR / p}{SST / (n-1)}$$

例:

根据经验表明, 一般在人的身高相等的情况下，血压的收缩压 Y与体 重 $X_1(kg)$、年龄 X_2有关. 现收集了13个男子的数据, 如下表所示:

序号	X_1	X_2	Y
1	76.0	50	120
2	91.5	20	141
3	85.5	20	124
4	82.5	30	126
5	79.0	30	117
6	80.5	50	125
7	74.5	60	123
8	79.0	50	125
9	85.0	40	132
10	76.5	55	123
11	82.0	40	132
12	95.0	40	155
13	92.5	20	147

对其进行回归，对应的程序如图所示

```
A=xlsread('6.2.xls');
for i=1:13
    x(i,1)=A(i,1);
    x(i,2)=A(i,2);
    y(i)=A(i,3);
end
x=[ones(13,1),x];
y=y';
[b,bint,r,rint,stats] = regress(y,x)
```

得到的结果为：

```
b =

  -62.9634
    2.1366
    0.4002

stats =

    0.9461   87.8404    0.0000    8.1430
```

回归方程为：

$$\hat{Y} = -62.9634 + 2.1366X_1 + 0.4002X_2$$

统计结果显示，回归效果较好，且自变量对因变量是显著的。

使用回归方法处理数据的情况

当我们的数据之间有明显的线性关系，且题目中想要让我们对数据未来的趋势进行预测时，我么就可以使用回归的方法来对数据进行处理。在得到回归的方程后并不意味着数据处理的结束，我们同样需要对方程进行显著性检验，可信度检验等。

方差分析

单因素方差分析

一、数据描述

设因素为 A，共有 r 个水平：

$$A_1, A_2, \ldots, A_r,$$

在水平 A_i 下做 n_i 次重复试验，n_i 可以不全相等.当

$$n_1 = n_2 = \ldots = n_r$$

时，实验称为等重复数单因素试验.我们可以得到一个单因素方差分析的数据表：

A_1	x_{11}	x_{12}	$\cdots$	x_{1n_1}
A_2	x_{21}	x_{22}	$\cdots$	x_{2n_2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rn_r}

二、数学模型

将水平 A_i 下的实验结果

$$x_{i1}, x_{i2}, \ldots, x_{in_i}$$

看做来自第 i 个正态总体

$$X_i \sim N(\mu_i, \sigma^2)$$

的样本观测值，其中 μ_i, σ^2 均未知，且每个总体 X_i 相互独立，考虑线性统计模型

$$\begin{cases} x_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{ 且相互独立} \end{cases} \tag{6.9}$$

其中，μ_i 是第 i 个总体的均值，ε_{ij} 是相应的实验误差.

我们的目标是要检验各处理或各水平对实验有无影响并估计他们的影响程度，比较因素 A 的 r 个水平的差异归结为比较这 r 个总体的均值，即检验假设

$$H_0: \ \mu_1 = \mu_2 = \cdots = \mu_r \ \ vs \ \ H_1: \ \mu_1, \mu_2, \cdots \mu_r \text{ 不全相等},$$

记

$$\mu = \frac{1}{n}\sum_{i=1}^{r} n_i\mu_i, \quad n = \sum_{i=1}^{r} n_i, \quad \alpha_i = \mu_i - \mu,$$

这里 μ 表示总平均，α_i 称为水平 A_i 对指标的效应，不难验证

$$\sum_{i=1}^{r} n_i\alpha_i = 0.$$

从而，上面的检验假设可以等价写成

$$H_0:\ \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \ \ vs \ \ H_1:\ \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 至少一个不为 } 0,$$

统计模型等价为

$$\begin{cases} x_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \ \ i = 1, 2, \cdots, r, \ \ j = 1, 2, \cdots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{ 且相互独立} \\ \sum_{i=1}^{r} n_i\alpha_i = 0 \end{cases} \tag{6.10}$$

此模型称为单因素方差分析的数学模型，它是一种线性模型.

三、平方和与自由度分解

1、偏差平方和

在单因素试验中，涉及三种偏差平方和，即总离差平方和（或总变差）、效应平方和（或组间平方和）和误差平方和（或组内平方和）.

总离差平方和

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij})$$

它是所有数据与总平均值差的平方和，描绘的是所有观测数据的离散程度.

误差平方和

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_{i\cdot}})^2, \quad \bar{x_{i\cdot}} = \frac{1}{n_i}\sum_{j=1}^{n_i}x_{ij},$$

误差平方和描述的是随机误差的影响.

效应平方和

$$S_A = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{x_{i\cdot}} - \bar{x})^2 = \sum_{i=1}^{r} n_i(\bar{x_{i\cdot}} - \bar{x})^2,$$

反映的是 r 个总体之间的差异.

$$S_T = S_E + S_A$$

由统计量的无偏估计相关内容，容易得到，在原假设 H_0 下，

$$\frac{S_A}{r-1}, \quad \frac{S_E}{n-r}$$

均是 σ^2 的无偏估计，即 V_A，V_E 均是 σ^2 的无偏估计.

且 $S_A/\sigma^2 \sim \chi^2(r-1)$，$S_E/\sigma^2 \sim \chi^2(n-r)$. S_A，S_E 相互独立，由 F 分布的定义，可构造 H_0 的检验统计量

$$F = \frac{S_A/(r-1)}{S_E/(n-r)} = \frac{V_A}{V_E} \sim F(r-1, n-r)$$

在原假设 H_0 成立的前提下，比值 F 的分子、分母都是总体方差 σ^2 的无偏估计量.故统计量 F 应当“很接近于 1”，如果因素 A 均方差 V_A 比误差均方差 V_E, 大的很多，即 F 值比“1”大的很多，则与原假设 H_0 相矛盾，这时有理由拒绝原假设，认为因素 A 的不同条件形成均值不完全相等的 r 个正态总体.

对于给定的显著性水平 α，用 $F_\alpha(r-1,n-r)$ 表示 F 分布上的 α 分位点.若 $F > F_\alpha(r-1,n-r)$，出现小概率事件，有理由拒绝原假设 H_0，认为单因素 A 的 r 个水平有显著差异.

若考虑 p 值：

$$p = P(F(r-1,n-r) > F),$$

则 p 值小于 α 等价于 $F > F_\alpha(r-1,n-r)$, 同样表示在显著性水平 α 下小概率事件发生了，应拒绝原假设；当 p 值大于 α 时，无法拒绝原假设，所以应该接受原假设.

从而，我们得到如下的单因素方差分析表：

方差来源	自由度	平方和	均方	F 值	P 值
因素 A	$r-1$	S_A	$V_A = \frac{S_A}{r-1}$	$F = \frac{V_A}{V_E}$	P值
误差	$n-r$	S_E	$V_E = \frac{S_E}{n-r}$		
总和	$n-1$	S_T			

如果 $F \geq F_{1-\alpha}(f_A, f_e)$ 则认为因子 A显著;若 $F < F_{1-\alpha}(f_A, f_e)$，则说明因子A不显著.

双因素方差分析

一、双因素无重复试验方差分析

1、双因素无重复试验 在双因素试验中，只有两个变动因素，记为 A 和 B.设因素 A 有 r 个 水平：$A_1, A_2, \ldots, A_r,$

因素 B 有 s 个水平：

$$B_1, B_2, \ldots, B_s,$$

则因素 A 与 B 之间有 rs 种不同水平搭配方式.对所有水平搭配方式均进行试验，称双因素全面实验，在每种水平下均进行一次实验，称双因素全面无重复试验，简称双因素无重复试验.

双因素试验比单因素试验要复杂，因为两个因素可能存在交互作用.但双因素无重复试验，即便存在交互作用的影响，也不能够对其分析，因为每一种实验条件下只有一个实验结果，这使得交互作用和实验误差混杂在一起，无法分解开来.故对双因素无重复试验来说，交互作用只好和误差和在一起当作误差来考虑.

	B_1	B_2	$\cdots$	B_s
A_1	x_{11}	x_{12}	$\cdots$	x_{1s}
A_2	x_{21}	x_{22}	$\cdots$	x_{2s}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rs}

2、方差分析

假定

$$x_{ij} \sim N(\mu_{ij}, \sigma^2),\ \ i = 1, 2, \cdots, r,\ \ j = 1, 2, \cdots, s,$$

且各 x_{ij} 相互独立.因为在双因素无重复试验中，可认为不考虑交互作用，故可建立如下数学模型：

$$\begin{cases} x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},\ \ i = 1, 2, \cdots, r,\ \ j = 1, 2, \cdots, s, \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{ 且相互独立} \\ \sum_{i=1}^{r} \alpha_i = 0,\ \ \sum_{j=1}^{s} \beta_j = 0 \end{cases} \tag{6.11}$$

其中

$$\mu = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} \mu_{ij}$$

为总平均，α_i 为因素 A 的第 i 个水平的效应，β_j 为因素 B 的第 j 个水平的效应.

我们的目标是分析因素 A 和 B 对试验指标影响的大小，在给出的显著性水平 α 下，提出假设：

$$H_{01}:\ \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0,$$

即因素 A 对试验指标影响不显著，

$$H_{02}:\ \beta_1 = \beta_2 = \cdots = \beta_s = 0.$$

即因素 B 对试验指标影响不显著.

双因素方差分析与单因素方差分析统计原理基本相同，也是基于平方和分解公式：

$$S_T = S_E + S_A + S_B.$$

其中，

$$S_T = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \bar{x})^2,\ \ \bar{x} = \frac{1}{rs}\sum_{i=1}^{r}\sum_{j=1}^{s}x_{ij},$$

$$S_A = s\sum_{i=1}^{r}(\bar{x_{i.}} - \bar{x})^2,\ \ \bar{x_{i.}} = \frac{1}{s}\sum_{j=1}^{s}x_{ij},\ i = 1, 2, \cdots, r,$$

$$S_B = r\sum_{j=1}^{s}(\bar{x_{.j}} - \bar{x})^2,\ \ \bar{x_{.j}} = \frac{1}{r}\sum_{i=1}^{r}x_{ij},\ j = 1, 2, \cdots, s,$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_{ij} - \bar{x_{i.}} - \bar{x_{.j}} + \bar{x})^2,$$

S_T 为总离差平方和，S_A 为因素 A 效应平方和，S_B 为因素 B 效应平方和，S_E 为误差平方和，同单因素方差分析一样，我们可以得到：

$$S_A/\sigma^2 \sim \chi^2(r-1),$$

$$S_B/\sigma^2 \sim \chi^2(n-r),$$

$$S_E/\sigma^2 \sim \chi^2((r-1)(s-1)).$$

并且，S_A 与 S_E 相互独立，S_B 与 S_E 相互独立.

另外还有，当 H_{01} 成立时，

$$F_A = \frac{S_A/(r-1)}{S_E/[(r-1)(s-1)]} \sim F(r-1,(r-1)(s-1)),$$

当 H_{02} 成立时

$$F_B = \frac{S_B/(s-1)}{S_E/[(r-1)(s-1)]} \sim F(s-1,(r-1)(s-1)).$$

以 F_A、F_B 分别作为 H_{01}、H_{02} 的检验统计量，将结果列成方差分析表：

方差来源	自由度	平方和	均方	F 值	P 值
因素 A	$r-1$	S_A	$V_A = \frac{S_A}{r-1}$	$\frac{V_A}{V_E}$	P_A
因素 B	$s-1$	S_B	$V_B = \frac{S_B}{s-1}$	$\frac{V_B}{V_E}$	P_B
误差	$(r-1)(s-1)$	S_E	$V_E = \frac{S_E}{(r-1)(s-1)}$		
总和	$rs-1$	S_T			

二、双因素等重复试验方差分析

1、双因素等重复试验

设因素 A 有 r 个水平：$A_1, A_2, \ldots, A_r$, 因素 B 有 s 个水平：$B_1, B_2, \ldots, B_s$, 在每种水平组合 (A_i, B_j) 下重复试验 t 次.记第 k 次观测值为 x_{ijk},将观测数据列表得

	B_1	B_2	$\cdots$	B_s
A_1	$x_{111}x_{112}\cdots x_{11t}$	$x_{121}x_{122}\cdots x_{12t}$	$\cdots$	$x_{1s1}x_{1s2}\cdots x_{1st}$
A_2	$x_{211}x_{212}\cdots x_{21t}$	$x_{221}x_{222}\cdots x_{22t}$	$\cdots$	$x_{2s1}x_{2s2}\cdots x_{2st}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	$x_{r11}x_{r12}\cdots x_{r1t}$	$x_{r21}x_{r22}\cdots x_{r2t}$	$\cdots$	$x_{rs1}x_{rs2}\cdots x_{rst}$

方差分析 假定

$$x_{ijk} \sim N(\mu_{ij}, \sigma^2), \quad i = 1, 2, \cdots, r, \quad j = 1, 2, \cdots, s, \quad k = 1, 2, \cdots, t,$$

且各 x_{ijk} 相互独立.建立如下数学模型：

$$\begin{cases} x_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} \sim N(0, \sigma^2), \text{ 且相互独立} \\ i = 1, 2, \cdots, r, \ \ j = 1, 2, \cdots, s, \ \ k = 1, 2, \cdots, t \end{cases} \tag{6.12}$$

其中 α_i 为因素 A 的第 i 个水平的效应，β_j 为因素 B 的第 j 个水平的效应.δ_{ij} 表示 A_i 和 B_j 的交互效应，从而有

$$\mu = \frac{1}{rs} \sum_{i=1}^{r} \sum_{j=1}^{s} \mu_{ij},$$

$$\sum_{i=1}^{r} \alpha_i = 0, \ \ \sum_{j=1}^{s} \beta_j = 0, \ \ \sum_{i=1}^{r} \delta_{ij} = \sum_{j=1}^{s} \delta_{ij} = 0$$

判断因素 A、B 及交互作用的影响是否显著等价于检验下列假设

$$H_{01}:\ \alpha_1=\alpha_2=\cdots=\alpha_r=0,$$

$$H_{02}:\ \beta_1=\beta_2=\cdots=\beta_s=0,$$

$$H_{03}:\ \delta_{ij}=0,\ \ i=1,2,\cdots,r,\ \ j=1,2,\cdots,s.$$

与前面所讲的方法类似，有平方和分解式

$$S_T=S_E+S_A+S_B+S_{A\times B},$$

其中：

$$S_T=\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk}-\bar{x})^2,\ \ \bar{x}=\frac{1}{rst}\sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}x_{ijk},$$

$$S_A=st\sum_{i=1}^{r}(\bar{x_{i..}}-\bar{x})^2,\ \ \bar{x_{i..}}=\frac{1}{st}\sum_{j=1}^{s}\sum_{k=1}^{t}x_{ijk},\ i=1,2,\cdots,r,$$

$$S_B=rt\sum_{j=1}^{s}(\bar{x_{.j.}}-\bar{x})^2,\ \ \bar{x_{.j.}}=\frac{1}{rt}\sum_{i=1}^{r}\sum_{k=1}^{t}x_{ijk},\ j=1,2,\cdots,s,$$

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(x_{ijk} - \bar{x_{ij.}})^2,\ x_{ij.} = \frac{1}{t}\sum_{k=1}^{t}x_{ijk},\ \ i = 1, 2, \cdots, r,\ j = 1, 2, \cdots, s.$$

$$S_{A\times B} = t\sum_{i=1}^{r}\sum_{j=1}^{s}(\bar{x_{ij.}} - \bar{x_{i..}} - \bar{x_{.j.}} + \bar{x})^2.$$

S_T 为总离差平方和，S_A 为因素 A 效应平方和，S_B 为因素 B 效应平方和，S_E 为误差平方和，$S_{A\times B}$ 为交互效应平方和.同样可以证明：

当 H_{01} 成立时

$$F_A = \frac{S_A/(r-1)}{S_E/[rs(t-1)]} \sim F(r-1, rs(t-1)),$$

当 H_{02} 成立时

$$F_B = \frac{S_B/(s-1)}{S_E/[rs(t-1)]} \sim F(s-1, rs(t-1)),$$

当 H_{03} 成立时

$$F_{A\times B} = \frac{S_{A\times B}/[(r-1)(s-1)]}{S_E/[rs(t-1)]} \sim F((r-1)(s-1), rs(t-1)).$$

分别以 $F_A, F_B, F_{A\times B}$ 作为 H_{01}, H_{02}, H_{03} 的检验统计量，将检验结果列成方差分析表：

方差来源	自由度	平方和	均方	F 值	P 值
因素 A	$r-1$	S_A	$V_A = \frac{S_A}{r-1}$	$\frac{V_A}{V_E}$	P_A
因素 B	$s-1$	S_B	$V_B = \frac{S_B}{s-1}$	$\frac{V_B}{V_E}$	P_B
交互效应 A*B	$(r-1)(s-1)$	S_{A*B}	$V_{A*B} = \frac{S_{A*B}}{(r-1)(s-1)}$	$F_{A*B} = \frac{V_{A*B}}{V_E}$	P_{A*B}
误差	$rs(t-1)$	S_E	$V_E = \frac{S_E}{rs(t-1)}$		
总和	$rs-1$	S_T			

例题

考察五名工人劳动生产率是否相同, 记录每人四天的产量及平均值, 推断他们的生产率有无显著差异。

	A_1	A_2	A_3	A_4	A_5
1	256	254	250	248	236
2	242	330	277	280	252
3	280	290	230	305	220
4	298	295	302	289	252

MATLAB中有关于单因子方差分析的函数，我们在进行单因子方差分析时，可以调用anova1函数，调用方法如下

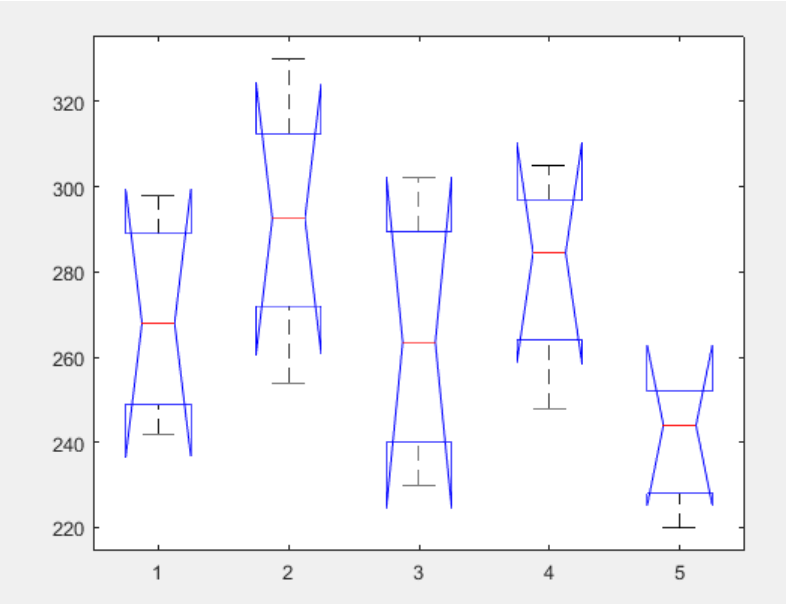
```
x=xlsread('6.3.xls');
p=anova1(x)
```

得到的方差分析表为：

Source	SS	df	MS	F	Prob>F
Columns	6125.7	4	1531.43	2.26	0.1109
Error	10156.5	15	677.1		
Total	16282.2	19			

可以看到 $p = 0.1109 > \alpha = 0.05$ 故接受 H_0 ，即五名工人的生产率没有显著差异

相应的盒装图为



方差分析使用情况

当我们拿到的数据在不同因素下有观测值，而题目要求我们判断因素是否对数据有影响时，需要用到单因子或多因子方差分析。

谢谢大家！