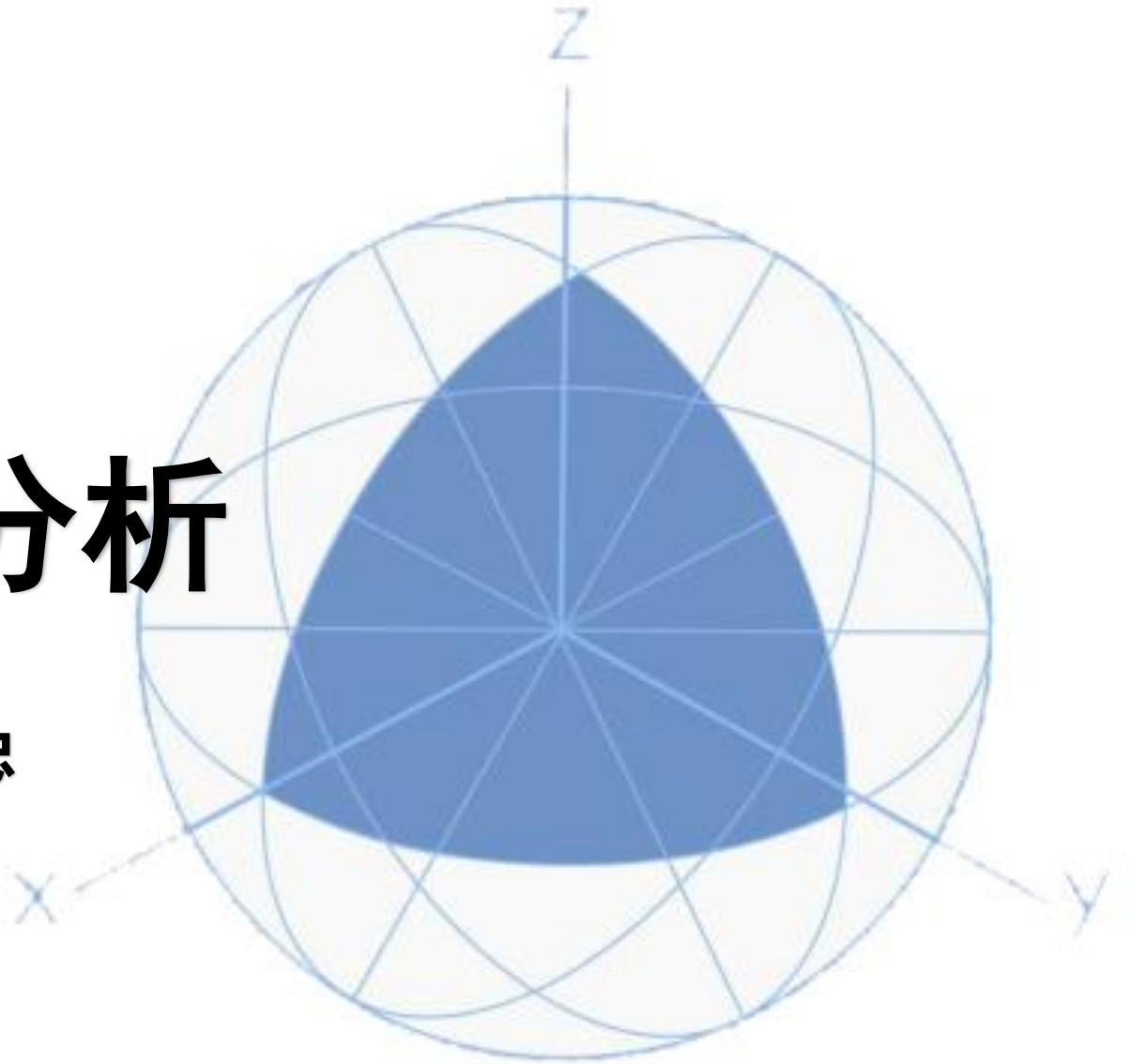
回归分析

谭　忠



厦门大学
XIAMEN UNIVERSITY

目录
CONTENTS

厦门大学
XIAMEN UNIVERSITY

Part 1

源头问题与当今应用

6.1源头问题与当今应用

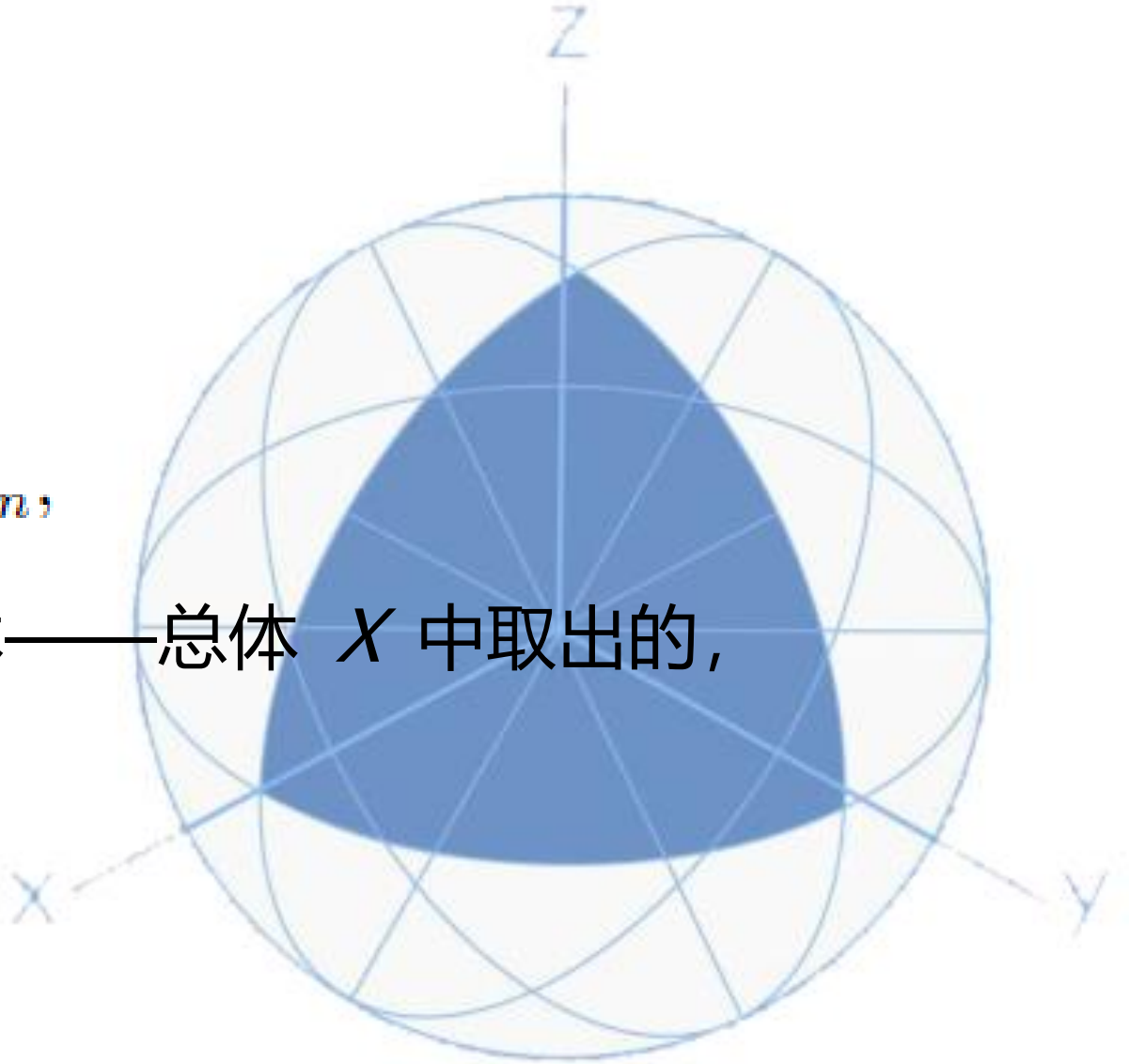
6.1.1数理统计基础知识

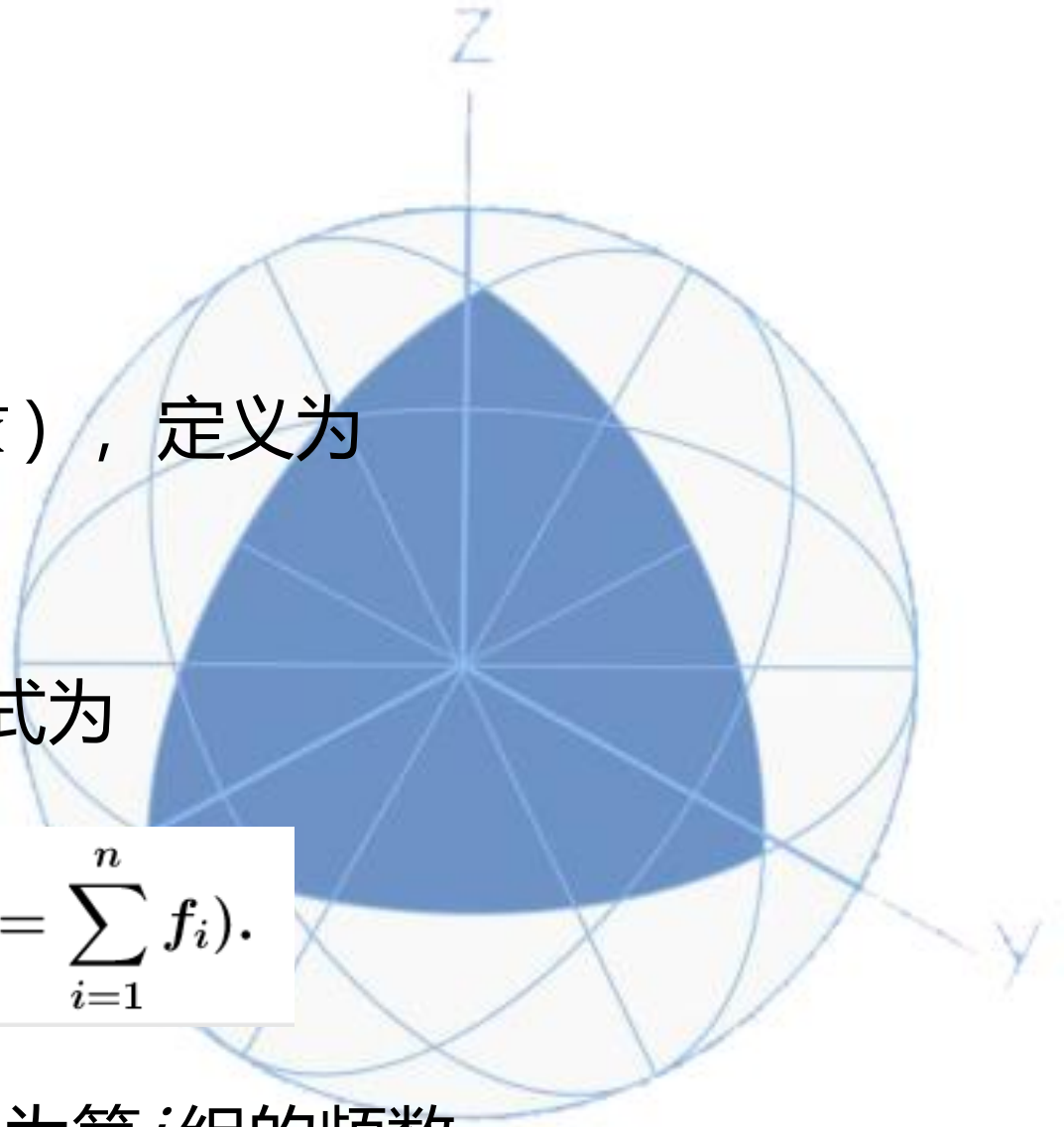
统计分析分为统计描述和统计推断两个部分. 统计描述是通过绘制统计图、编制统计表、计算统计量等方法来表述数据分布特征. 它是数据分析的基本步骤，也是进行统计推断的基础.

已知一组试验（或观测）数据为

$$x_1, x_2, \ldots, x_n,$$

它们是从所要研究的对象的全体——总体 X 中取出的，这n 个观测值就构成一个样本.





(1)均值

均值（mean）是数据的平均数（记为 $\bar{x}$），定义为

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

在分组样本场合，样本的均值的近似公式为

$$\bar{x} = \frac{x_1 f_1 + \ldots + x_k f_k}{n} \quad (n = \sum_{i=1}^{n} f_i).$$

其中 k 为组数，x_i 为第 i 组的组中值，f_i 为第 i 组的频数.

(2)方差

设*x*1,*x*2,...,*xn* 为取自某总体的样本，则它关于样本均值 $\bar{x}$ 的平均偏差平方和

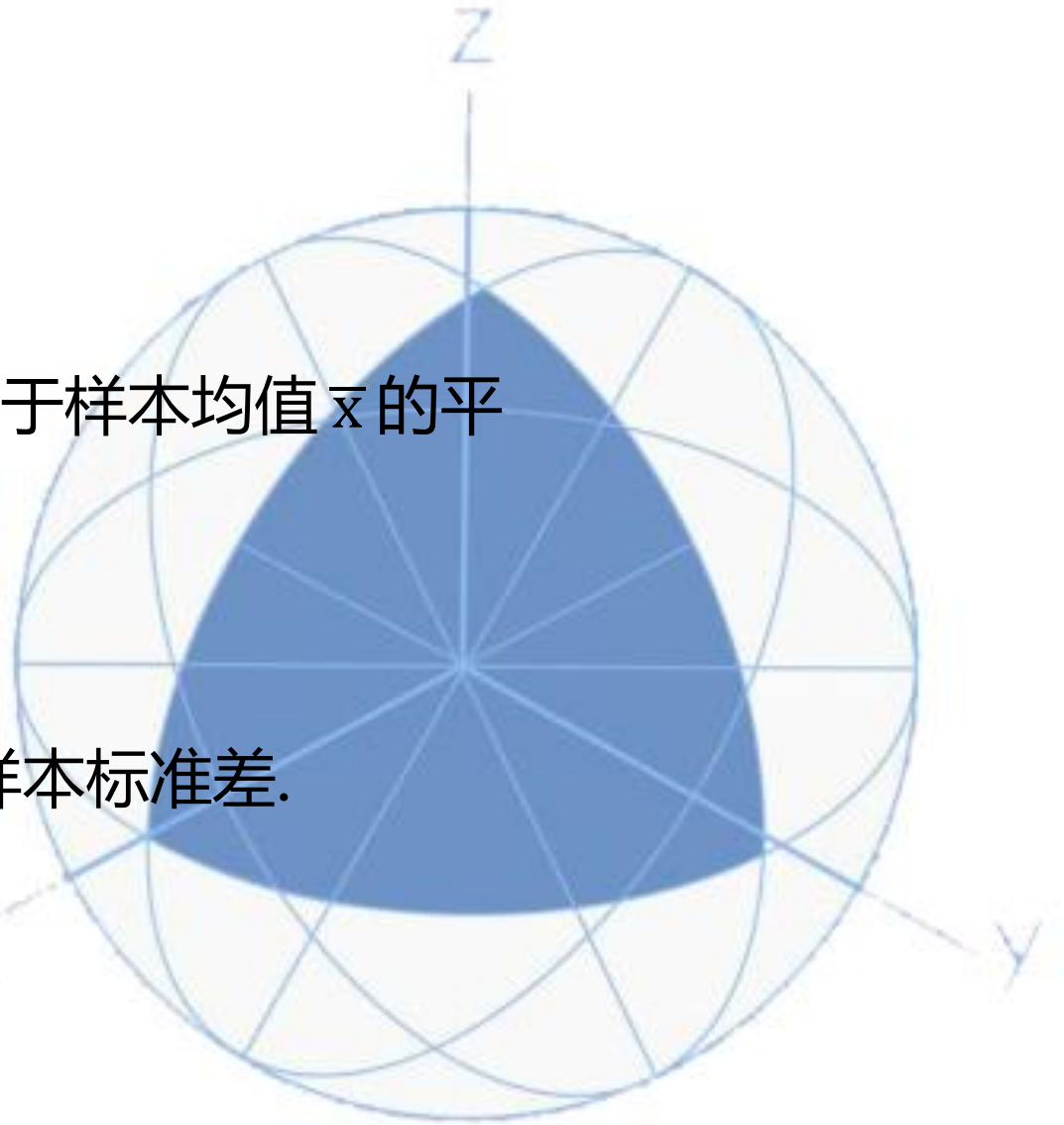
$$s^{*2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

称为样本方差. 其算术根 $s^* = \sqrt{s^{*2}}$ 称为样本标准差.

在*n* 不大时，常用

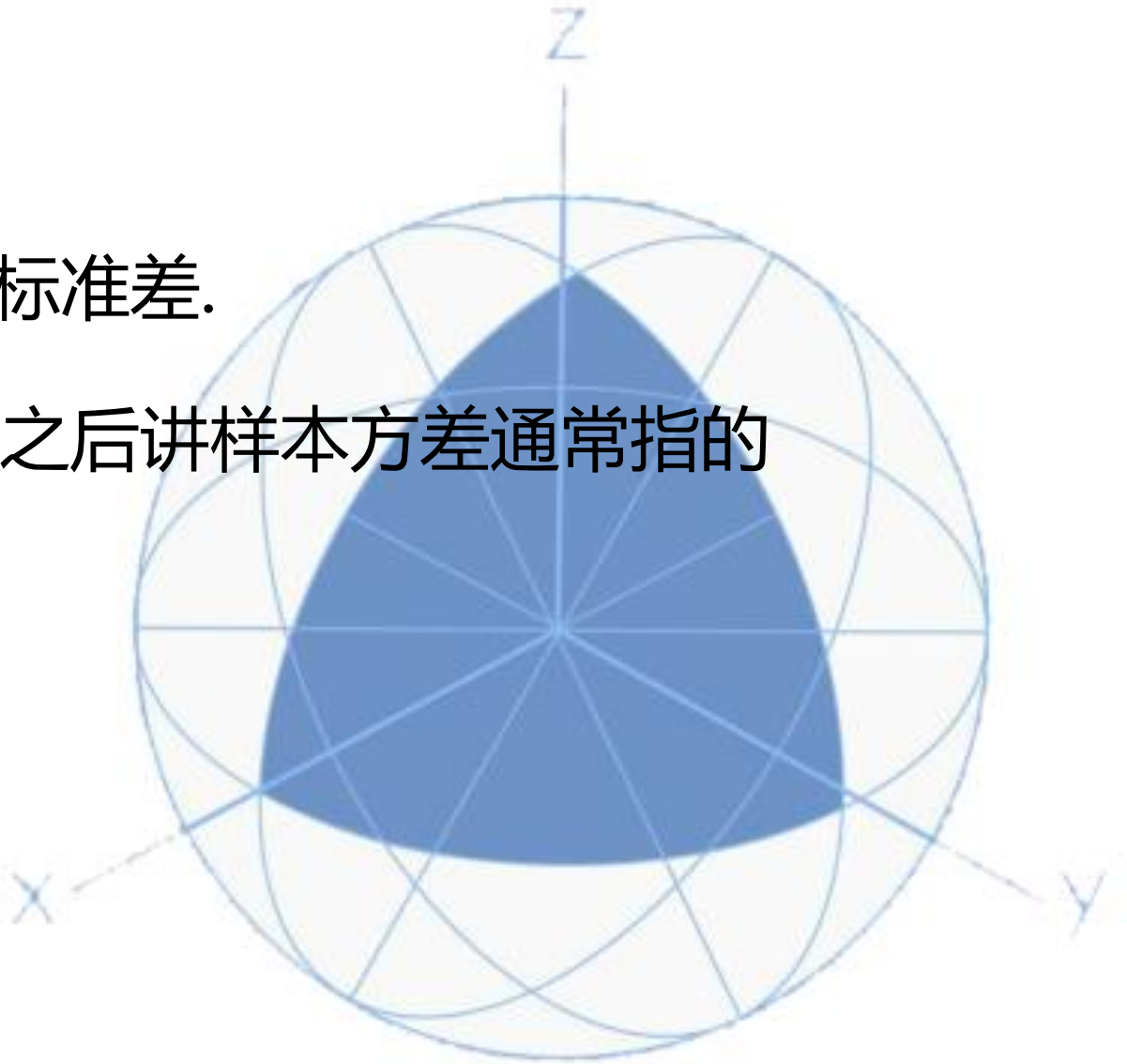
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

作为样本方差（也称为无偏方差）.



其算术根 $s=\sqrt{s^2}$ 也称为样本标准差.

在实际中，$s2$ 比s^*2更常用，在之后讲样本方差通常指的是$s2$.



(3)变异系数

变异系数又称“标准差率”．标准差与平均数的比值称为变异系数，记为C.V. 其计算公式为

$$CV = \frac{s}{\bar{x}} \times 100\%,$$

它是一个无量纲的量，用百分数表示. 变异系数越小，变异(偏离)程度越小；反之，变异系数越大，变异(偏离)程度越大.

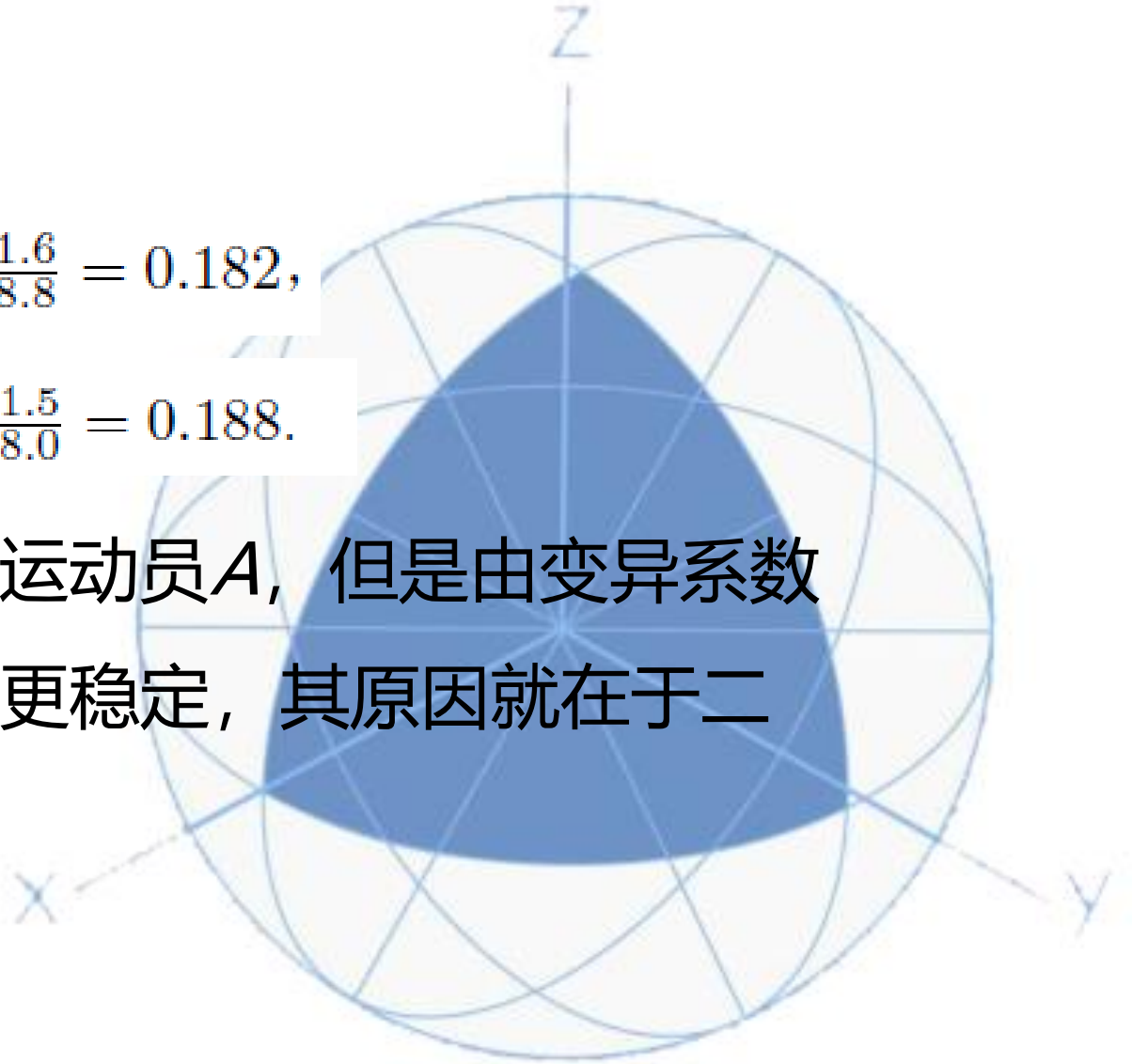
例 两个射击运动员参加射击比赛，运动员A 的数据统计是平均环数为8.8 环，标准差为1.6 环，运动员B 的平均环数为8.0 环，标准差为1.5 环，试比较两名运动员比赛的稳定程度.

此例中观测值虽然都是射击的环数，单位相同，但是它们的平均数不同，只能用变异系数来刻画稳定大小.

运动员A的变异系数为 $C.V = \frac{1.6}{8.8} = 0.182,$

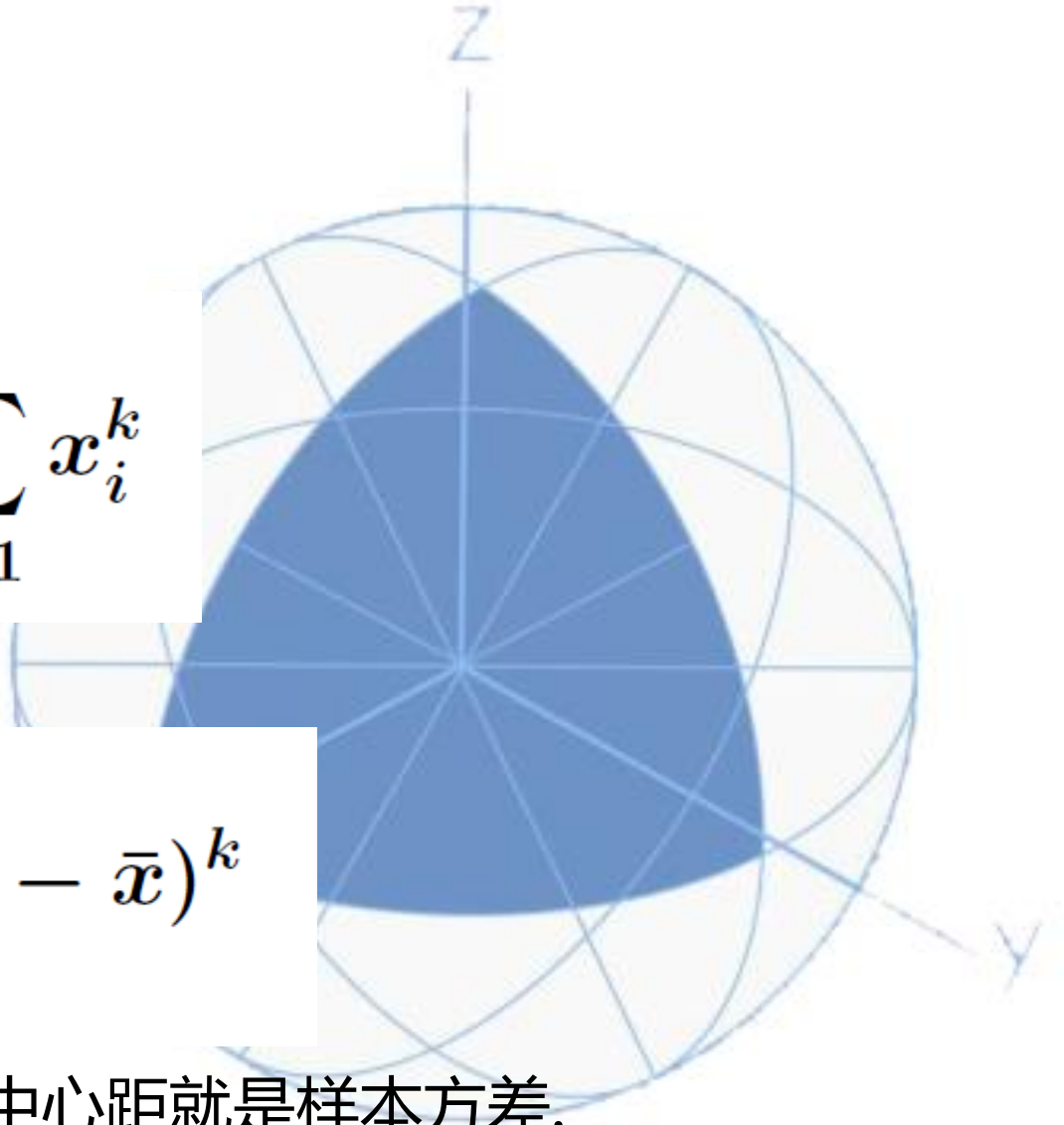
运动员B的变异系数为 $C.V = \frac{1.5}{8.0} = 0.188.$

可以看出运动员B的标准差小于运动员A，但是由变异系数反应出的稳定程度而言，A比B更稳定，其原因就在于二者的均值不同.



（4）样本的k阶原点矩

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$



（5）样本的k阶中心矩

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

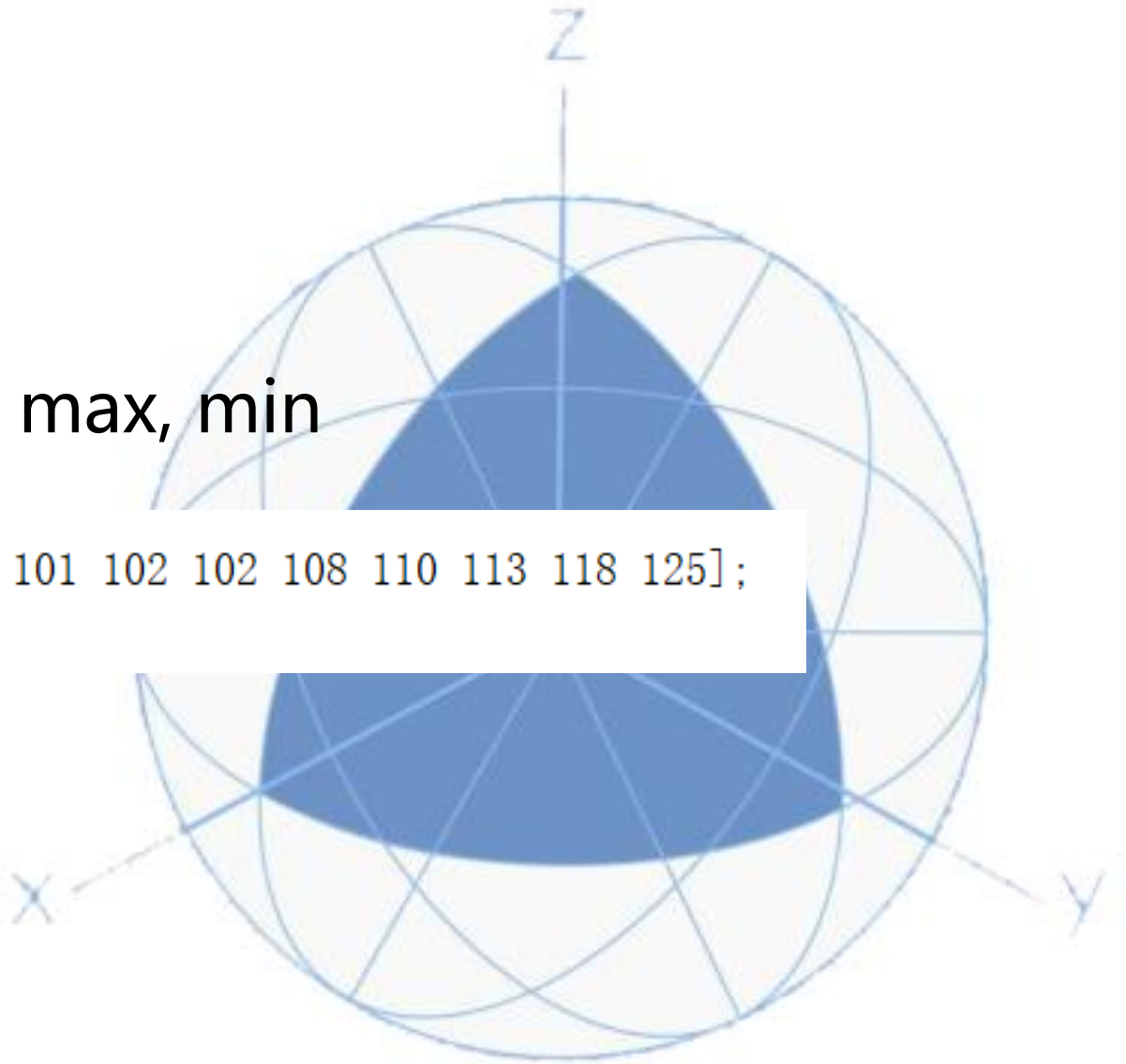
样本一阶原点矩就是样本均值，样本二阶中心距就是样本方差.

补：Matlab中统计量的计算

sum， mean, median, var, std, max, min

```
>> x=[79 84 84 88 92 93 94 97 98 99 100 101 101 102 102 108 110 113 118 125];
mean(x),var(x),std(x),max(x),min(x)
```

```
ans =

   99.4000


ans =

  133.9368


ans =

   11.5731


ans =

   125


ans =

    79
```



（6）统计中几个重要的概率分布

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

卡方分布 $\chi^2(n)$ 若X1,X2,..., Xn 为标准正态分布，那么$\sum_{i=1}^{n} x_i^2$

服从卡方分布 $\chi^2(n)$

***t*分布** $t(n)$ 若X服从标准正态分布N(0,1)， Y服从卡方分布 $\chi^2(n)$，那么

$\frac{X}{\sqrt{Y/\mathrm{n}}}$ 服从*t*分布 $t(n)$.

***F*分布** $F(m, n)$ 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 $F(m,n)$

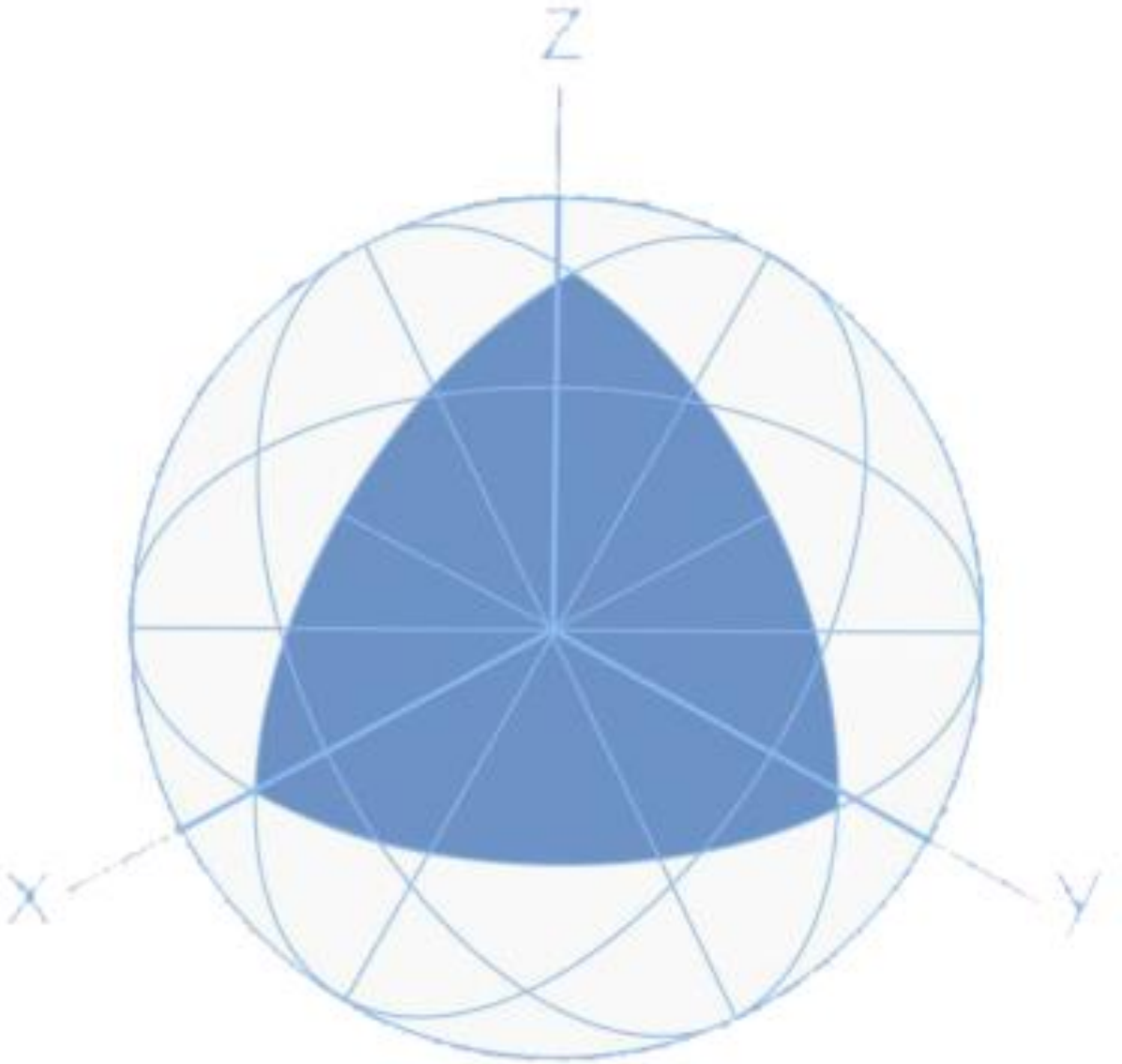
（7）抽样分布定理

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

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

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

6.1.2 参数估计

参数估计就是根据从总体中抽取的样本对估计总体分布中未知参数的方法.

当我们估计总体的某一个参数时，涉及到的最主要的两个问题就是：

（1）如何给出估计，即估计的方法问题

（2）如何对不同的估计进行评价，即估计的好坏判断标准.

一、点估计

点估计的两种常用方法

矩估计

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

极大似然估计

$$\widehat{\theta} : L(\theta) = \sup_{\theta\in\Theta} \prod_{i=1}^{n} f(x_i;\theta)$$

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 $E\widehat{\theta} = \theta$

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

一致性

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

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

二、区间估计

由于点估计值只是估计量的一个近似值，因而点估计本身既没有反映出这种近似值的精度，即指出用估计值去估计的误差范围有多大，而且也没有指出这个误差范围以多大的概率包括未知参数，这些问题正是区间估计要讨论的问题. 区间估计解决了这二个问题，它给出了估计的可信程度，是一种重要的统计推断形式.

当$P(\widehat{\theta}_1 < \theta < \widehat{\theta}_2) = 1-\alpha$ 时，称 $[\widehat{\theta}_1, \widehat{\theta}_2]$ 为参数 θ 置信水平为$1-\alpha$ 置信区间. 当置信区间越小，估计的精度越高；置信水平越大，估计的可信程度越高.

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

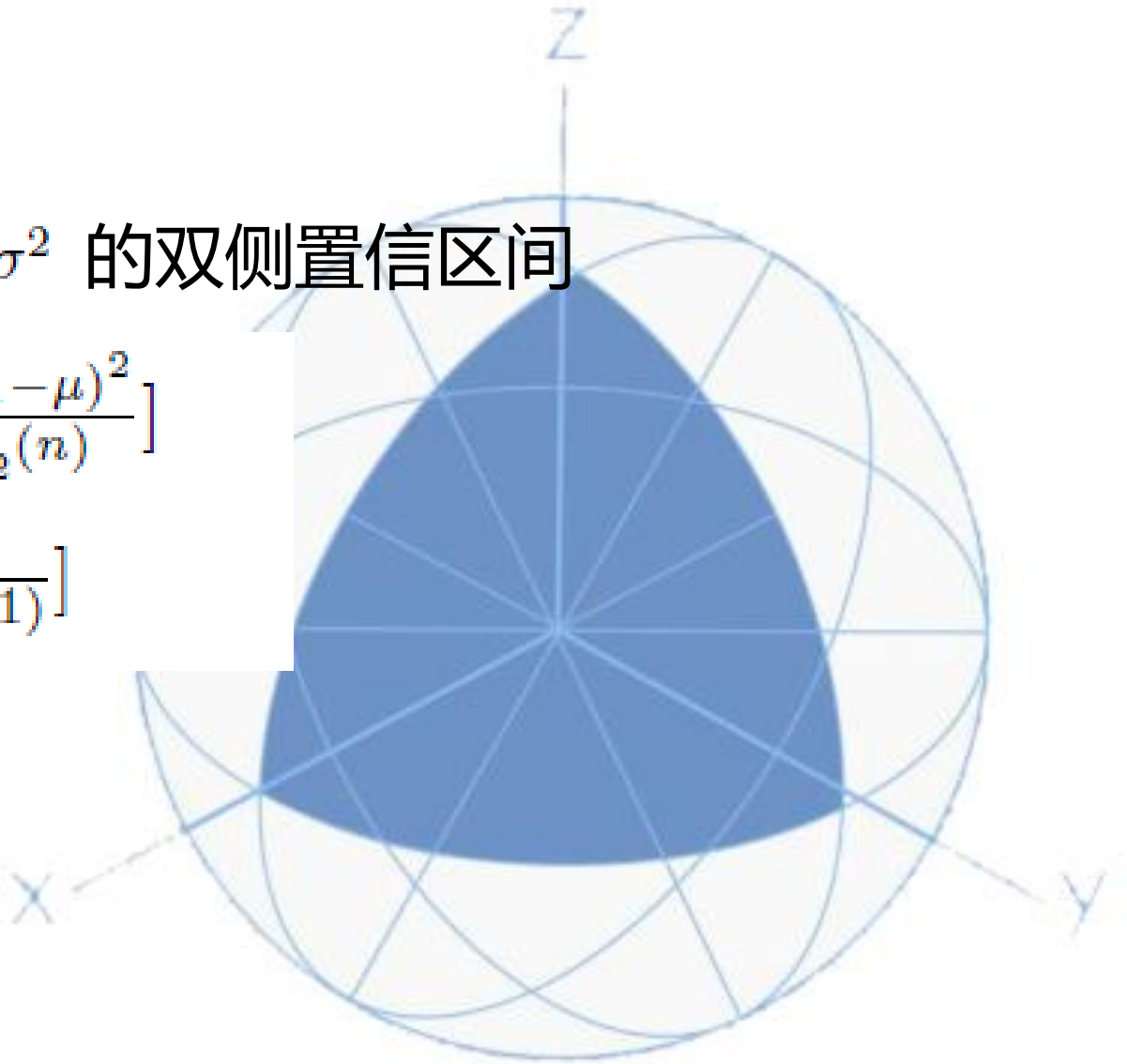
方差 σ^2 已知　　$[\overline{x} - \mu_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + \mu_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$

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 $X \sim N(\mu, \sigma^2), \sigma^2$ 的双侧置信区间

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

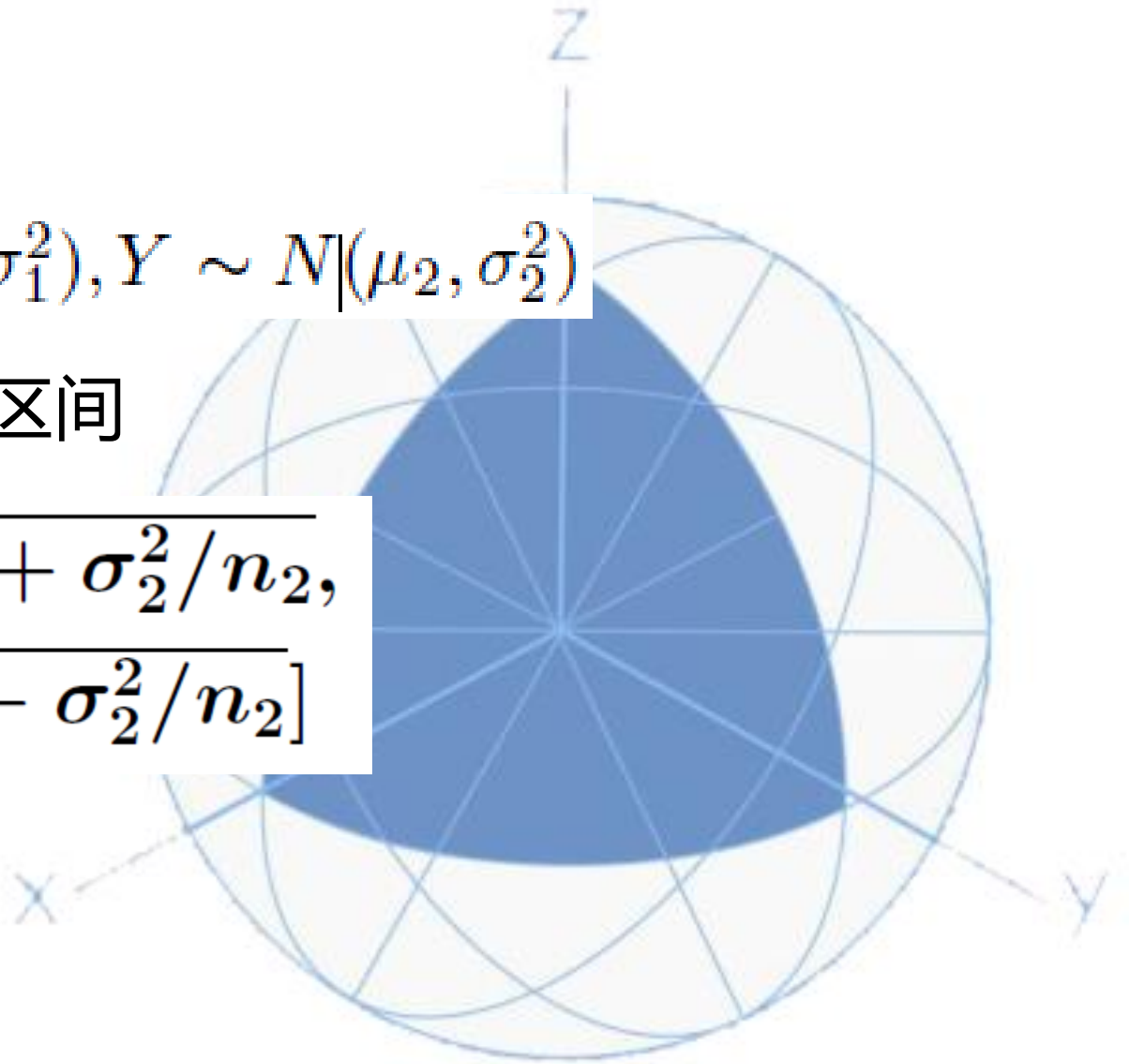
μ 未知 $\quad \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$



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

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



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

方差均未知但是相等时， $\mu 1 - \mu 2$ 的置信区间

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

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

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

μ_1, μ_2均未知时，方差比σ_1^2/σ_2^2的置信区间

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

例： 用天平称量某物体的质量9次，得平均值$\bar{x} = 15.4(g)$, 已知天平称量结果为正态分布，其标准差为0.1(g). 试求该物体质量的0.95置信区间.

解： σ^2已知，则μ的置信区间为 $[\overline{x} - u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + u_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}]$.

此处 $1-\alpha = 0.95, \alpha = 0.05,$ 查表知 $u_{0.975} = 1.96.$

由已知 $\bar{x} = 15.4, \sigma = 0.1, n = 9$ 代入上式可得其置信区间为 $[15.4 - 1.96 \times 0.1/\sqrt{9},\ 15.4 + 1.96 \times 0.1/\sqrt{9}]$ 即 $[15.3347,\ 15.4653]$.

6.1.3 假设检验

基本步骤

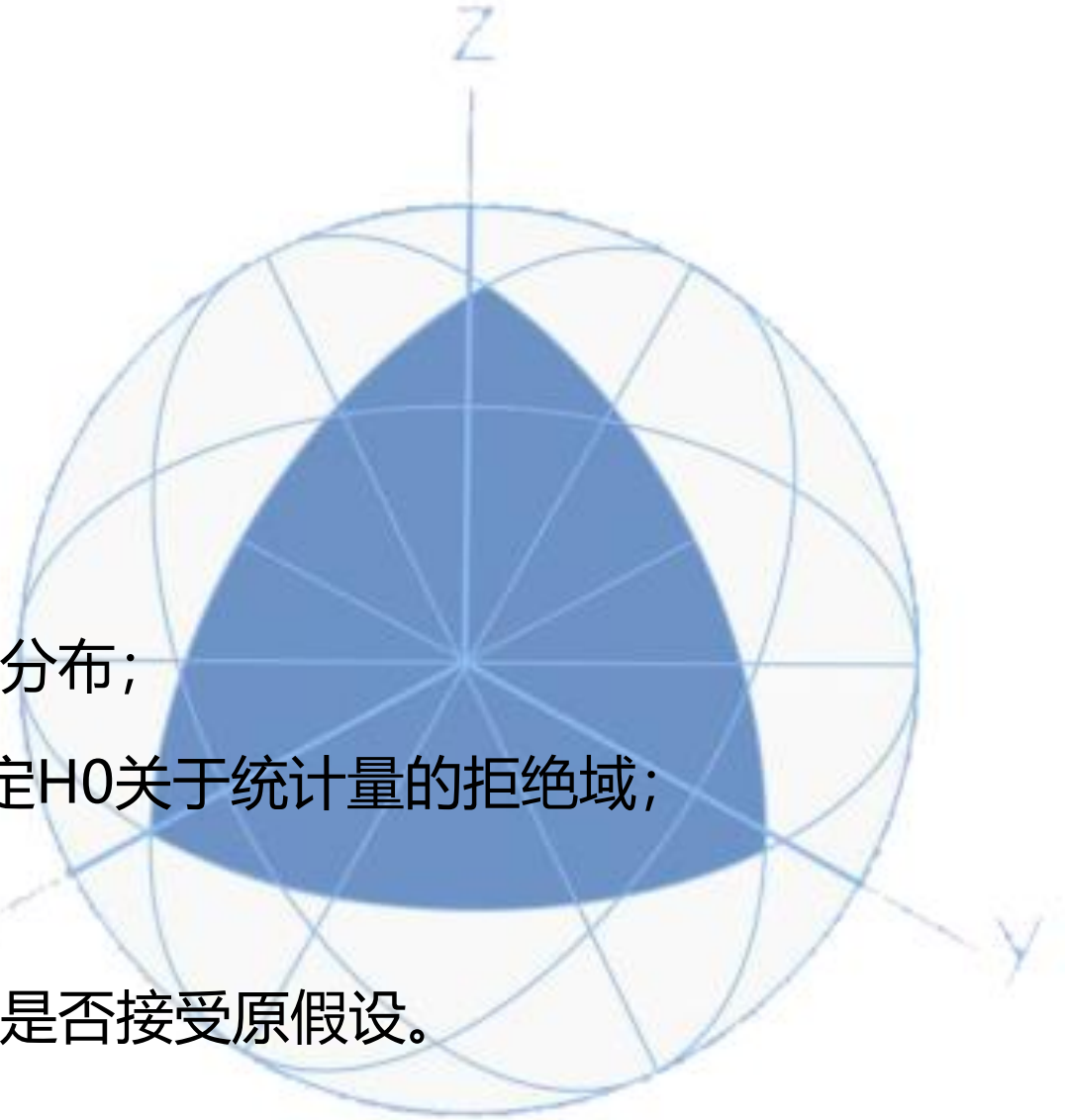
(1)提出假设：提出原假设H0与备择假设H1；

(2)建立检验统计量：选择检验统计量并确定其分布；

(3)确定拒绝域：在给定的显著性水平α下，确定H0关于统计量的拒绝域；

(4)计算：算出样本点对应的检验统计量的值；

(5)判断：统计值是否落在拒绝域内，从而判断是否接受原假设。



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?

解： 总体 $X \sim N(\mu, 0.015^2)$

待检验的原假设和备择假设分别为

$$H_0 : \mu = 0.5 \quad H_1 : \mu \neq 0.5.$$

检验的拒绝域为 $\{|u| \geq u_{1-\alpha/2}\}$. 取显著性水平$\alpha =$ 0.05，则查表可知$u_{0.975} = 1.96.$

由该例中的观测值可计算得出

$$\bar{x} = 0.511, u = \frac{0.511 - 0.5}{0.015/\sqrt{9}} = 2.244$$

u值落入拒绝域 $|u| \geq 1.96$ 中，故拒绝原假设，即认为这天的包装不正常。

也可采用p值完成此次检验. 此处u0=2.244, 则

$$\begin{aligned} p &= P(|u| \geq |u_0|) = 2(1-\Phi(|u_0|) \\ &= 2(1-\Phi(2.244)) = 0.0248 \end{aligned}$$

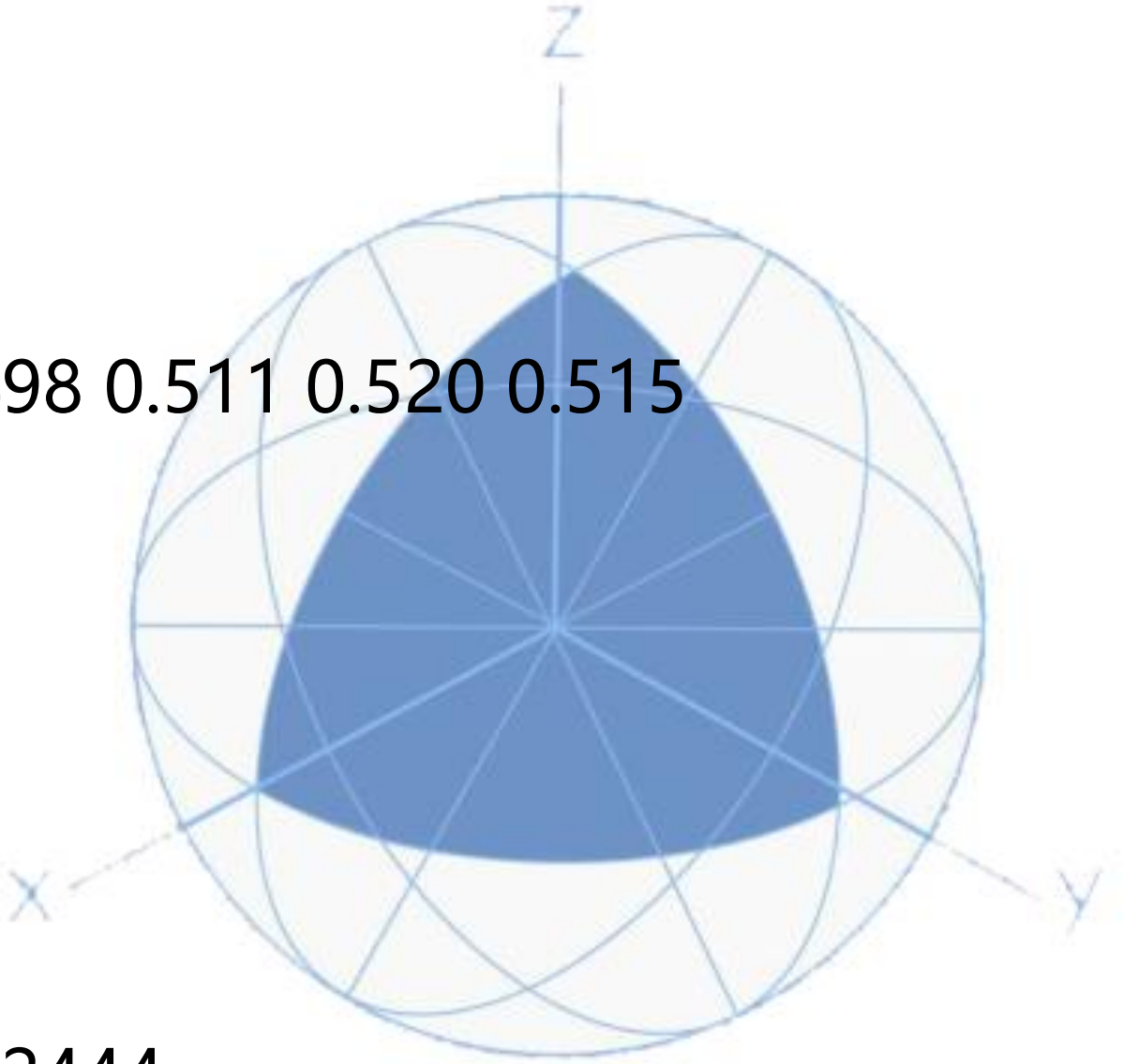
由于p值小于事先给定的水平0.05，故拒绝原假设，结论相同。

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,zval]=ztest(x,0.5,0.015)
```

(运行结果)

```
h =1    p =0.0248

ci =0.5014   0.5210      zval = 2.2444
```



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

x样本数据, m均值, sigma标准差, alpha置信水平（默认0.05）

tail=0, 即备择假设： $H_1 : \mu \neq \mu_0$ (默认，双边检验)

tail=1，表示备择假设： $H_1 : \mu > \mu_0$ (单边检验)；

tail=-1，表示备择假设： $H_1 : \mu < \mu_0$ (单边检验)

h=0 表示在显著性水平下，不能拒绝原假设，

h=1 表示在显著性水平下可以拒绝原假设

p 为真正观察值的概率，p为小概率时则对原假设提出质疑

ci 为真正均值的置信区间，zval 为统计量的值。

(matlab求解结果)

h=1，p=0.0248，ci =[0.5014 0.5210], zval=2.2444

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.

厦门大学
XIAMEN UNIVERSITY

Part 2

回归分析思想与建模方法

6.2.1 回归的基本概念

1、回归分析名称的由来

回归分析的基本思想和方法以及回归名称的由来归功于英国统计学家高尔顿，1886年他在研究家族成员之间的遗传规律时发现：

虽然高个子父亲有生高个孩子的趋势, 但一群高个子父亲的儿子平均身高却低于父亲们的身高；反之，一群矮个子父亲的儿子们的平均身高却高于父亲们的平均身高.换而言之，即子代的平均身高有向同龄人平均身高回归的趋势. 这也才使得人类身高在一定时间内相对稳定，而不会出现父辈高其子女更高，父辈矮其子女更矮的两极分化现象.

2、回归分析的一般模型

若变量x1,x2,...,xp与变量y之间存在某种相关关系，通常意味着当x1,x2,...,xp取定值后，y便有相应的概率分布与之对应. 因此我们可以将回归模型的一般形式表示为

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon$$

其中随机变量y 被称为因变量，也被称为被解释变量，对应的x1,x2,...,xp称为自变量，也称为解释变量.

此外在计量经济学中也称因变量为内生变量，自变量为外生变量. 只有一个自变量的回归分析,称之为**一元回归分析**，多于一个自变量称之为**多元回归分析**.

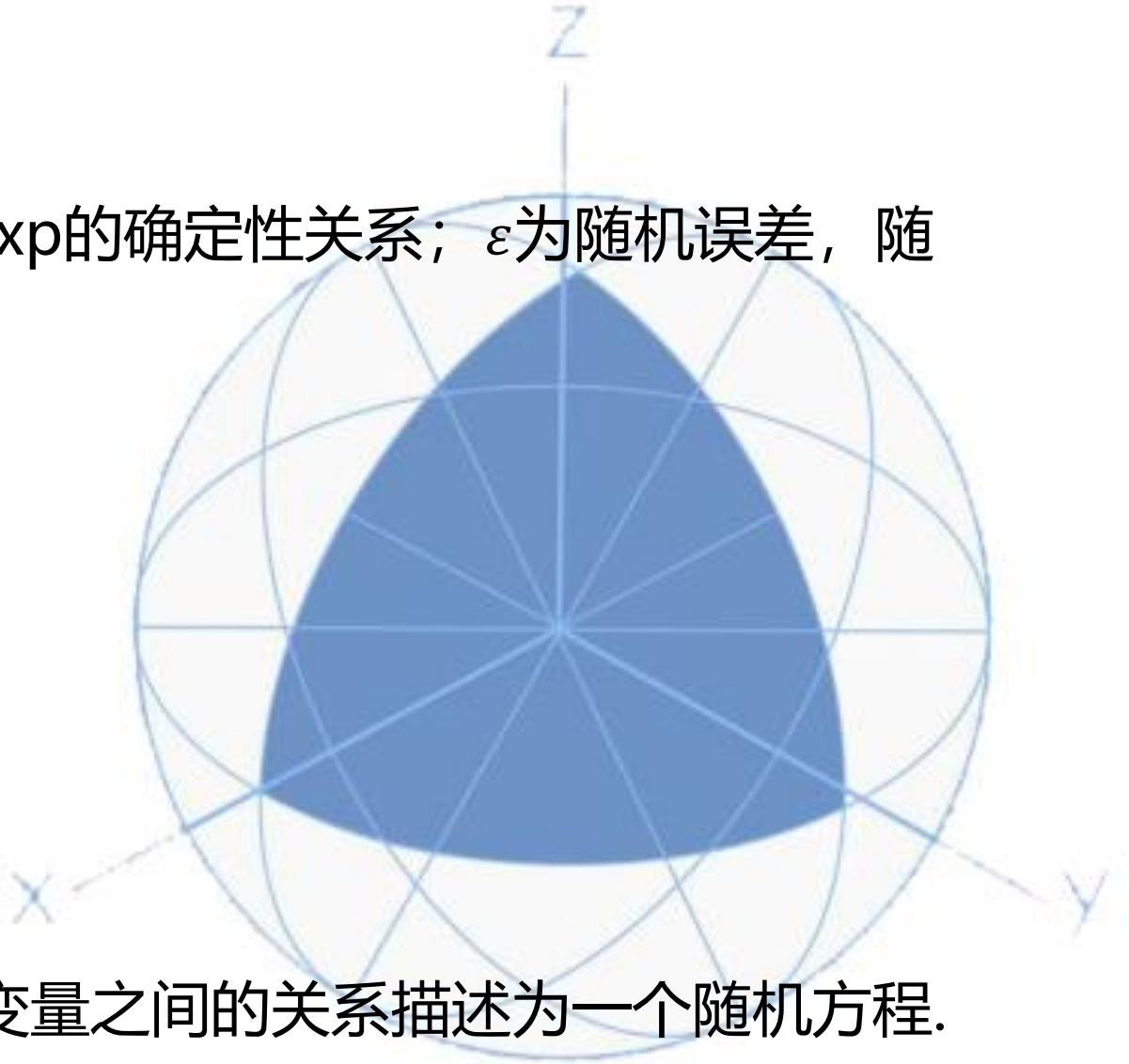
另外f(x1,x2,...,xp) 一般为变量x1,x2,...,xp的确定性关系；ε为随机误差，随机误差项一般包括下列因素的影响：

（1）解释变量中被忽略的因素的影响；

（2）样本数据采集过程中的观测误差；

（3）理论模型设定偏差的影响；

（4）其他因素的影响.



正是由于随机误差项ε的引入，才使得变量之间的关系描述为一个随机方程.

3、回归分析研究的主要问题

(1)确定y与x1,x2,...,xp间的定量关系表达式，这种表达式称为回归方程;

(2)对所得方程进行可信度检验，包括参数的显著性检验，置信区间的估计等等;

(3)判断自变量xj(j=1,2,...,p)对y有无影响;

(4)利用所得回归方程进行预测和控制.

6.2.2 一元线性回归

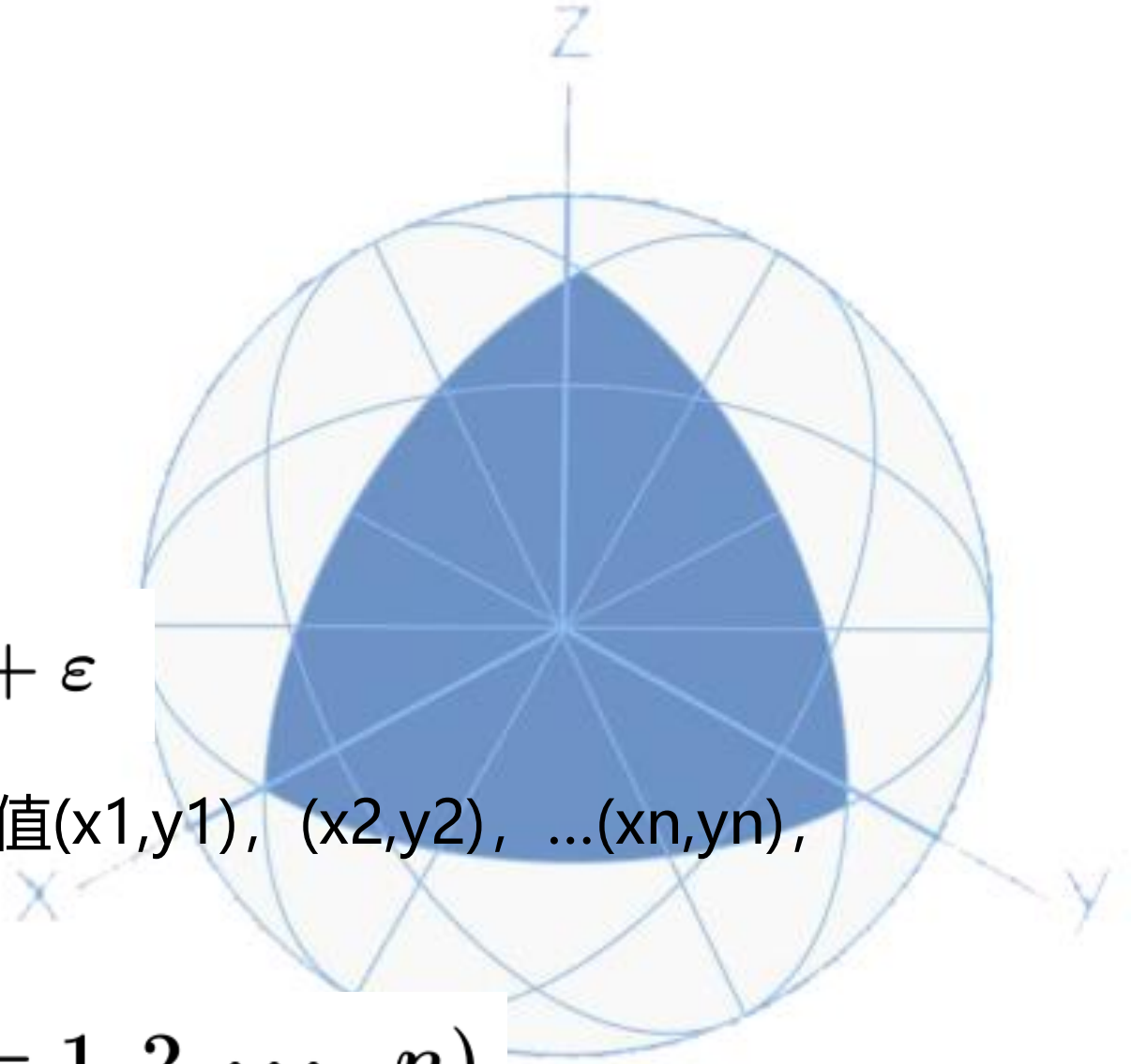
1、一般形式

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(x1,y1)，(x2,y2)，...(xn,yn)，则上述模型可以等价的写为：

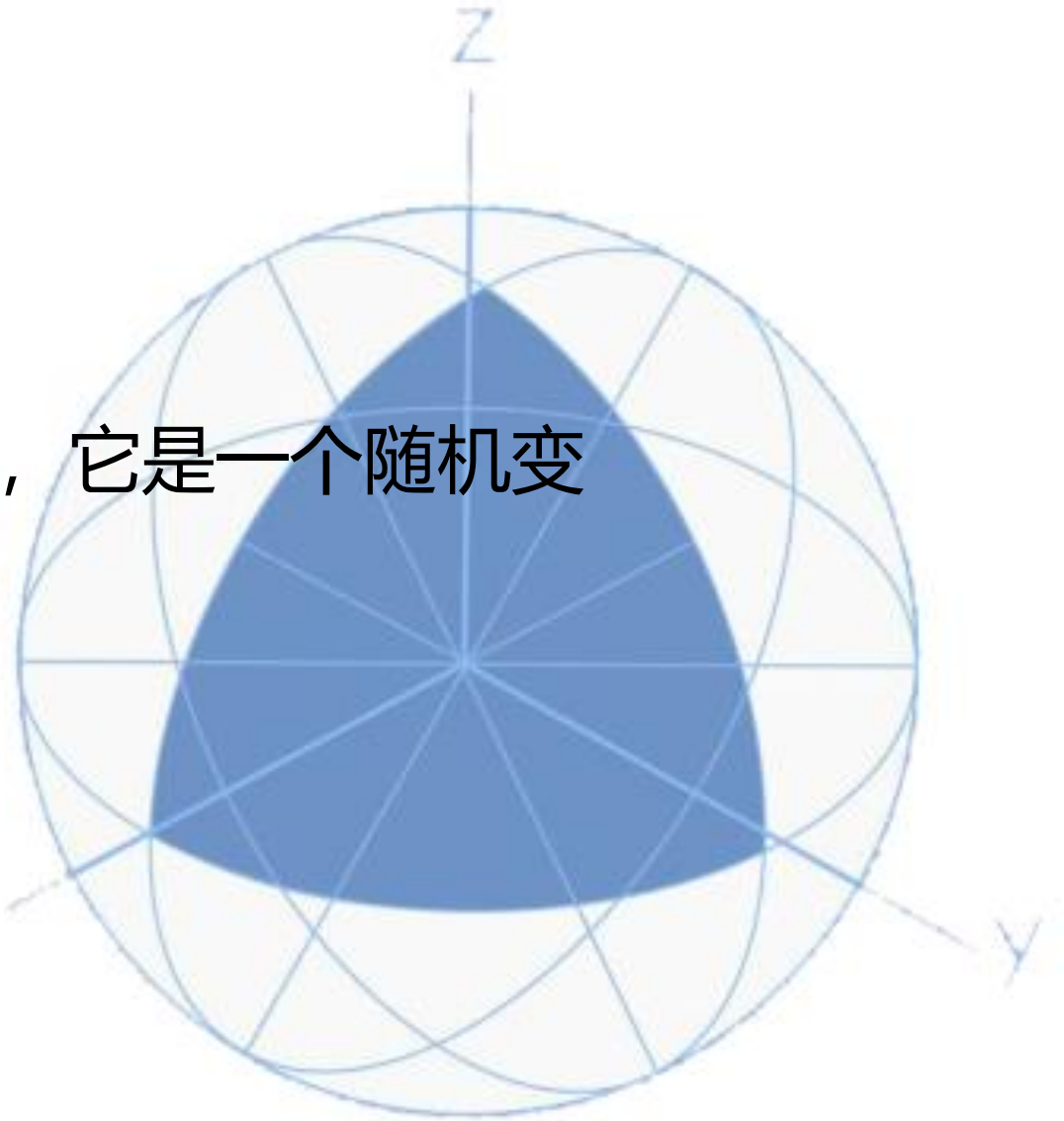
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, (i = 1, 2, \cdots, n)$$



模型假设:

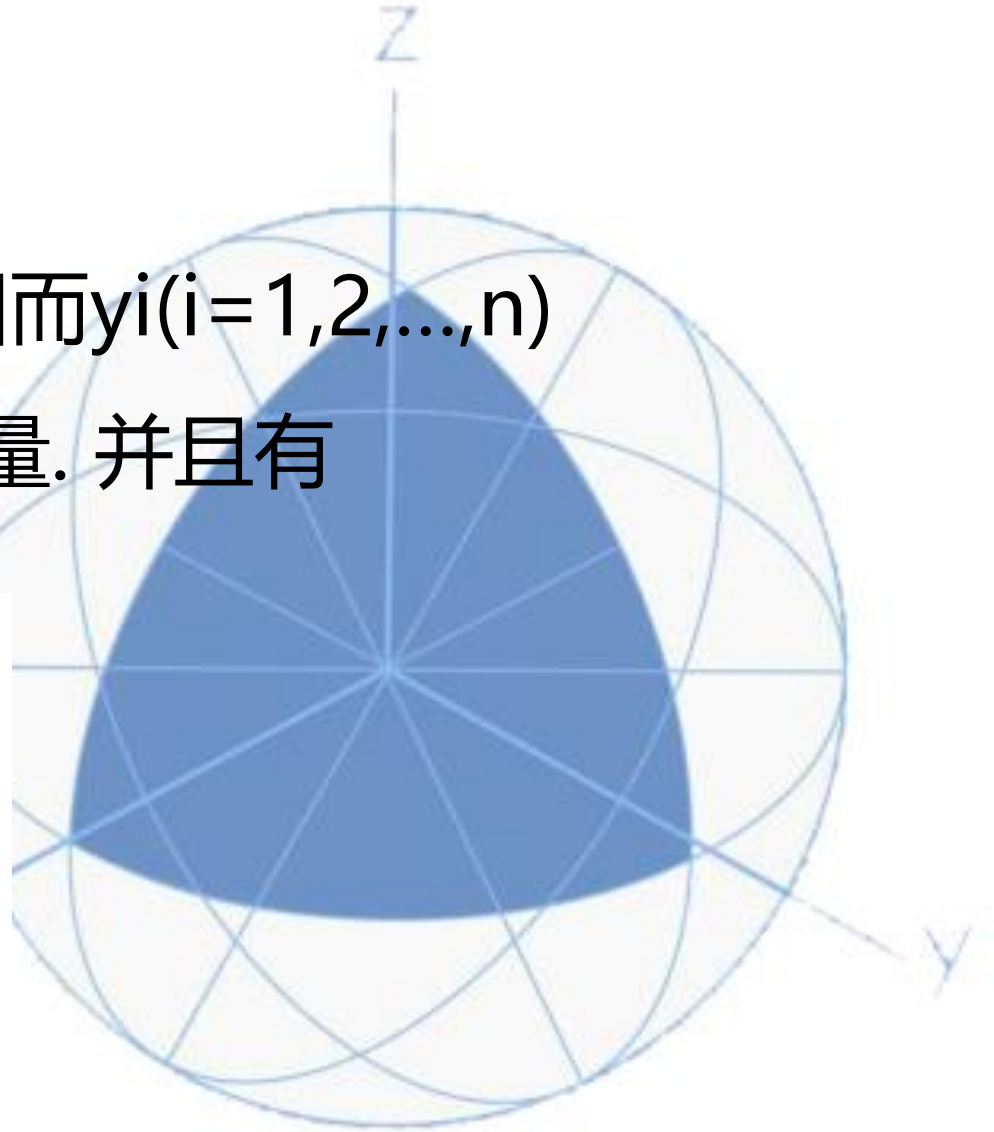
（1）假定ε是不可观测的随机误差项，它是一个随机变量，满足

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



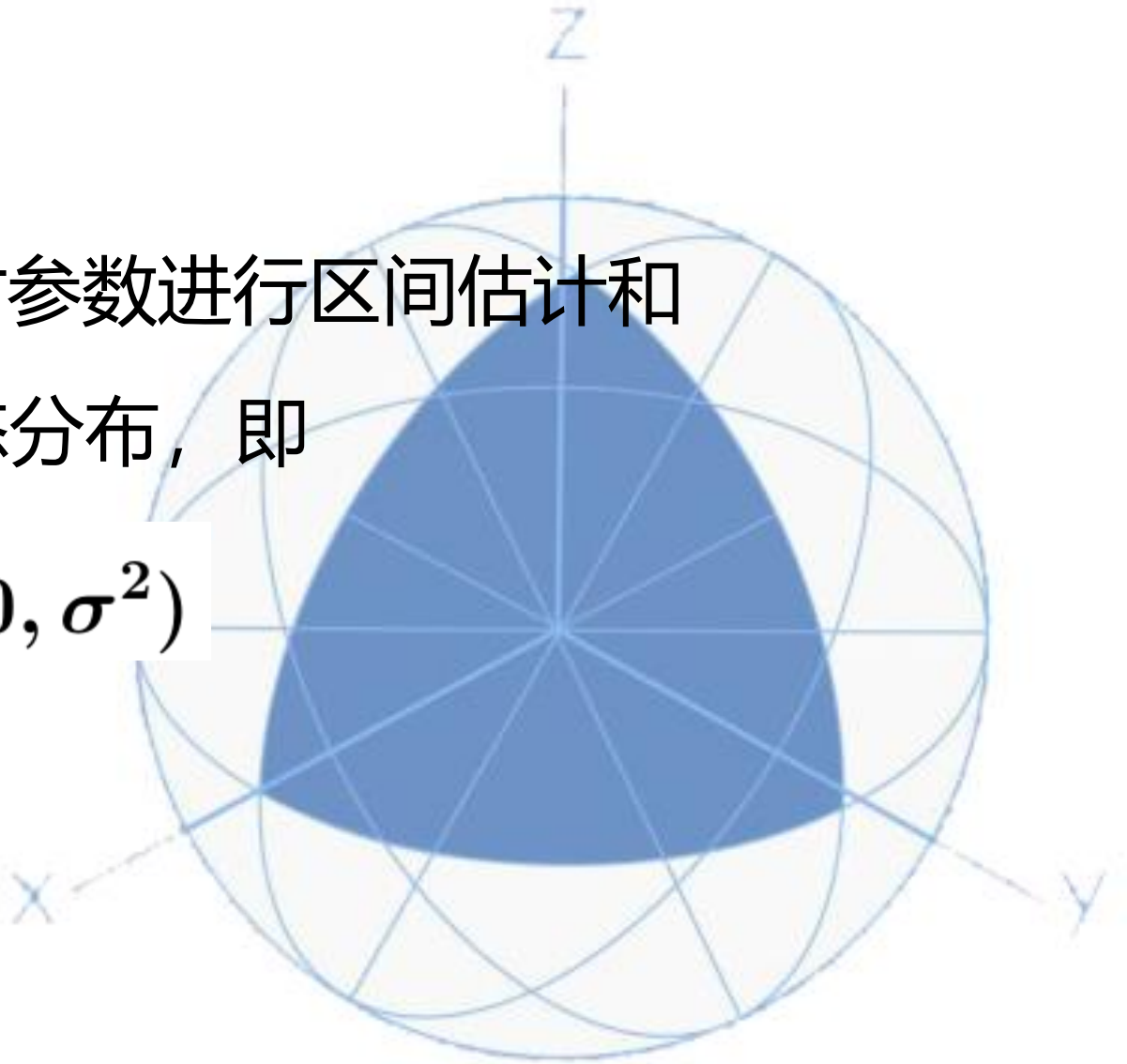
（2）假设n组数据是独立观测的，因而yi(i=1,2,...,n)与ε_i(i=1,2,...,n)是相互独立的随机变量. 并且有

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



（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

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



矩阵形式

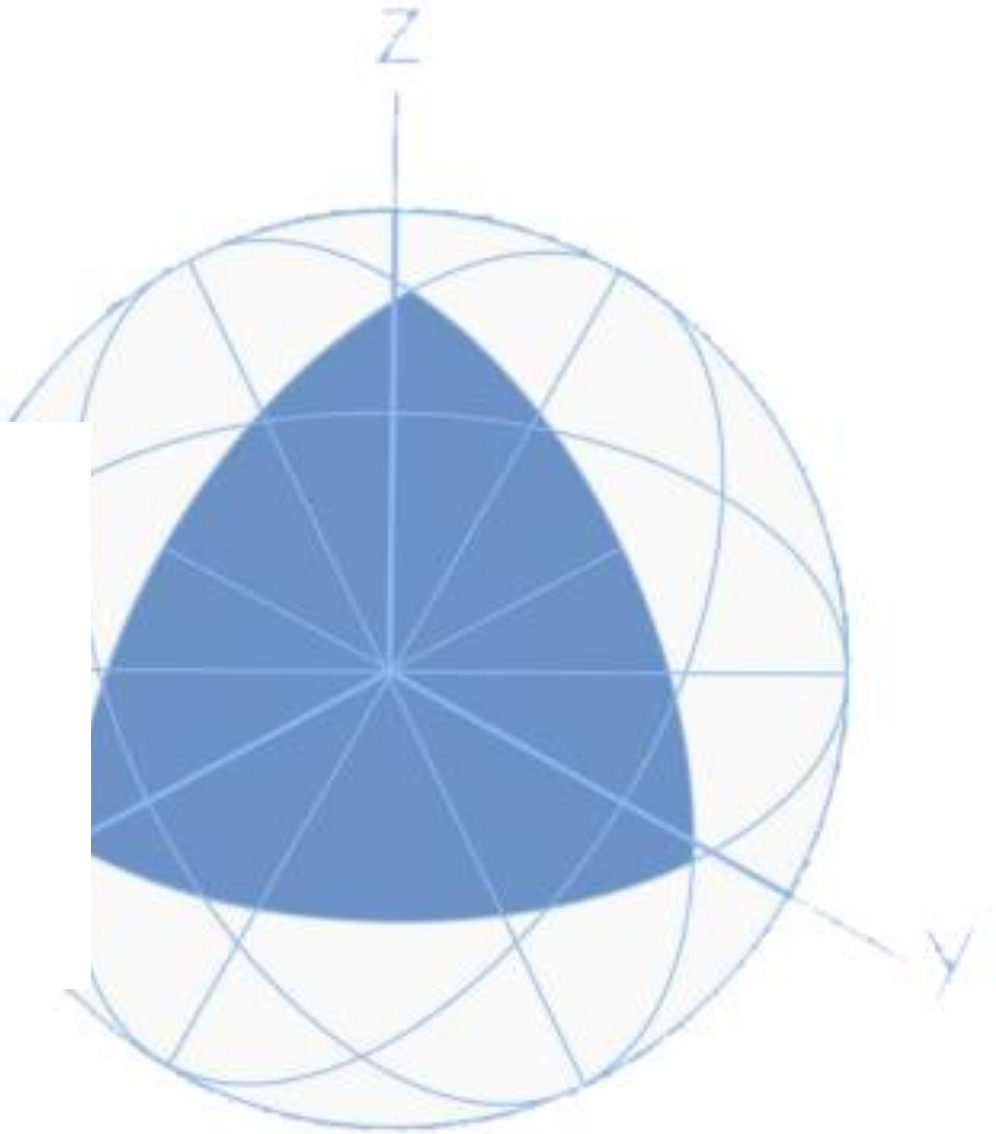
设

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$$\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

所以模型可以等价的转化为：

$$\begin{cases} \boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon \\ \boldsymbol{E}(\varepsilon) = 0 \\ \boldsymbol{var}(\varepsilon) = \boldsymbol{\sigma^2} \boldsymbol{I}_n \end{cases}$$



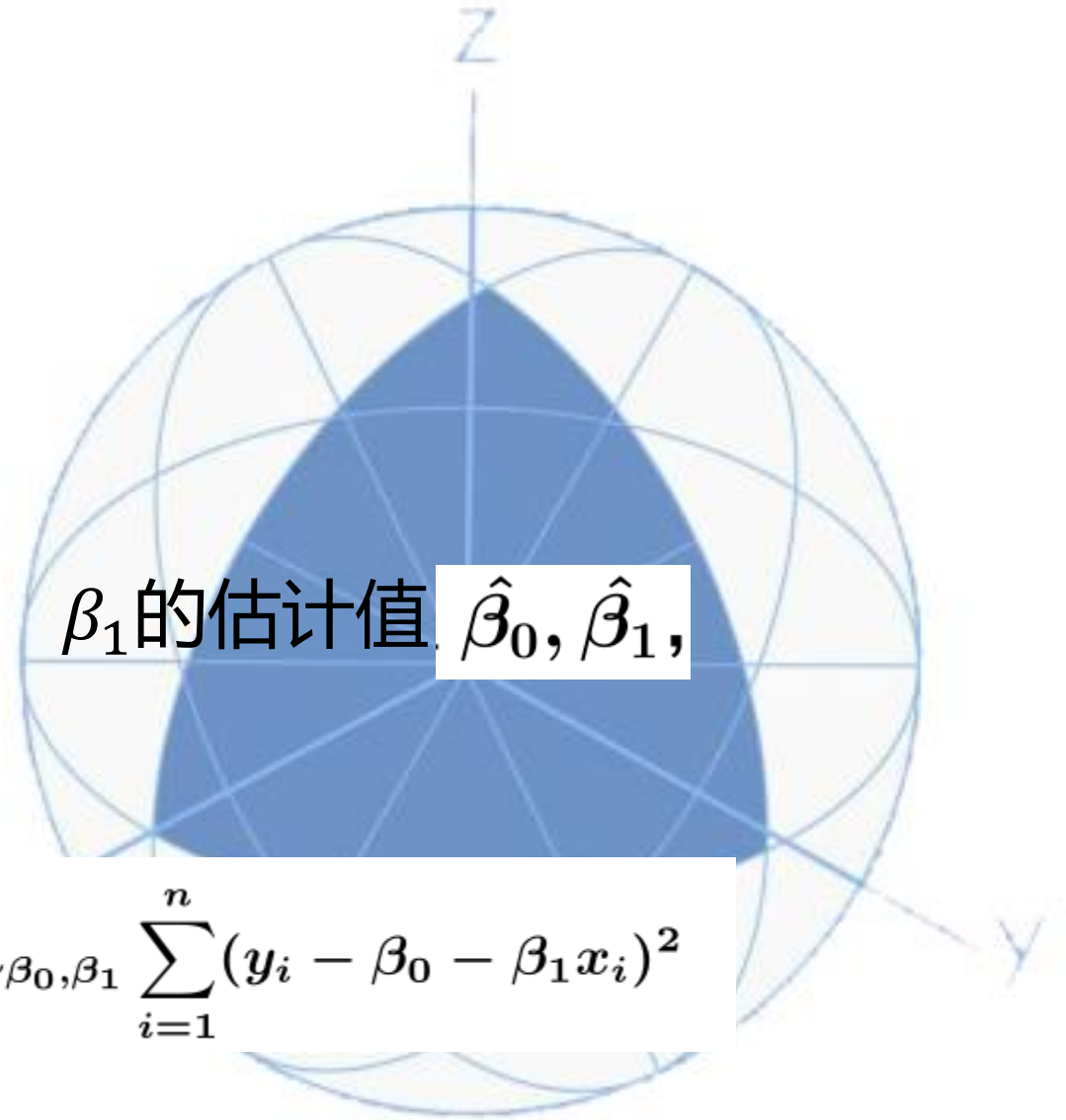
2、参数估计

(1)普通最小二乘估计(OLSE)

所谓最小二乘估计就是寻找参数β_0，β_1的估计值$\hat{\beta}_0, \hat{\beta}_1$，

使得离差平方和最小，即

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



由于Q 是关于 $\hat{\beta}_0, \hat{\beta}_1$ 的非负二次函数，因而极小值总是存在的. 对离差平方和 $Q(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

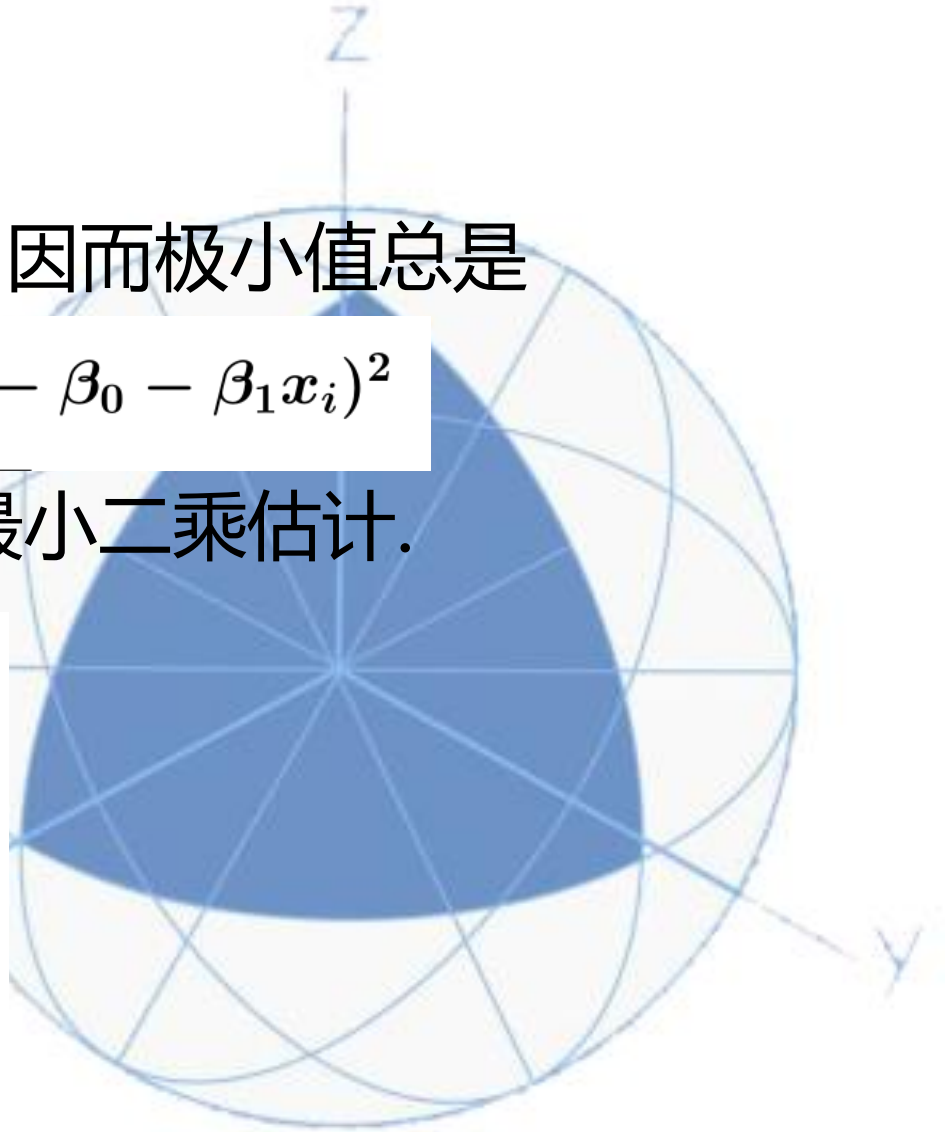
分别关于β_0， β_1求导即可得β_0， β_1的最小二乘估计.

记

$$L_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

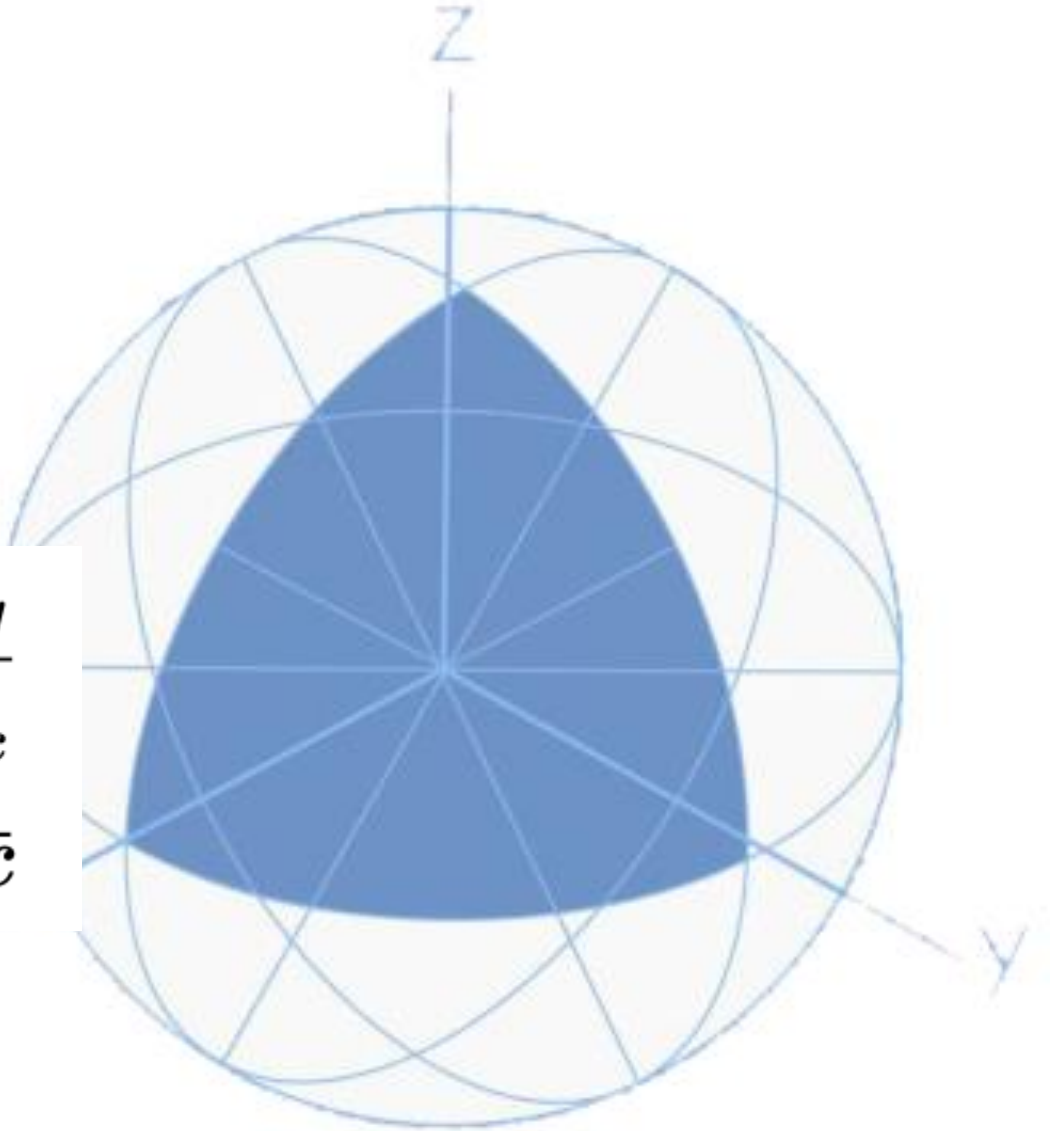
$$L_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$L_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$



则最小二乘估计可表示为

$$\begin{cases} \hat{\beta}_1 = \dfrac{L_{xy}}{L_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$



(2)极大似然估计(MLE)

由之前的模型假设条件有

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

所以y1,y2,...,yn的似然函数为

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f_i(y_i)$$

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

相应的对数似然函数为

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

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.

同时最大似然估计还可以得出σ^2 的估计值为

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

不过这个估计量是有偏的，实际应用中，常用的无偏估计量为

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

此外由 $E(y_i) = \beta_0 + \beta_1 x_i,\ var(y_i) = \sigma^2,$

我们很容易可以得到以下结论：

$$E(\hat{\beta}_1) = \beta_1, var(\hat{\beta}_1) = \frac{\sigma^2}{L_{xx}}$$
$$E(\hat{\beta}_0) = \beta_0, var(\hat{\beta}_0) = [\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}]\sigma^2$$

同时在高斯马尔科夫条件（误差满足均值为零、同方差且互不相关）下，我们可以证明还是最小方差无偏估计量.

3、显著性检验

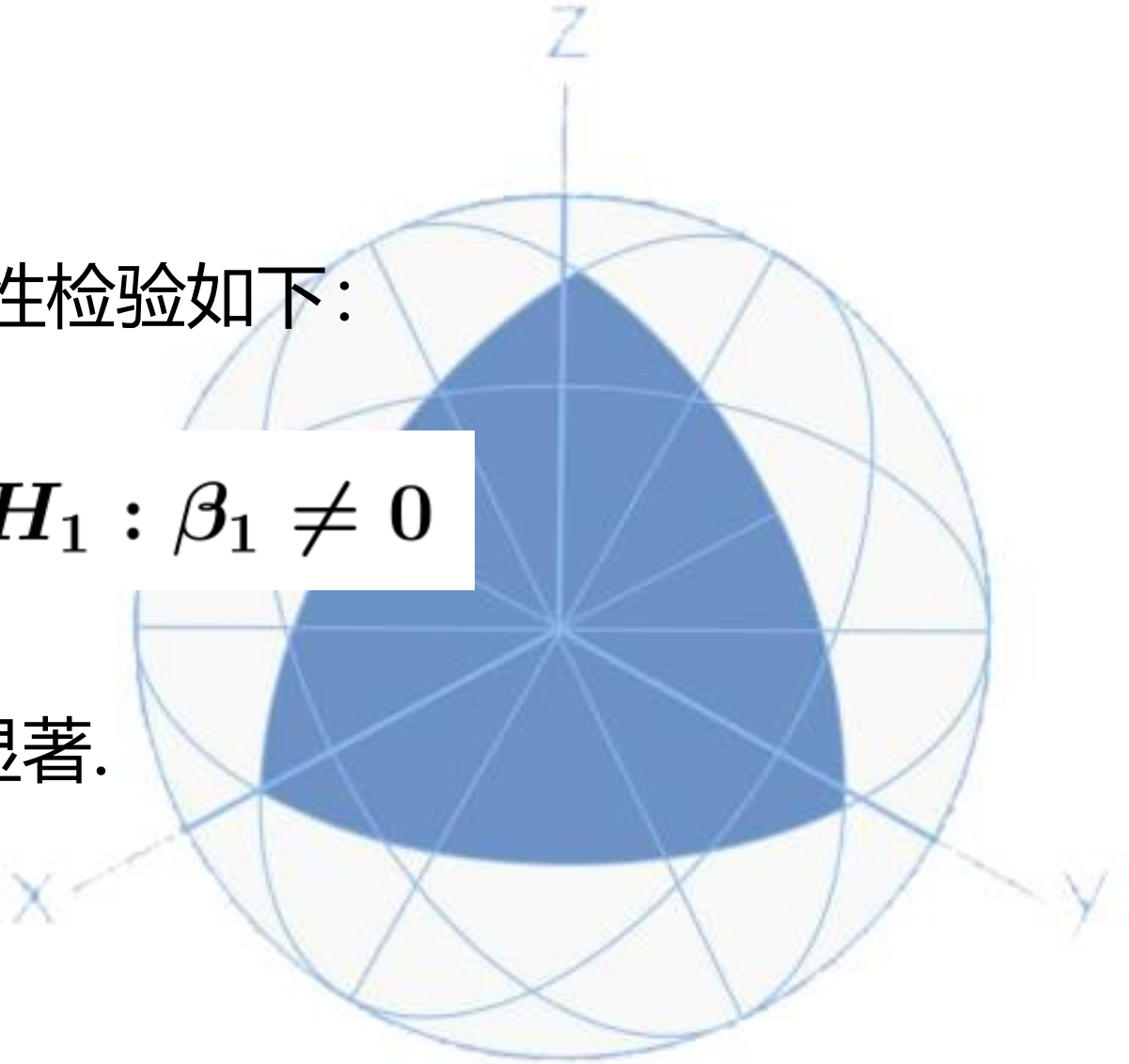
对于一元线性回归方程 $E(y) = \beta_0 + \beta_1 x$，当 $\beta_1 = 0$ 时，不管x如何变化，E(y)都不随x的变化做线性变化，这是求得的回归方程就没有意义，也称回归方程不显著.

反之，若 $\beta_1 \neq 0$ 时，E(y)都随x的变化做线性变换，称为回归方程显著.

综上，一元线性回归方程的显著性检验如下：

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

拒绝原假设H0 即表示回归方程显著.



（1）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 $F = \frac{SSR/1}{SSE/(n-2)}$

这里不证明给出的两个结论（概率统计书）

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

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

$$W = \{F \geq F_{1-\alpha}(1, n-2)\}$$

相应的回归分析表如下：

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			

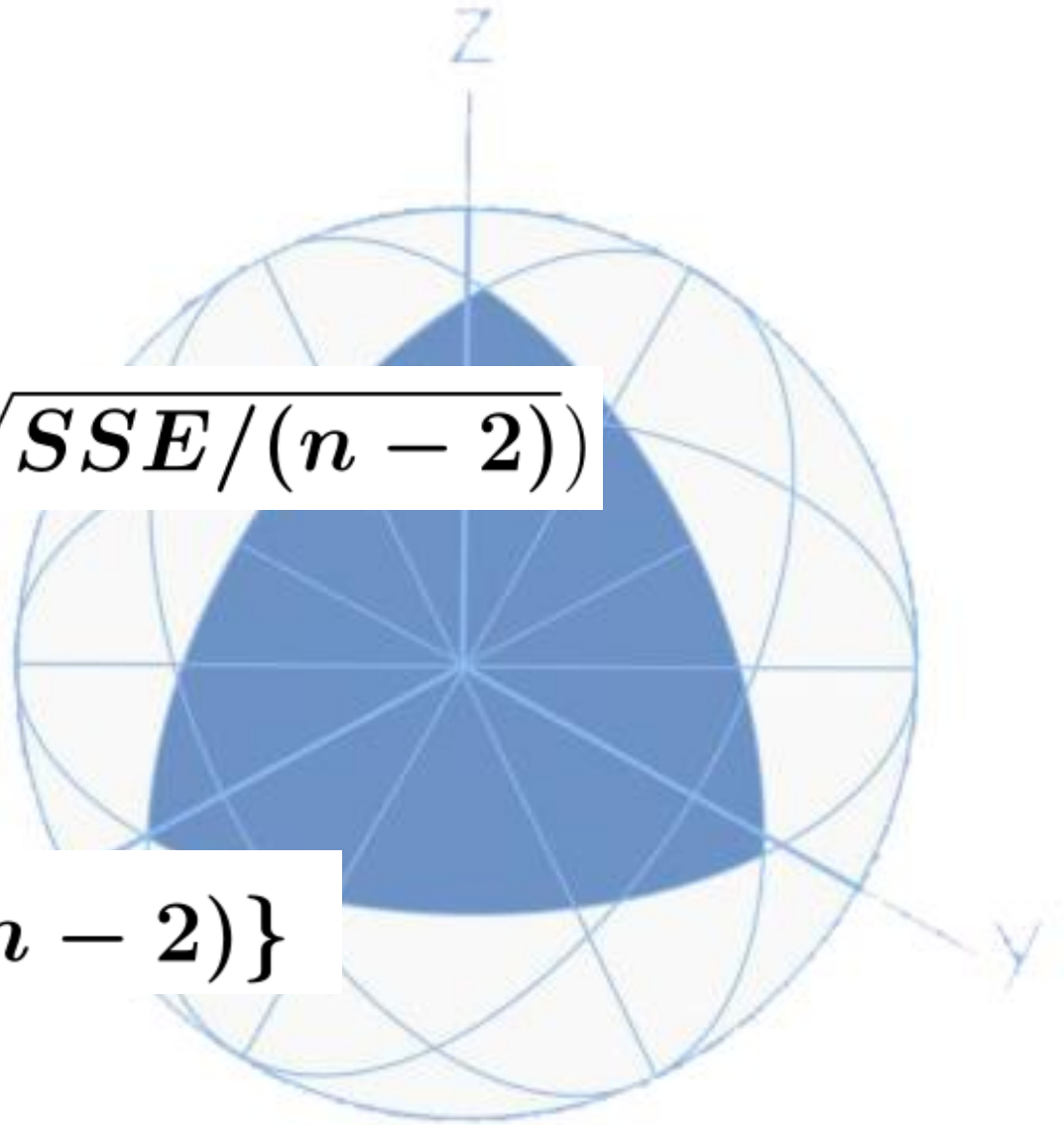
（2）t检验

构造 t 统计量 $t = \frac{\hat{\beta}_1\sqrt{L_{xx}}}{\hat{\sigma}}(\hat{\sigma} = \sqrt{SSE/(n-2)})$

在假设条件下有 $t \sim t(n-2)$.

当显著水平为α时，拒绝域为

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



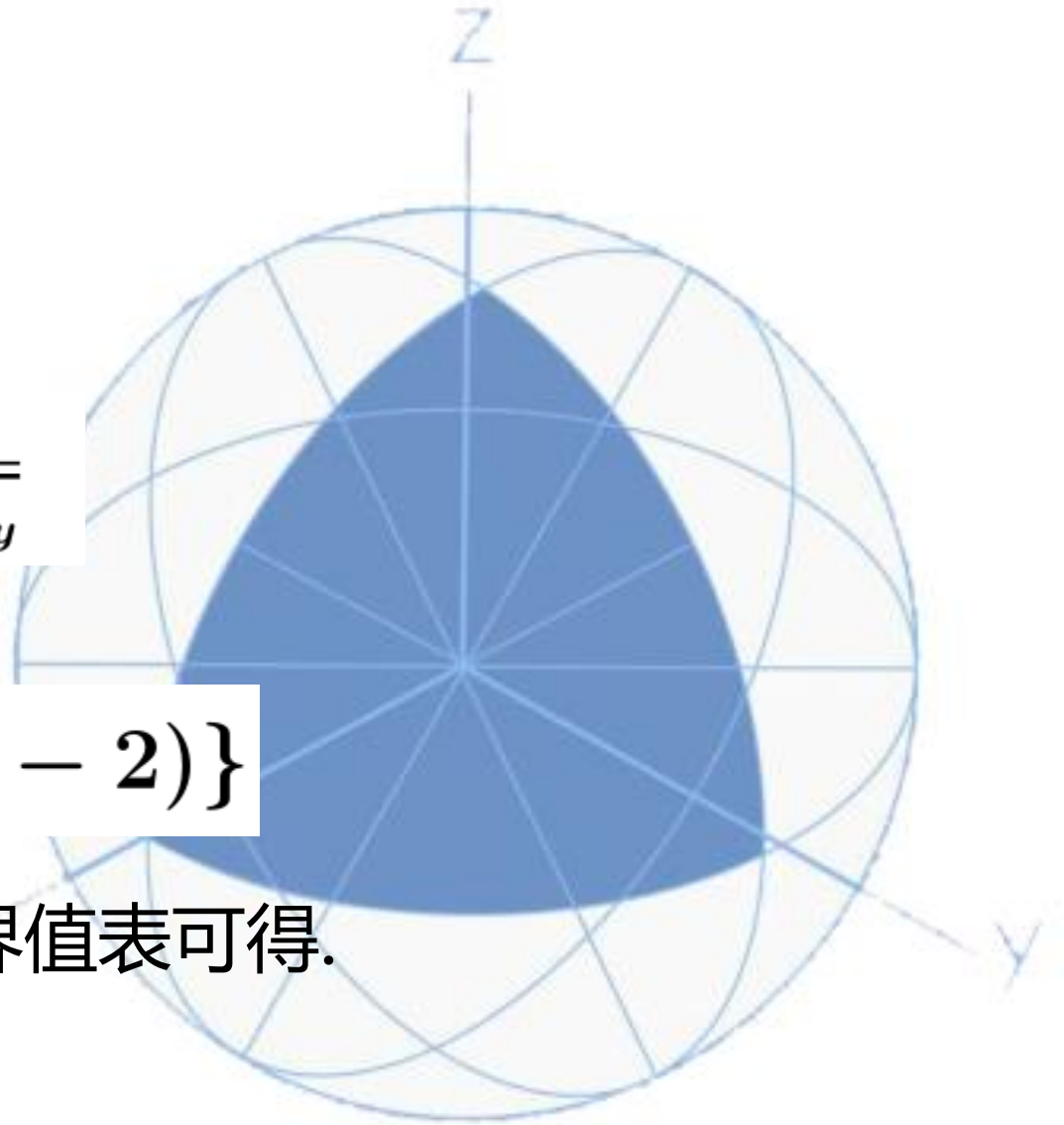
（3）相关系数检验

记样本相关系数为 $r = \dfrac{L_{xy}}{\sqrt{L_{xx}}\sqrt{L_{yy}}}$

检验的拒绝域为

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

其中 $r_{\alpha}(n-2)$ 查相关系数的临界值表可得.



4、回归系数的区间估计

由 $\hat{\beta}_0$、$\hat{\beta}_1$ 的统计性质可得：

$$t_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{var(\hat{\beta}_i)}} \sim t(n-2), (i = 1, 2)$$

则置信水平为$1-\alpha$的β_i的区间估计为：

$$[\hat{\beta}_i - \sqrt{var(\hat{\beta}_i)}t_{1-\alpha/2}(n-2),\ \hat{\beta}_i + \sqrt{var(\hat{\beta}_i)}t_{1-\alpha/2}(n-2)]$$

5、估计和预测

当回归方程经过检验是显著的后，才可以用来做估计和预测，这是两个不同的问题：

(1)估计问题：当给定x=x0时，求相应平均值 $E(y_0) = \beta_0 + \beta_1 x_0$ 的点估计与其置信水平为$1-\alpha$的区间估计；

对于均值E(y0) 而言，因为 $E(y_0) = \beta_0 + \beta_1 x_0$ 为常数，其一个直观的估计就是

$$E(\hat{y}_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

简单起见我们记为 $\hat{y}_0$. 则显然 $\hat{y}_0$ 是E(y0)的无偏估计.又因为是

$$\hat{y}_0 - E(y_0) \sim N(0, [\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{L_{xx}}]\sigma^2)$$

进而可得E(y0)的$1-\alpha$置信区间为

$$[\hat{y}_0 - t_{1-\alpha/2}(n-2)\sqrt{h_0}\hat{\sigma}, \hat{y}_0 + t_{1-\alpha/2}(n-2)\sqrt{h_0}\hat{\sigma}]$$

其中 $h_0 = \frac{1}{n} + \frac{(\bar{x}-x_0)^2}{L_{xx}}$

(2)预测问题：当给定x=x0 时，对应的 $y_0 = \beta_0 + \beta_1 x_0 + \epsilon$

是一个随机变量，求其置信水平为$1-\alpha$的预测区间.

对于x=x0, y=y0 时，因为：

$$y_0 - \hat{y}_0 \sim N(0, [1 + \frac{1}{n} + \frac{(\bar{x}-x_0)^2}{L_{xx}}]\sigma^2)$$

进而统计量

$$t = \frac{y_0-\hat{y}_0}{\sqrt{1+h_0}\hat{\sigma}} \sim t(n-2)$$

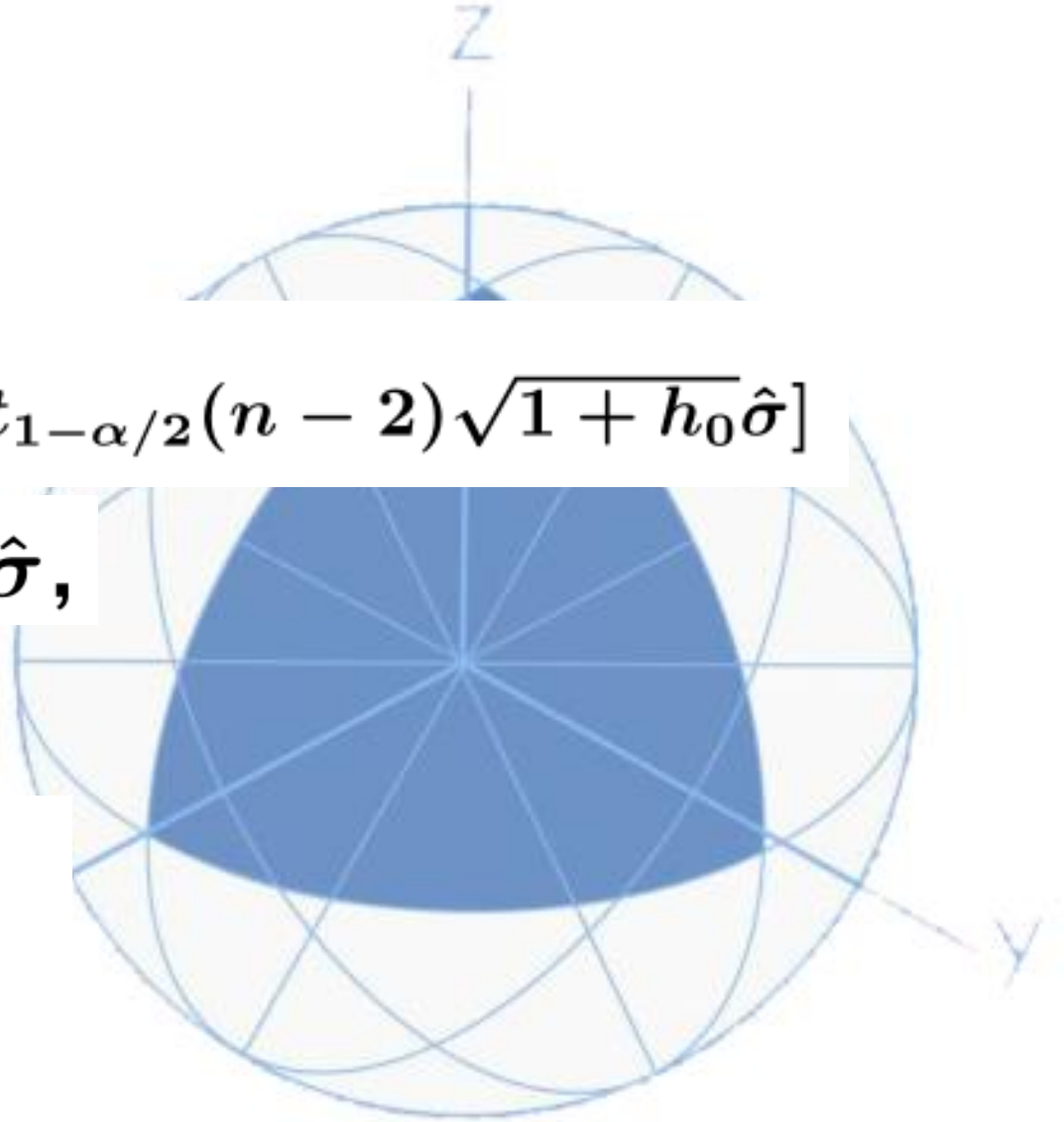
所以置信度为$1-\alpha$的预测区间为:

$$[\hat{y_0}-t_{1-\alpha/2}(n-2)\sqrt{1+h_0}\hat{\sigma},\hat{y_0}+t_{1-\alpha/2}(n-2)\sqrt{1+h_0}\hat{\sigma}]$$

记 $h=t_{1-\alpha/2}(n-2)\sqrt{1+h_0}\hat{\sigma}$,

则预测区间为

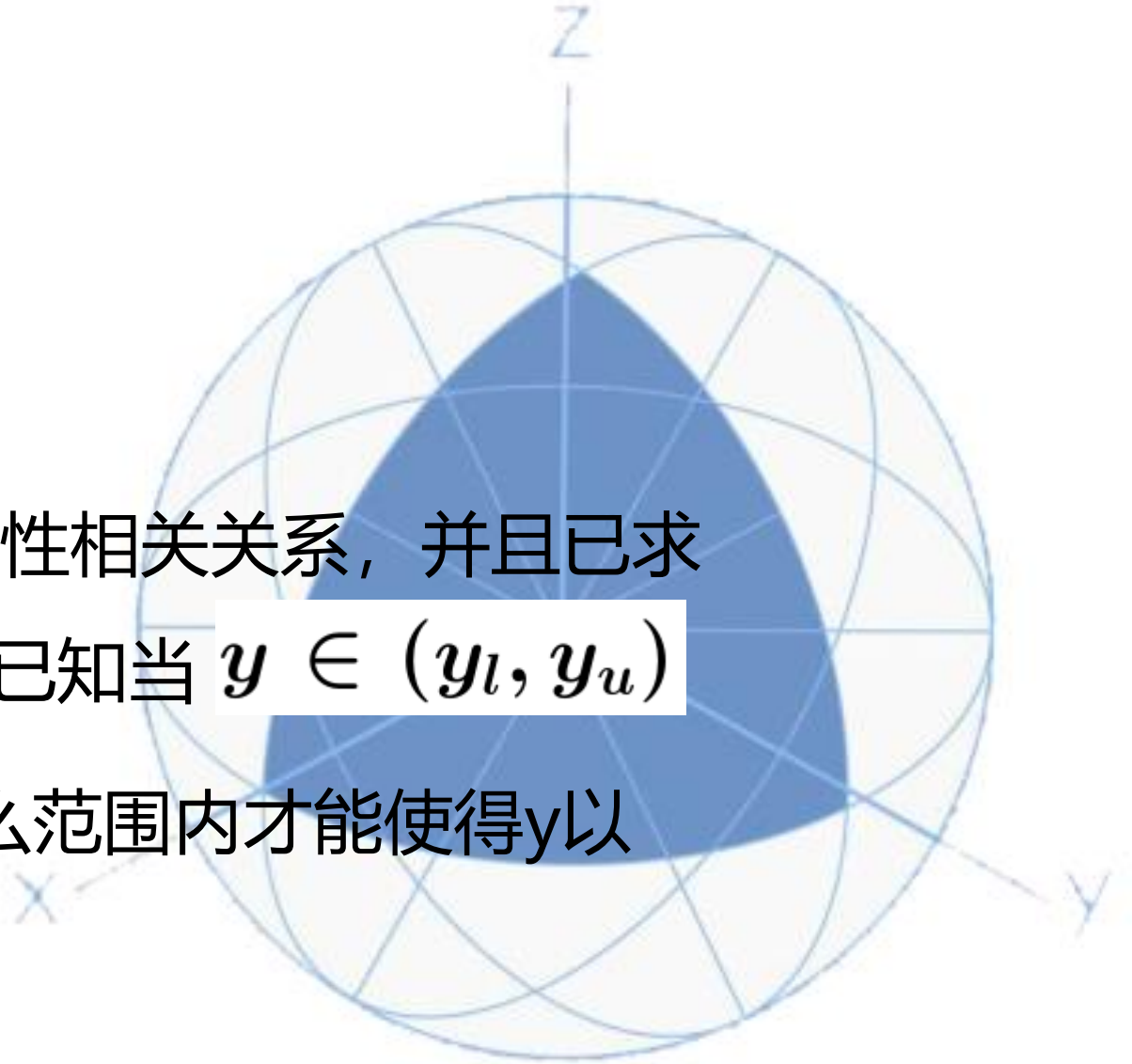
$$[\hat{y_0}-h,\hat{y_0}+h]$$



6、控制

问题提出：

假设质量指标y和某一自变量x有线性相关关系，并且已求得回归方程为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. 已知当 $y \in (y_l, y_u)$ 时为质量合格，那么x应控制在什么范围内才能使得y以 $1-\alpha$的概率合格呢?



控制可以看成是预测的反问题，因而有不等式组：

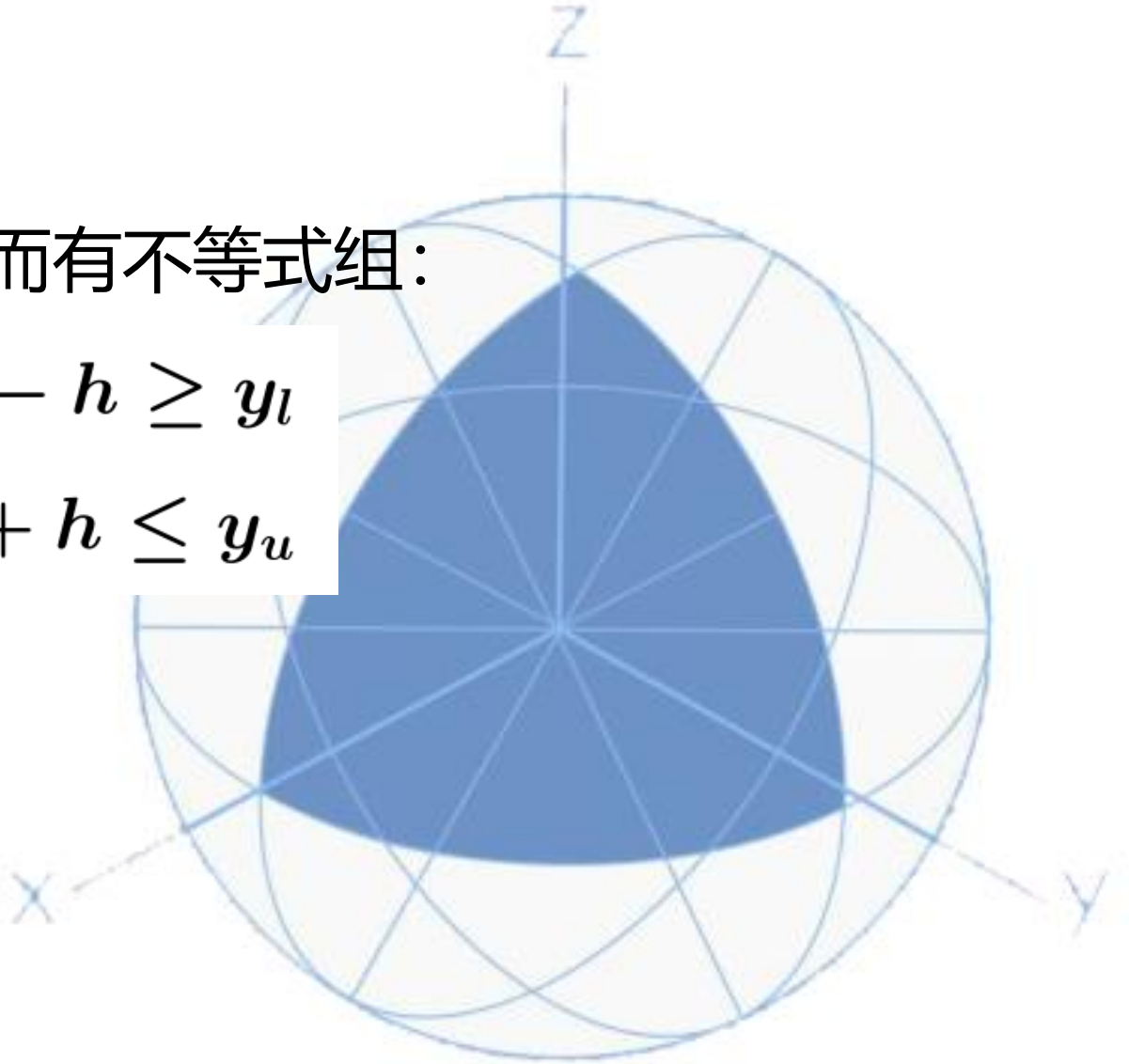
$$\begin{cases} \hat{y} - h = \hat{\beta}_0 + \hat{\beta}_1 x - h \geq y_l \\ \hat{y} + h = \hat{\beta}_0 + \hat{\beta}_1 x + h \leq y_u \end{cases}$$

当 $\hat{\beta}_1 > 0$ 时，

有 $\frac{y_l - \hat{\beta}_0 + h}{\hat{\beta}_1} \leq x \leq \frac{y_u - \hat{\beta}_0 - h}{\hat{\beta}_1}$

当 $\hat{\beta}_1 < 0$ 时，

有 $\frac{y_u - \hat{\beta}_0 - h}{\hat{\beta}_1} \leq x \leq \frac{y_l - \hat{\beta}_0 + h}{\hat{\beta}_1}$



案例分析

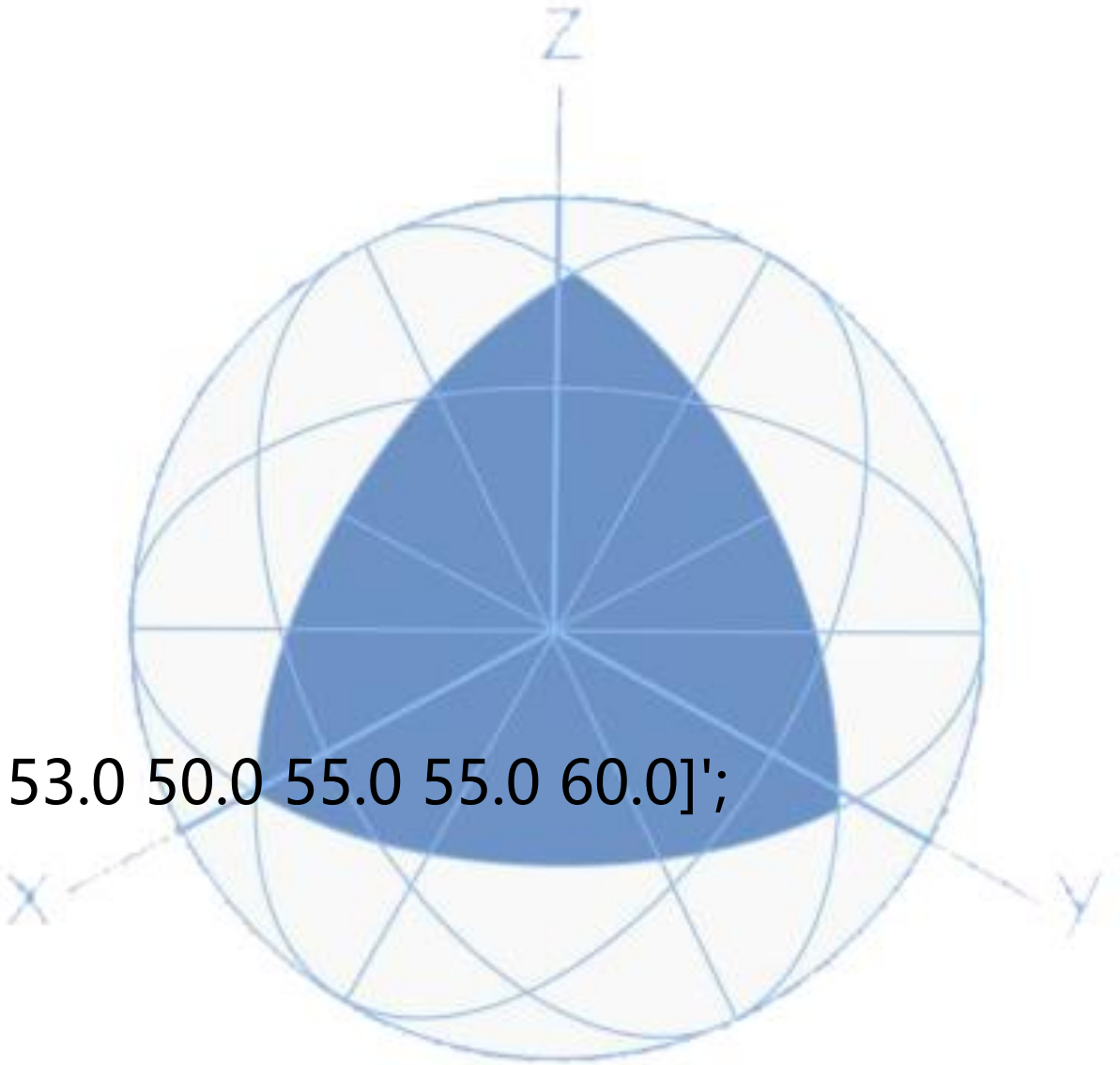
1、问题背景：

由专业知识知道，合金强度Y(N/mm2)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(xi,yi)(i=1,2,...,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

2、模型建立和参数估计

```
x1=0.1:0.01:0.18;

x2=[x1 0.20 0.21 0.23]';

x=[ones(12,1) x2];

y=[42.0 43.5 45.0 45.5 45.0 47.5 49.0 53.0 50.0 55.0 55.0 60.0]';

plot(x2',y','r*')%作散点图

[b,bint,r,rint,stats]=regress(y,x);
```



%缺省时显著性水平为 0.05

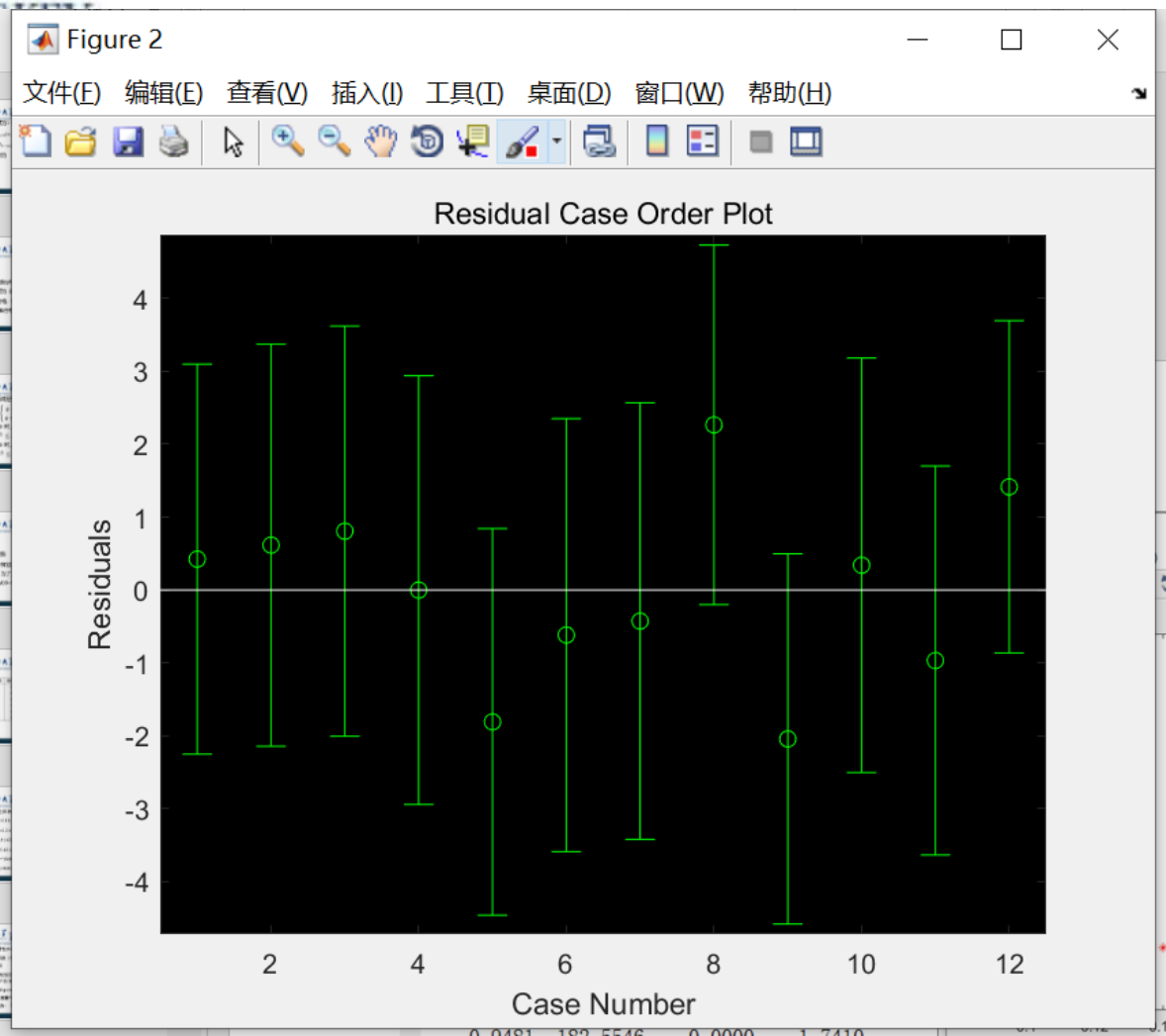
%b 为回归系数（升幂）， bint 返回回归系数的置信区间，r ，rint 为残差及其置信区间

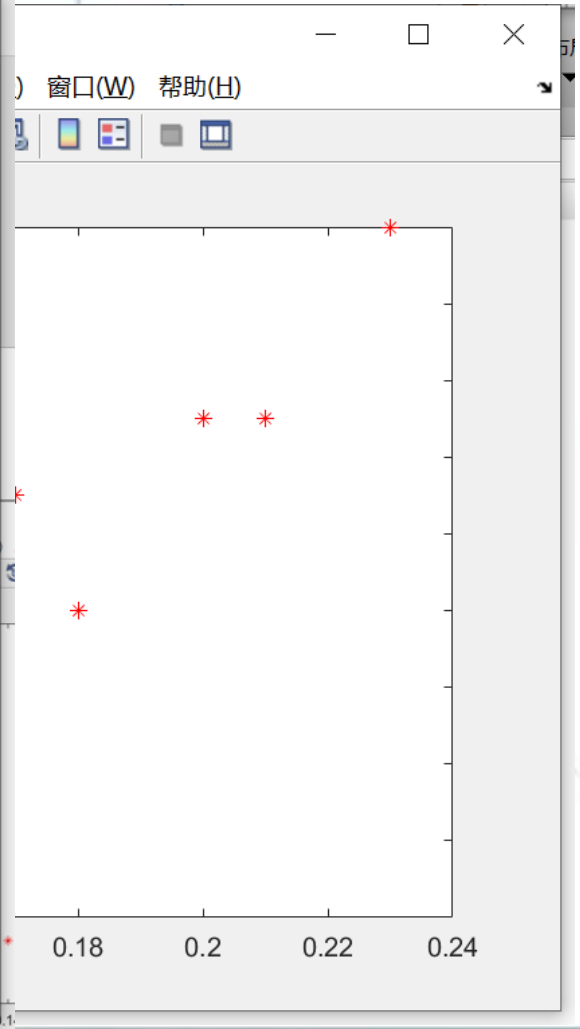
%stats 返回检验回归模型的4个统计量，分别为R的平方值，F 值，和与 F 对应的概率 P 和 残差方差， P<alpha, 时拒绝原假设，即模型是显著的。

b,bint,stats,figure(2),rcoplot(r,rint)　%rcoplot（r,rint）画出regress()拟合后的数据残差图，圆圈代表残差的值，竖线代表置信区间的范围

得到回归模型为： $\hat{y} = 28.4928 + 130.8348x$

```
b =

   28.4928
  130.8348


bint =

   24.9728   32.0
  109.2589  152.4


stats =

    0.9481  182.5
```





6.2.3 多元线性回归

1、多元线性回归模型

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

对于实际问题，设 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)\ (i = 1, \cdots, n)$

是我们获得n次独立观测值，则有

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

令

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}, \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_p \end{bmatrix},$$

所以模型等价转化为矩阵形式 $\boldsymbol{Y}=\boldsymbol{X}\beta+\varepsilon$

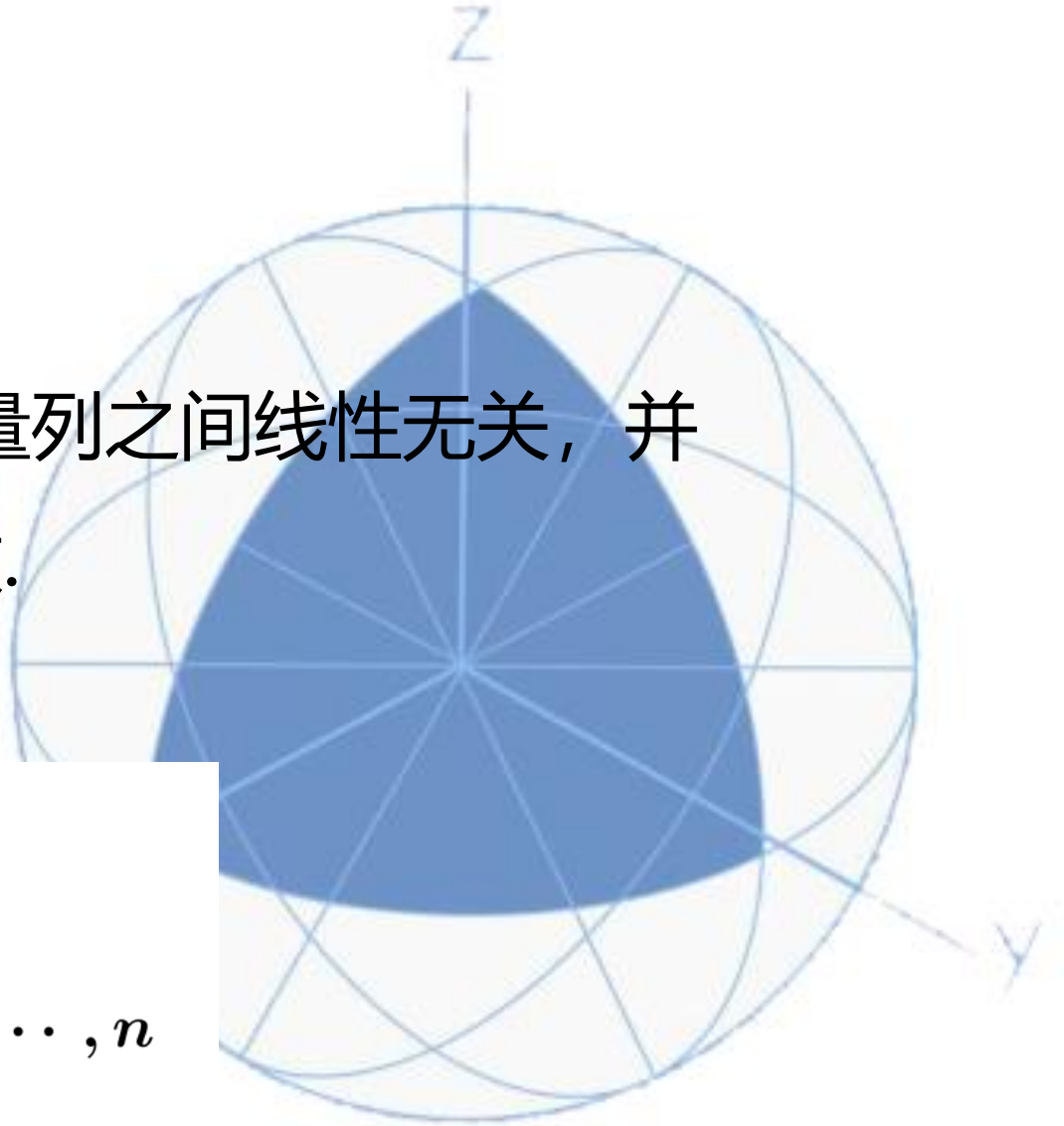
模型假设

(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:

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

这个条件称为高斯—马尔可夫条件.



(3)正态分布假定条件:

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

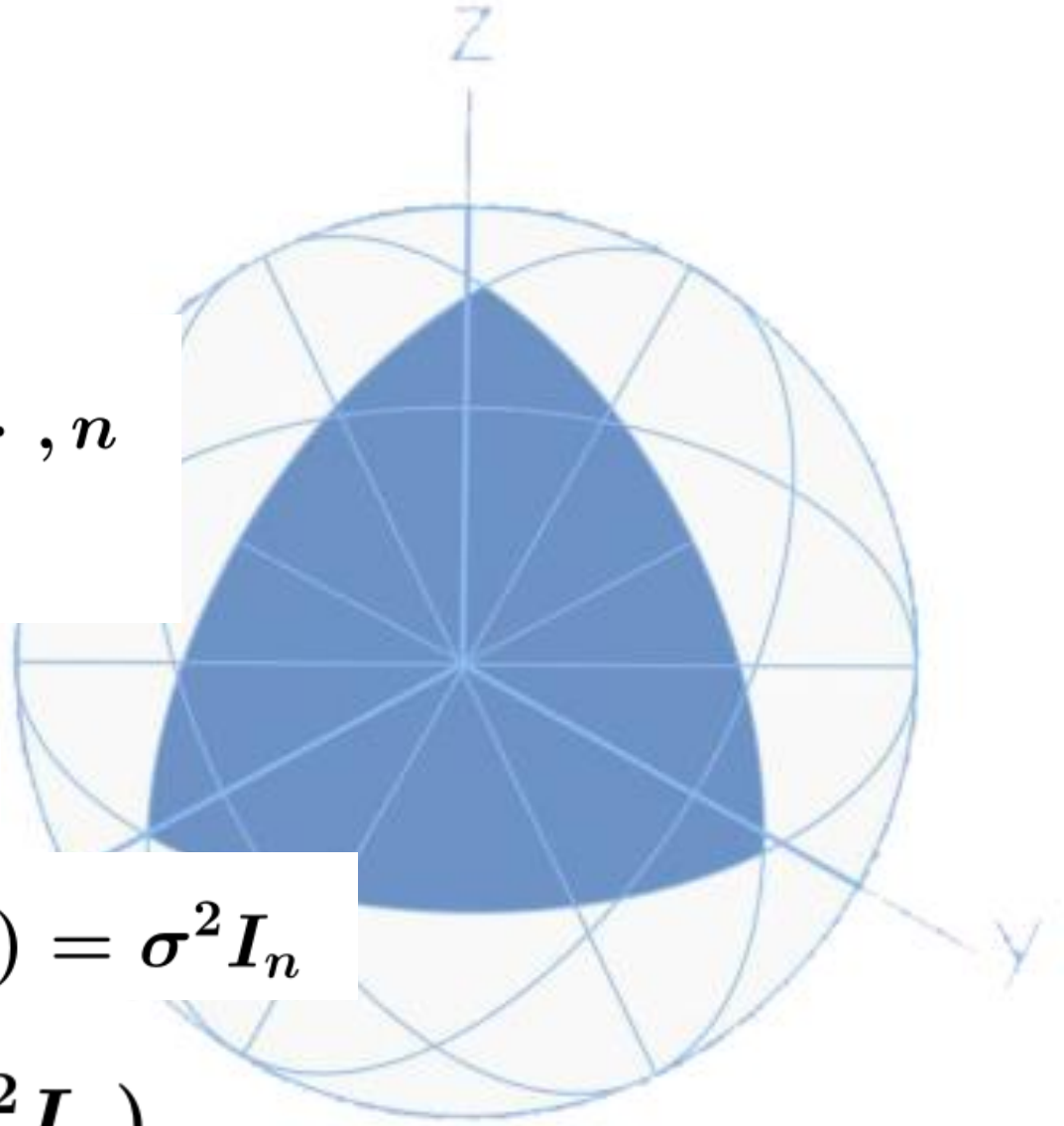
也可以用矩阵形式表示为

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

相应的我们也有 $E(Y) = X\beta, var(Y) = \sigma^2 I_n$

因此

$$Y \sim N(X\beta, \sigma^2 I_n)$$



2、参数估计

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

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

误差项方差σ^2的无偏估计为

$$\hat{\sigma^2} = \frac{1}{n-p-1}(ee')$$

其中

$$e' = (e_1, e_2, \cdots, e_n) = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, \cdots, y_n - \hat{y}_n)$$

此外

$$\begin{aligned} E(\hat{\beta}) &= E((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

$$
\begin{aligned}
D(\hat{\beta}) &= cov(\hat{\beta}, \hat{\beta}) \\
&= cov((X'X)^{-1}X'Y, (X'X)^{-1}X'Y) \\
&= (X'X)^{-1}X'\sigma^2 I_n((X'X)^{-1}X')' \\
&= \sigma^2(X'X)^{-1}
\end{aligned}
$$

3、显著性检验

（1）F检验　设原假设为

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

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			

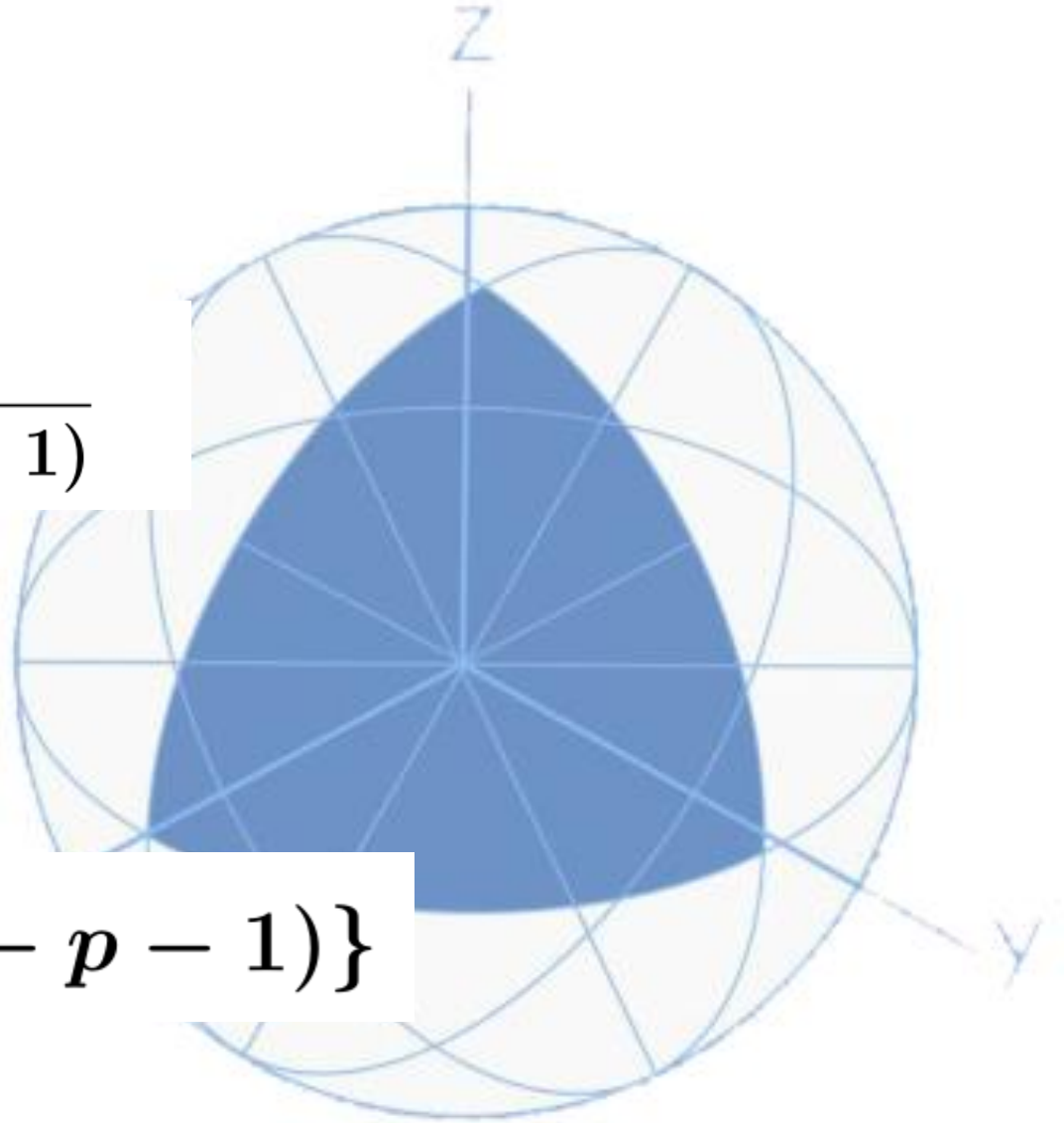
构造F统计量

$$F = \frac{SSR/p}{SSE/(n-p-1)}$$

则有 $F \sim F(p, n-p-1)$

显著水平为α时，拒绝域为

$$W = \{F > F_\alpha(p, n-p-1)\}$$



(2)t检验

为了检验某个自变量xj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

$$t_j=\frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}}$$

其中

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

$$\hat{\sigma}=\sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}e_i^2}$$

则有

$$t_j \sim t(n-p-1)$$

显著水平为α时，拒绝域为

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

同理对于回归系数的置信区间，因为我们有 $\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$，从而我们有

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1)$$

因此β_j的置信水平为$1-\alpha$的置信区间为

$$(\hat{\beta}_j - t_{1-\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma},$$
$$\hat{\beta}_j + t_{1-\alpha/2}(n-p-1)\sqrt{c_{jj}}\hat{\sigma})$$

也与一元线性回归类似，这里就不再详细介绍.

（3）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R2取值在[0,1]区间内，R2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，R2 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R2进行自由度的调整：

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

6.2.4 多重共线性

一、多重共线性的概念

设回归模型为

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

存在完全多重共线性，即设计矩阵X的列向量存在不全为零的一组数c0,c1,…,cp，使得

$$c_0 + c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip} = 0, i = 1, 2, \cdots, n$$

此时rank(X)<p+1，所以|X'X|=0，正规方程组 $\boldsymbol{X'X\hat{\beta}=X'Y}$ 的解不唯一，并且回归参数的最小二乘估计不成立

在实际问题中，我们常常遇见的是近似共线性的情形，

即存在不全为零的一组数c0,c1,...,cp，使得

$$c_0+c_1x_{i1}+c_2x_{i2}+\cdots+c_px_{ip}\approx 0, i=1,2,\cdots,n$$

我们称之为**多重共线性.**

二、多重共线性产生的原因

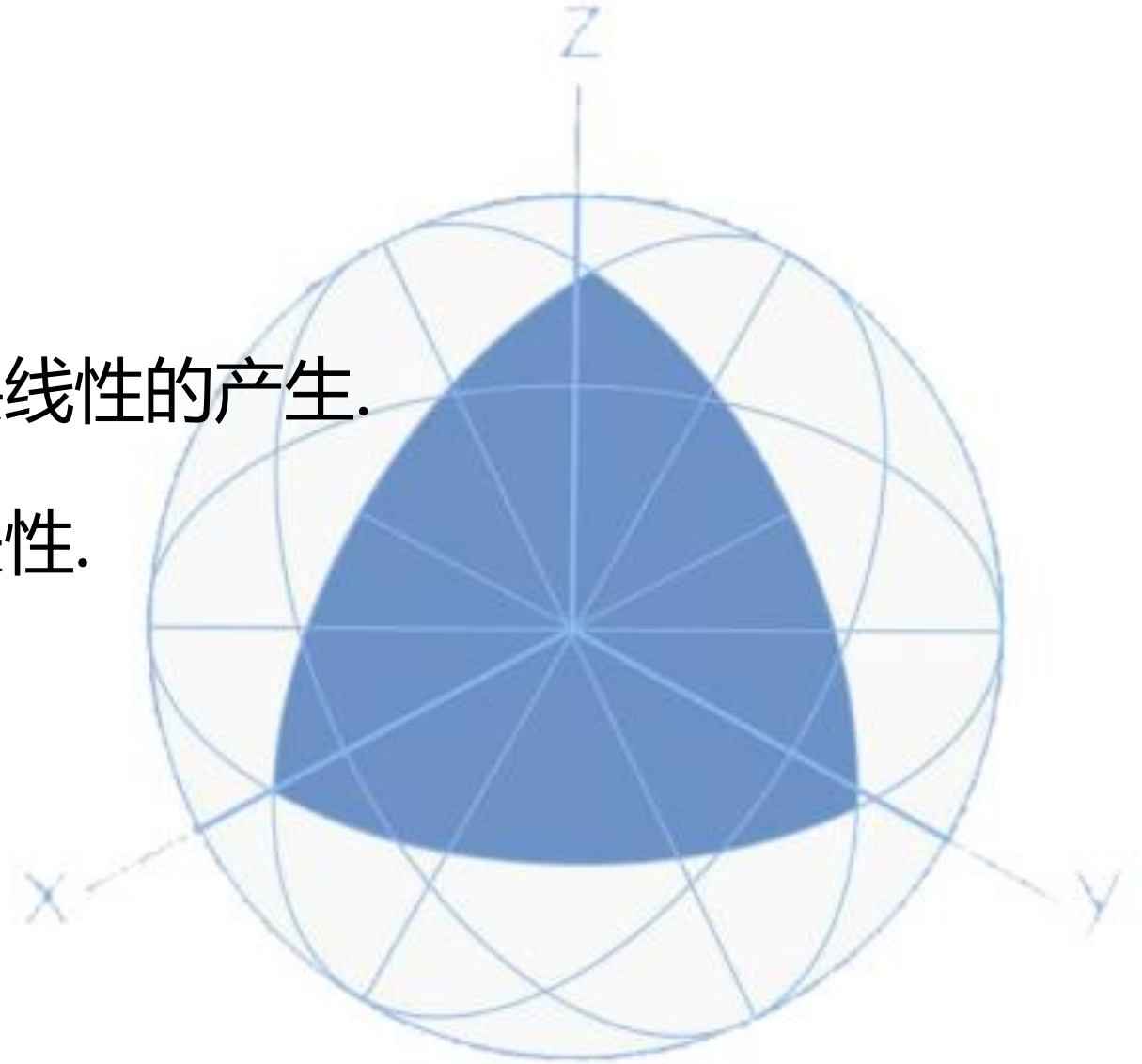
(1)变量之间的内在联系，是产生多重共线性的根本原因.

比如说影响我国居民消费的因素有职工平均工资、银行利率、零售物价指数、货币发行量和储蓄额等，而这些因素本身之间也有很强的相关性.

(2)变量之间有同方向变动的趋势，也是产生多重共线性的重要原因，如对于时间序列数据.

(3)滞后变量的引入也会导致多重共线性的产生.

(4)样本数据之间也存在一定的相关性.



三、多重共线性的诊断方法

1直观判定法：根据经验，通常以下情况的出现可能提示存在多重共线性的影响：

(1)当增加或删除一个解释变量，或者改变一个观测值时，回归参数的估计值发生较大变化，回归方程可能存在严重的多重共线性.

(2)定性分析认为，一些重要的解释变量的回归系数的标准误差较大，在回归方程中没有通过显著性检验时，可初步判断可能存在严重的多重共线性.

(3)有些解释变量的回归系数所带正负号与定性分析结果违背时，很可能存在多重共线性.

(4)解释变量的相关矩阵中，解释变量之间的相关系数较大时，可能会存在多重共线性问题.

一般而言，如果每两个解释变量的简单相关系数比较高，如大于0.8，则可认为存在着较严重的多重共线性. 但是，较高的简单相关系数只是多重共线性存在的充分条件，而不是必要条件.

特别是在多于两个解释变量的回归模型中，有时较低的简单相关系数也可能存在多重共线性. 因此并不能简单地依据相关系数进行多重共线性的准确判断.

2、条件数判定法：

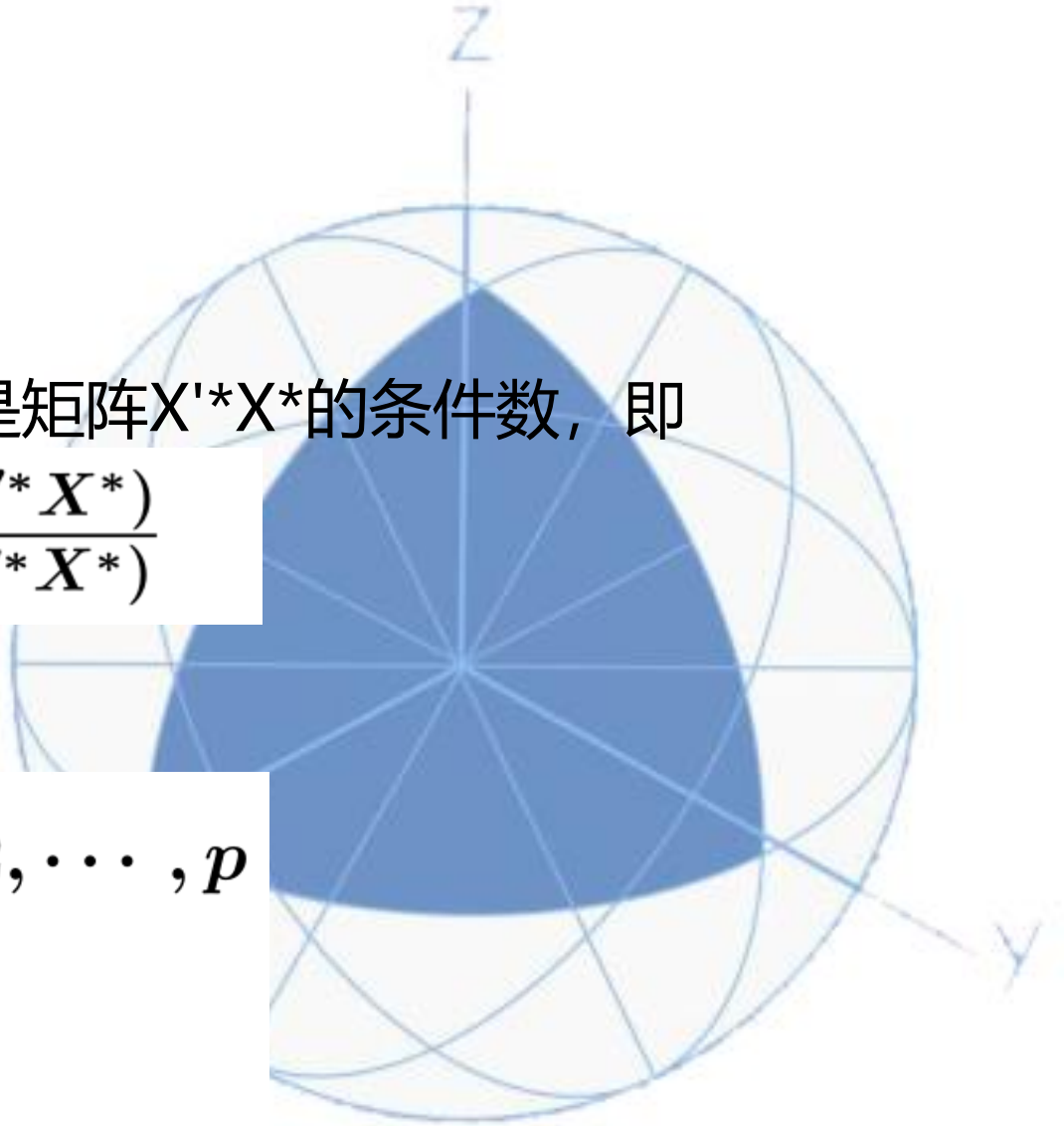
度量多重共线性严重程度的一个重要指标就是矩阵X'*X*的条件数，即

$$\kappa(X'^*X^*) = \frac{\lambda_{max}(X'^*X^*)}{\lambda_{min}(X'^*X^*)}$$

其中X*是指X经过中心标准化形成的，即

$$x_{ij}^* = \frac{x_{ij}-\bar{x_j}}{\sqrt{L_{jj}}}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, p$$

其中$L_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2$

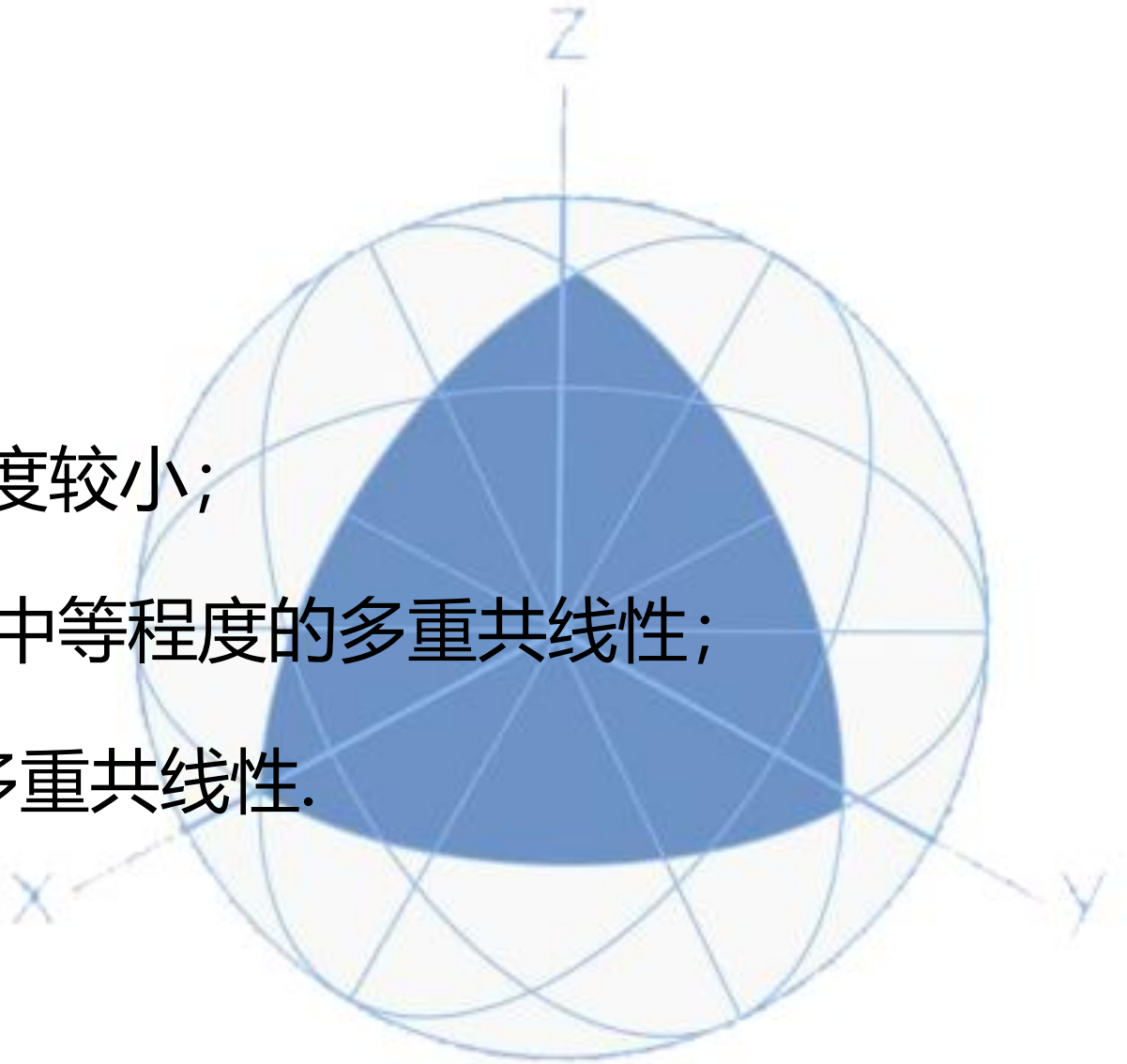


实际问题中，一般认为：

当 $\kappa < 100$ 时，多重共线性程度较小；

当 $100 \leq \kappa \leq 1000$ 时，存在中等程度的多重共线性；

当 $\kappa > 1000$ 时，存在严重的多重共线性.



6.2.5 逐步回归

在实际问题中，当我们进行自变量的选择的时候，如果忽略了一个对y有显著影响的自变量，那么建立的回归方程与实际必然有较大的偏差. 但是变量选的过多，使用起来又不方便，特别是当方程中含有对y影响不大的变量时.

常见的选择自变量的方法有前进法，后退法和逐步回归法. 但是前进法和后退法都有较为明显的缺点和不足.

前进法可能存在当引入一个变量之后其他自变量变得不显著了，但却没有进行剔除．后退法则是一开始将全部自变量引入，计算量很大．并且一旦某个自变量被剔除，他就再也没有机会重新进入回归方程．逐步回归法则吸收了两者的优点，克服了他们的不足，把它们结合起来．

具体做法是将变量一个个引入，每引入一个，就对已选入的变量进行检验，剔除不显著的自变量，这个过程要用到F 检验．

反复进行，直到既无显著的自变量选入回归方程也没有不显著的变量从回归方程中剔除.

值得注意的是引入自变量和剔除自变量的显著性水平α值不同，要求 $\alpha_{进} < \alpha_{出}$，否则可能形成“死循环”.

下面举个具体多元线性回归的例子，并做多重共线性逐步回归以及回归诊断.

案例：香港股票市场的变化

问题背景：为了研究香港股市的变化，我们以恒生指数为例，建立回归方程，分析影响股票价格趋势变动的因素. 这里的恒生指数是反应股票市场上所有上市股票价格整体水平变化的指标.

我们设

y表示因变量恒生指数；

x1表示成交额(百万港元)，用来反映市场状况；

x2表示99金价(港元/两)，从贵金属方面反映金融环境的影响；

x3表示港汇指数，从汇率方面反映金融环境的影响；

x4表示人均生产总值(港元)，用来反映整体经济情况；

x5表示建筑行业总开支(百万港元)，也用来反映整体经济状况

x6表示房地产买卖金额(百万港元)，用来反映整体经济状况；

x7表示优惠利率，用来反映金融环境的影响.

为了估计模型参数，我们收集了上述变量1974年到1988年的间共15年的数据，如下图所示：

年份	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1974	172.9	11246	681	105.9	10183	4110	11242	9
1975	352.94	10335	791	107.4	10414	3996	12693	6.5
1976	447.67	13156	607	114.4	13134	4689	16681	6
1977	404.02	6127	714	110.8	15033	6876	22131	4.75
1978	409.51	27419	911	99.4	17389	8636	31353	4.75
1979	619.71	25633	1231	91.4	21715	12339	43528	9.5
1980	1121.17	95684	2760	90.8	27075	16623	70752	10
1981	1506.94	105987	2651	86.3	31827	19937	125989	16
1982	1105.79	46230	2105	125.3	35393	24787	99468	10.5
1983	933.03	37165	3030	107.4	38823	25112	82478	10.5
1984	1008.54	48787	2810	106.6	46079	24414	54936	8.5
1985	1567.56	75808	2649	115.7	47871	22970	87135	6
1986	1960.06	123128	3031	110.1	54372	24403	129884	6.5
1987	2884.88	371406	3644	105.8	65602	30531	153044	5
1988	2556.72	198569	3690	101.6	74917	37861	215033	5.25

2、模型建立与分析

设定多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

用regress（y,X）可得回归方程为

$$\hat{y} = -723.289 + 0.003x_1 + 0.157x_2 + 6.151x_3 + 0.025x_4 - 0.045x_5 + 0.006x_6 + 16.501x_7$$

$$R^2 = 0.9904, \bar{R}^2 = 0.9808,$$
$$F = 103.409, P = 0.000002$$

整体效果的F检验通过.

3、模型修正

通过计算相关系数我们可以看出 $r_{y,3} = -0.0425, \quad r_{y,7} = -0.0955$

这说明港汇指数和优惠利率对恒生指数的影响不大. 另外通过matlab软件的条件数计算可得 $\kappa = 330.3587 > 100$ ，说明存在中等程度的多重共线性.

4、逐步回归

从上述分析中，我们首先可以剔除对y影响不大的解释变量x3, x7. 然后分别对x1,x2,x4,x5,x6 做一元回归. 其中在逐步回归法中，我们设 $\alpha_{进} = 0.05, \alpha_{出} = 0.10$

用matlab软件得到一元回归方程分别为：

$$\hat{y} = 514.103 + 0.008x_1, \bar{R}^2 = 0.8289, F = 68.843$$

$$\hat{y} = -193.951 + 0.638x_2, \bar{R}^2 = 0.7648, F = 46.532$$

$$\hat{y} = -107.176 + 0.038x_4, \bar{R}^2 = 0.8710, F = 95.547$$

$$\hat{y} = -80.079 + 0.008x_5, \bar{R}^2 = 0.7544, F = 44.011$$

$$\hat{y} = -136.863 + 0.013x_6, \bar{R}^2 = 0.8689, F = 93.818$$

从上述结果我们可以看出第一步应该选入的自变量为：x4.

同理我们在x4基础上分别在对x1,x2,x5,x6做二元线性回归，进行比较可得第二步应选入的自变量为：x4, x1.

接着继续进行逐步回归，在x4, x1的基础上分别对x2, x5, x6做三元线性回归，进行比较可得出第三步应选入的自变量为：x4, x1, x6

当继续分别对x2, x5做逐步回归时，我们发现得不出更优的结果.

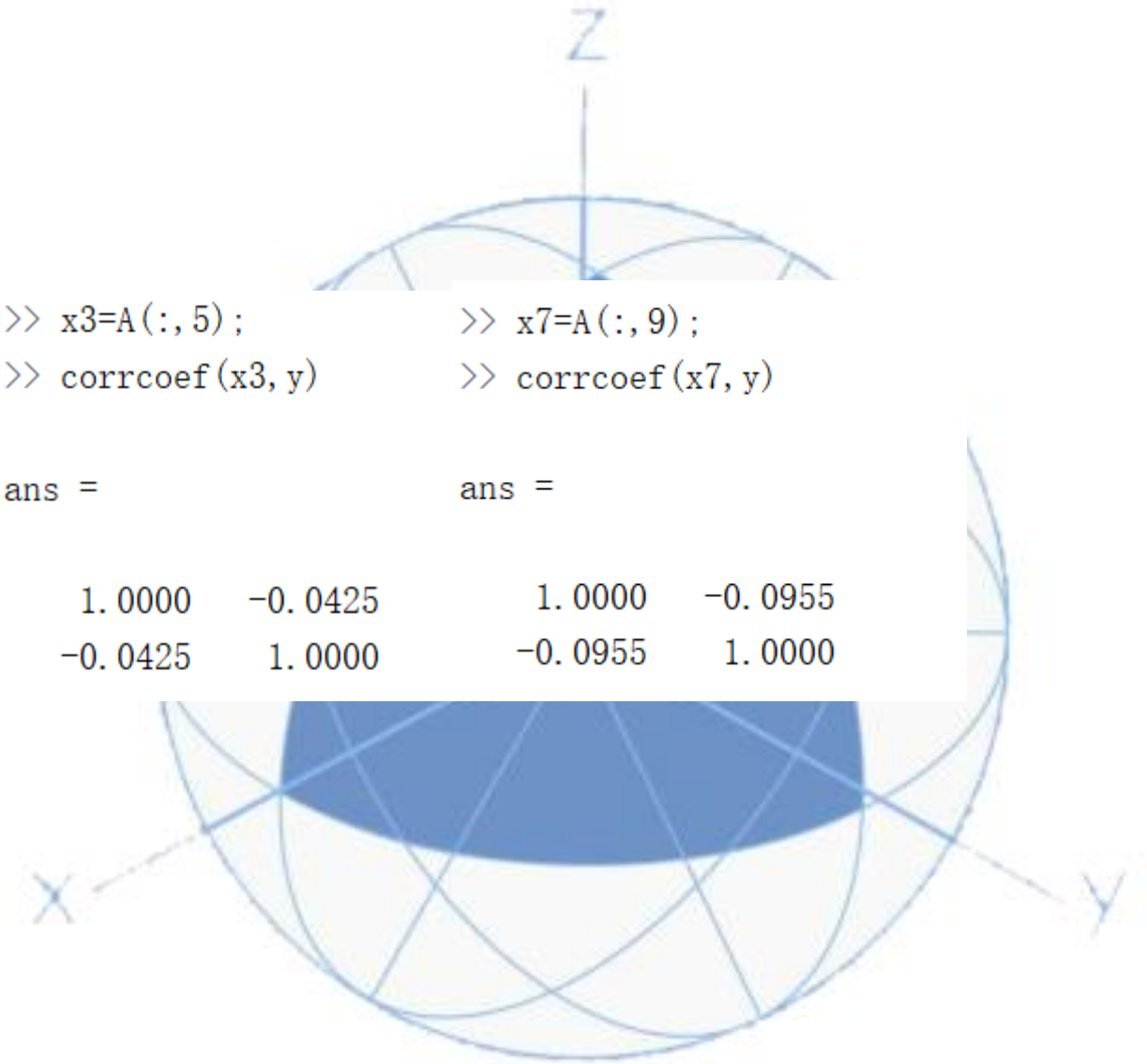
因此，综上最终修正后的回归模型为：

$$\hat{y} = 75.807 + 0.0036x_1 + 0.0129x_4 + 0.0044x_6$$

```
>> A=load('D:\\股票变化.txt')
[n,p]=size(A);
X=[ones(n,1),A(:,3:p)]
y=A(:,2);
[b,bint]=regress(y,X)
```

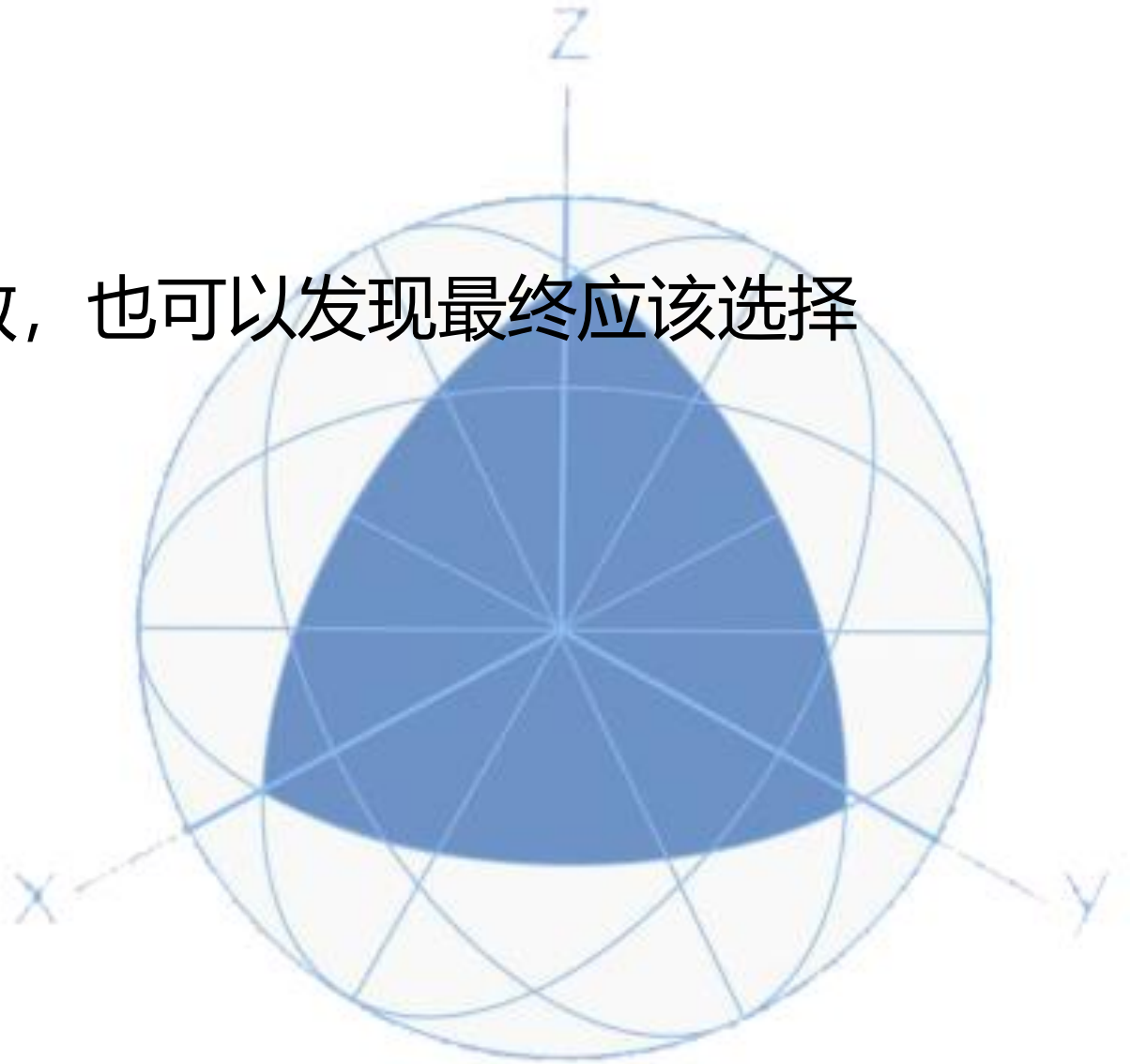
```
>> x3=A(:,5);
>> corrcoef(x3,y)

ans =

    1.0000   -0.0425
   -0.0425    1.0000
```

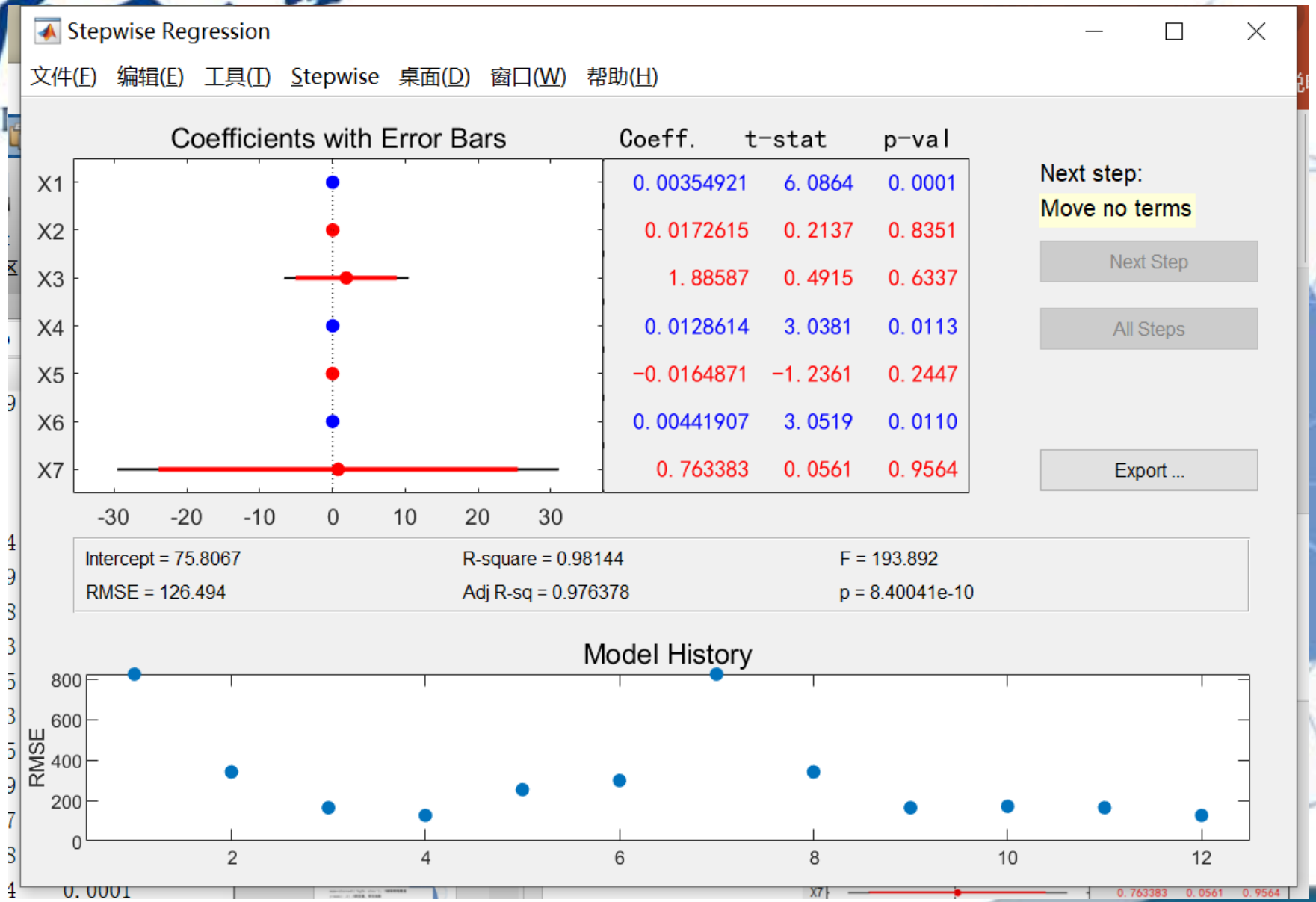
```
>> x7=A(:,9);
>> corrcoef(x7,y)

ans =

    1.0000   -0.0955
   -0.0955    1.0000
```



或者用matlab中的stepwise()函数，也可以发现最终应该选择的自变量为x1, x4, x6.

```
>> clear
>> A=load('D:\\股票变化.txt');
[n,p]=size(A);
X=A(:,3:p);
y=A(:,2);
r=corrcoef(X);
 [v,d]=eig(r);
stepwise(X,y)
```




Stepwise Regression
文件(F) 编辑(E) 工具(T) Stepwise 桌面(D) 窗口(W) 帮助(H)
Coefficients with Error Bars
X1
X2
X3
X4
X5
X6
X7
-30 -20 -10 0 10 20 30
Coeff. t-stat p-val
0.00354921 6.0864 0.0001
0.0172615 0.2137 0.8351
1.88587 0.4915 0.6337
0.0128614 3.0381 0.0113
-0.0164871 -1.2361 0.2447
0.00441907 3.0519 0.0110
0.763383 0.0561 0.9564
Next step:
Move no terms
Next Step
All Steps
Export ...
Intercept = 75.8067
R-square = 0.98144
F = 193.892
RMSE = 126.494
Adj R-sq = 0.976378
p = 8.40041e-10
Model History
RMSE
800
600
400
200
0
2 4 6 8 10 12

模型的经济学解释：

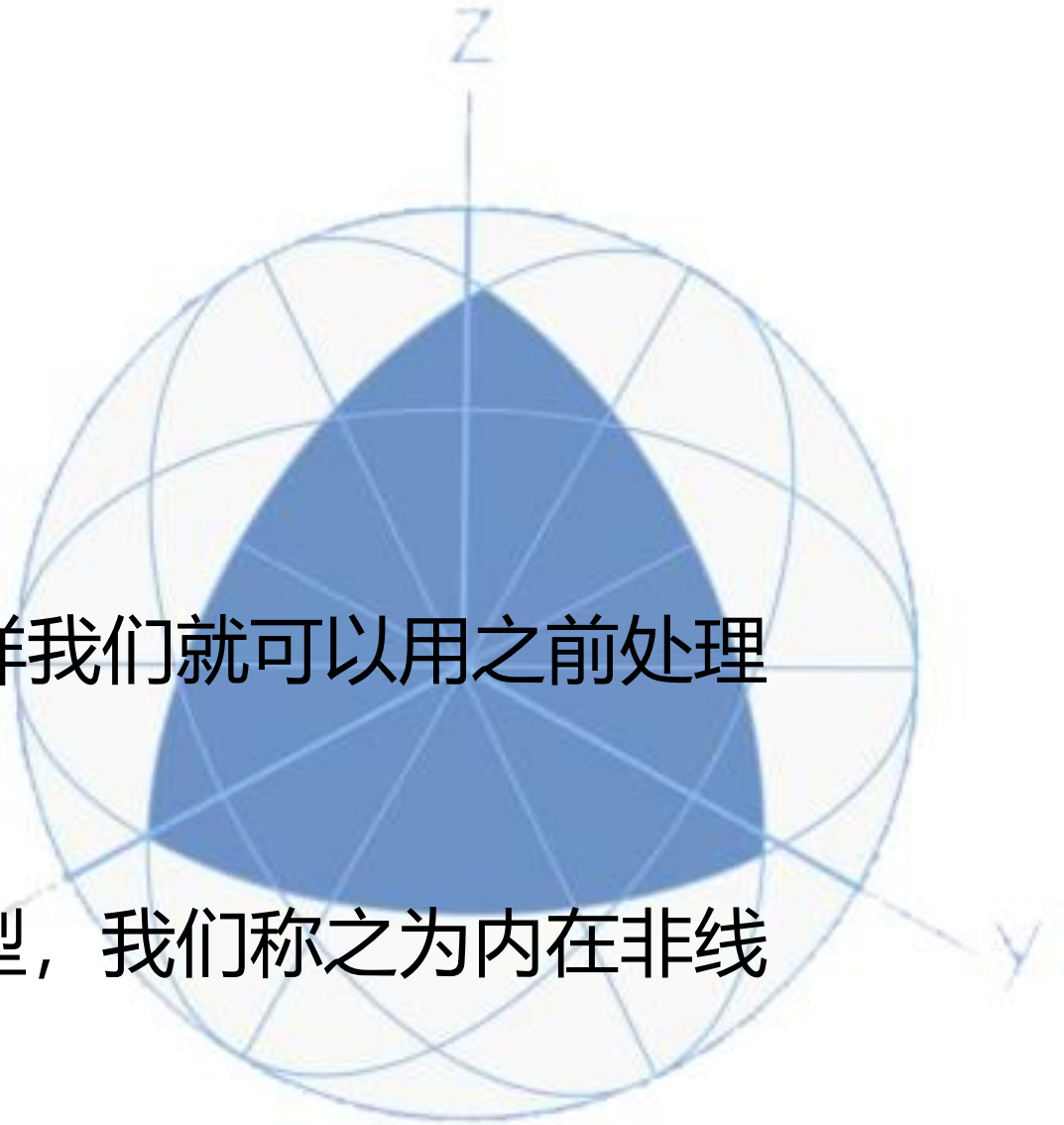
从上述回归方程我们可以看出，影响恒生指数的主要因素为成交额、人均生产总值和房地产买卖金额. 在香港股市中，成交额每增长100 万港元，恒生指数平均上涨0.0036个百分点；人均生产总值每增长100港元，恒生指数平均上涨1.29个百分点；房地产买卖金额每增加100万港元，恒生指数平均上涨0.0044个百分点.

6.2.6 非线性回归

我们往往分为两类：

一类是可以转化为线性回归模型，这样我们就可以用之前处理线性回归模型的方法；

另一类则属于不可转化为线性回归模型，我们称之为内在非线性回归模型.



1. 可转化为线性回归的曲线回归:

常见的可转化为线性回归模型的有道格拉斯生产函数以及多项式回归模型.

这里我们简单的介绍一元多项式回归模型.

设已收集到n组样本(xi, yi)，假定原模型为

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i$$

$$i = 1, 2, \cdots, n, \text{ 其中 } \varepsilon_i \sim N(0, \sigma^2)$$

现令 $z_{i1} = x_i, z_{i2} = x_i^2, \cdots, z_{ik} = x_i^k$

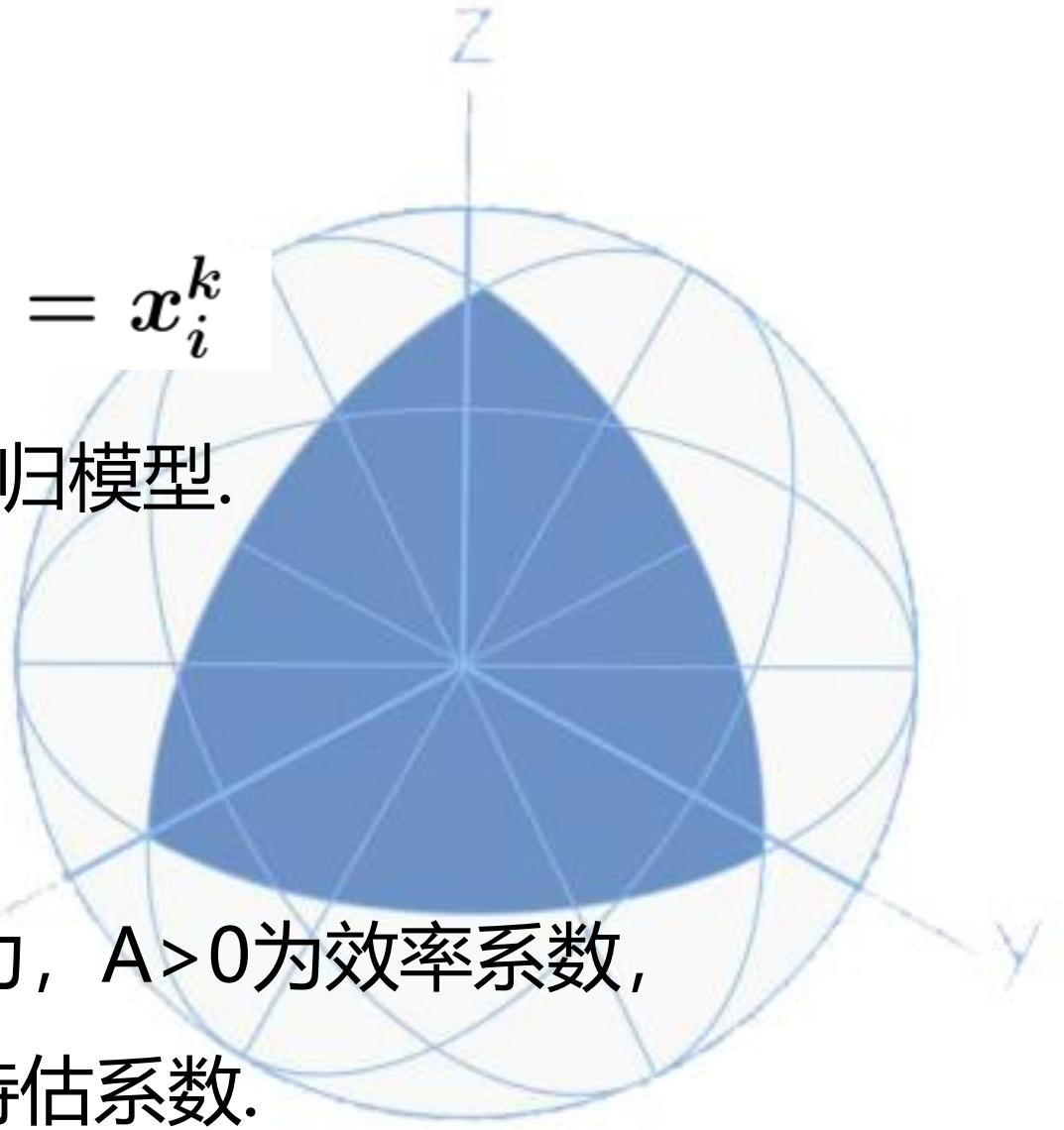
则多项式回归模型可转化为k元线性回归模型.

道格拉斯生产函数模型为:

$$y = AK^{\alpha}L^{\beta}$$

其中y为产出，K为资本，L 表示劳动力，A>0为效率系数，

α, β分别为K和L的产出弹性. A,α, β为待估系数.



现有n组样本 (yi, Ki, Li)，模型两边取对数建立回归模型有：

$$lny_i = lnA + \alpha lnK_i + \beta lnL_i + \varepsilon_i$$

令 $y_i' = y_i, \beta_0 = lnA, x_{i1} = lnK_i, x_{i2} = lnL_i,$

可转化为线性回归模型

$$y_i' = \beta_0 + \alpha x_{i1} + \beta x_{i2} + \varepsilon_i$$

2. 内在非线性模型:

设 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i), i = 1, 2, \cdots, n$ 是(x1,x2,...,xp;y)的n次独立观测值，则多元非线性回归模型可以表示为

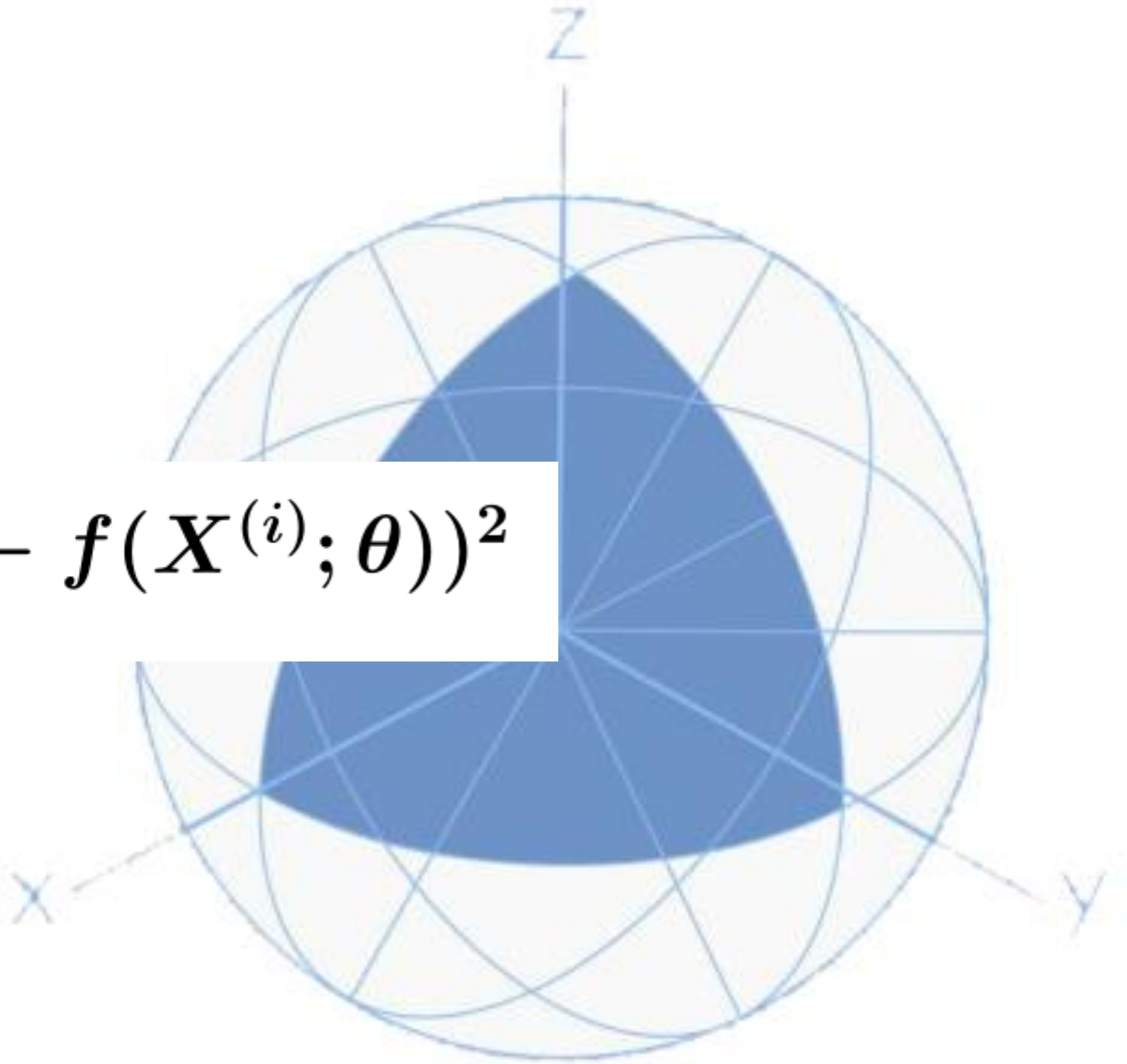
$$y_i = f(x_{i1}, x_{i2}, \cdots, x_{ip}, \theta_1, \theta_2, \cdots, \theta_k) + \varepsilon_i,$$

其中 $\varepsilon_i \sim N(0, \sigma^2)$， 并且独立同分布. 记

$$X^{(i)} = (x_{i1}, x_{i2}, \cdots, x_{ip})^T, \theta = (\theta_1, \theta_2, \cdots, \theta_k)^T$$

其最小二乘估计为

$$minQ(\theta) = \sum_{i=1}^{n}(y_i - f(X^{(i)};\theta))^2$$



厦门大学
XIAMEN UNIVERSITY

Part 3

方差分析思想与案例分析

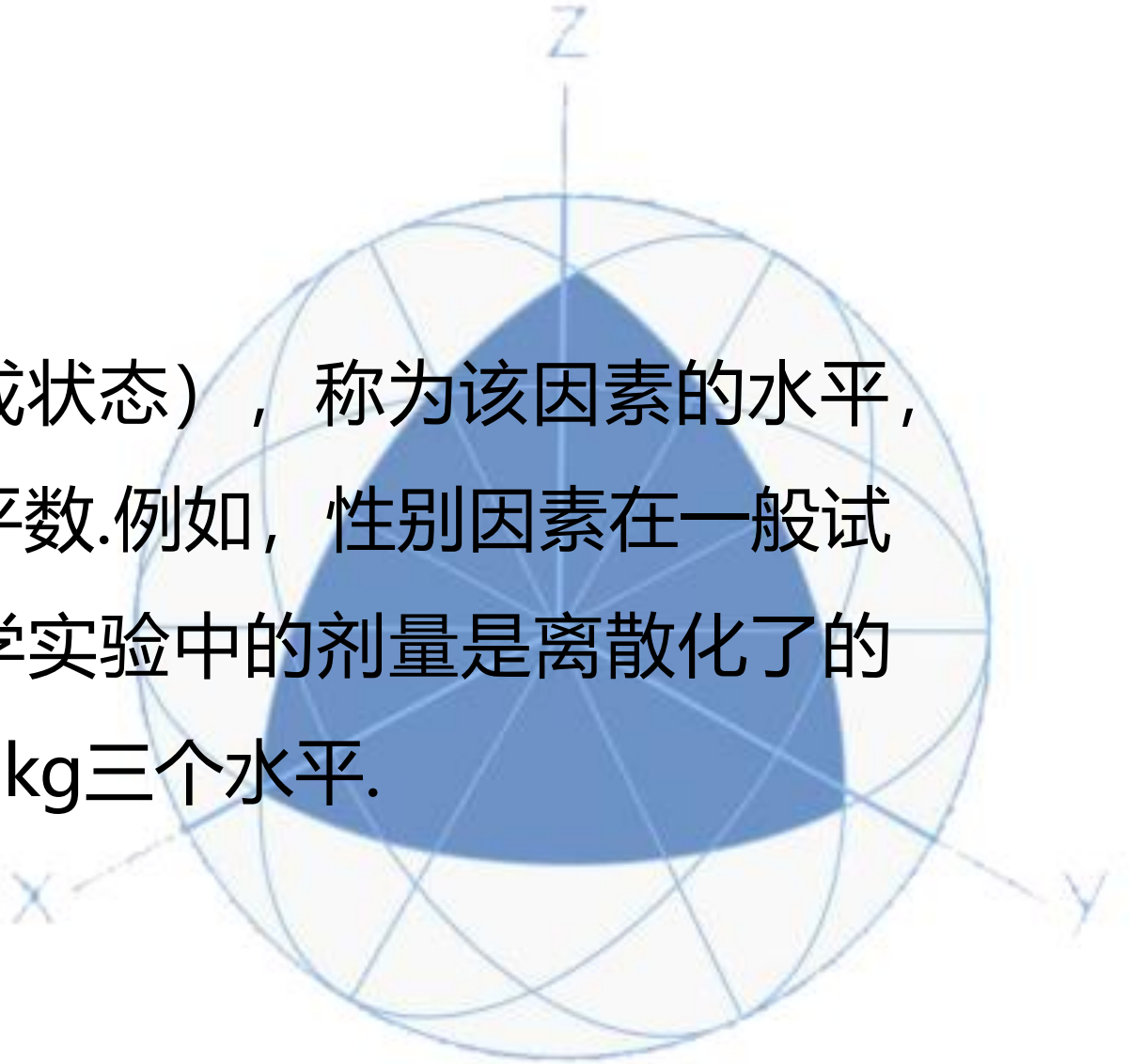
一、基本概念

(1)因素:

方差分析中的因素是指，从对实验产生影响较多的众多因素中挑选出的，可通过不同的条件或处理以供考察的因素.方差分析中因素不包括测不到的数值的因素和试验中不可控制或难以预测的因素.

(2) 水平与水平数:

因素在实验中所取的不同条件（或状态），称为该因素的水平，水平也叫位级.水平的个数称为水平数.例如，性别因素在一般试验下研究两个水平：男和女；化学实验中的剂量是离散化了的几个有限水平，如：1kg，2kg，3kg三个水平.

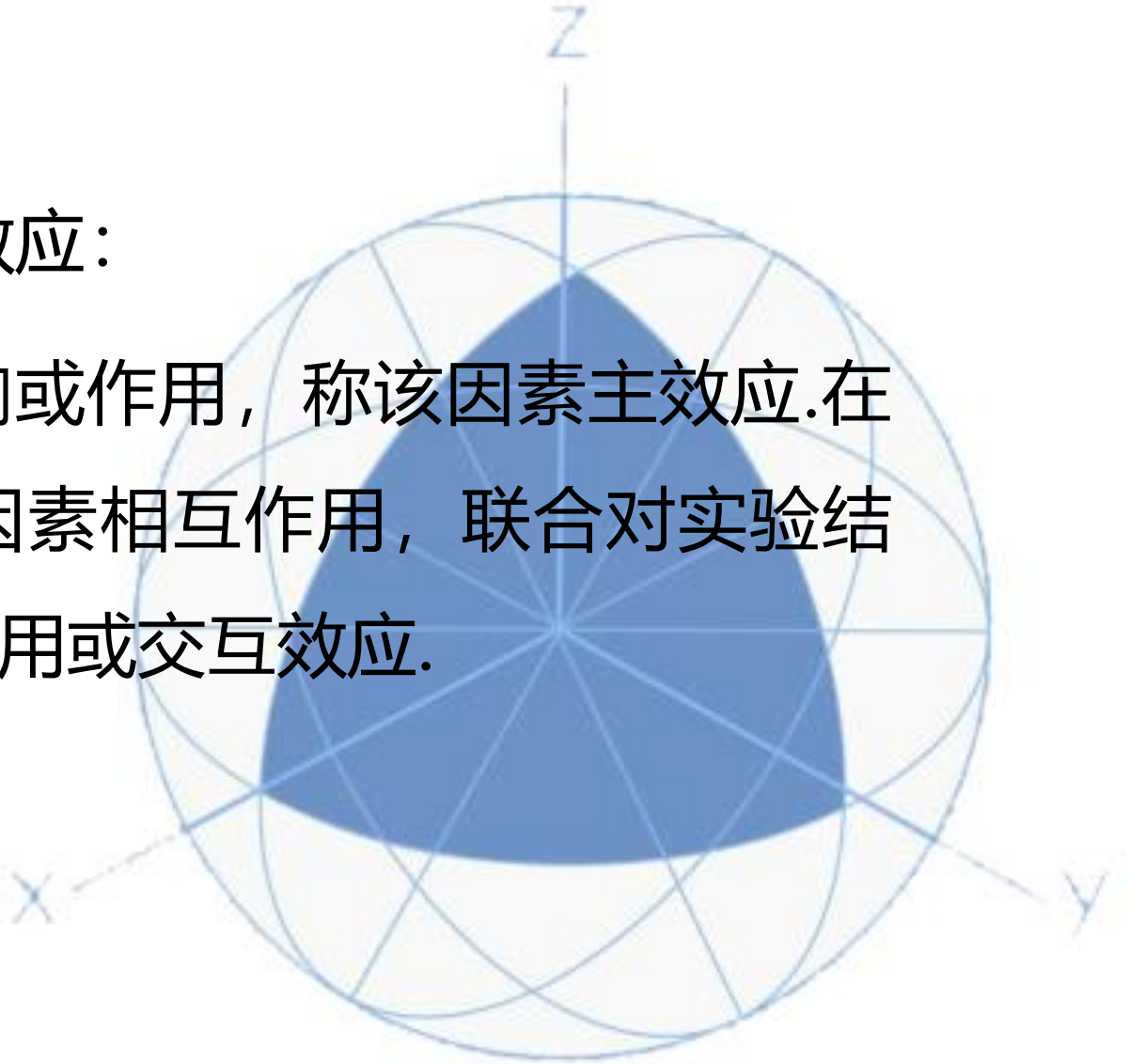


(3) 重复与重复数:

在相同条件下进行2次或者2次以上的实验，称为重复试验或者有重复试验.相同条件下重复试验的次数简称重复数.在实验的各因素、各水平下均做同样多次试验，称为等重复试验.

(4)因素的主效应与因素间的交互效应：

某因素单独对实验结果产生的影响或作用，称该因素主效应.在多因素试验中，两个或两个以上因素相互作用，联合对实验结果产生的影响或作用，称为交互作用或交互效应.



6.3.1单因素方差分析

单因素方差分析则是仅仅讨论一种实验因素对实验结果有无显著影响的分析.在许多因素中，调查某一影响最大的特定因素的效果时，或在诸多因素的分析已有结果或者进展，现在需要的是调查剩下诸因素中影响最大的因素时，一般运用单因素方差分析.

一、 数据描述

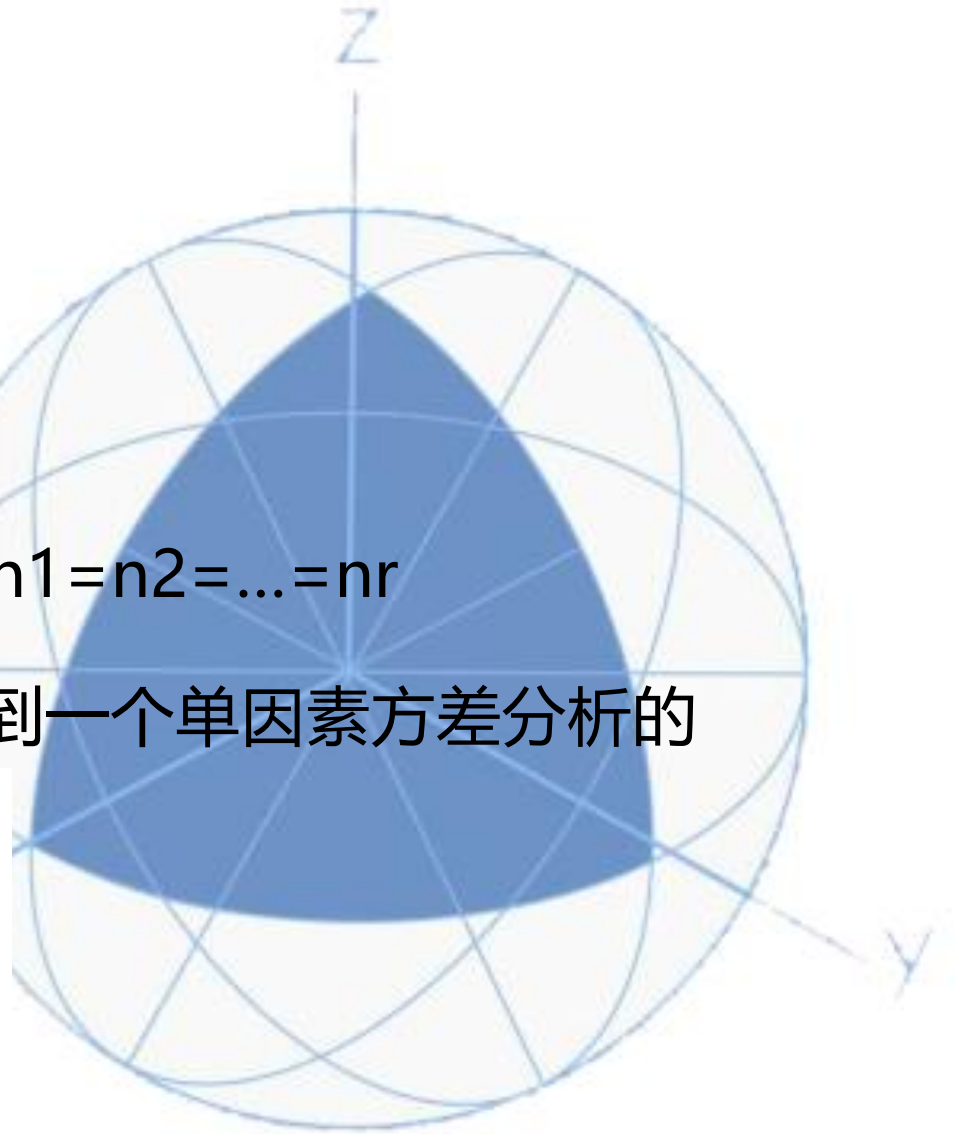
设因素为A，共有r个水平：A1,A2,...,Ar,

在水平Ai下做ni次重复试验，ni可以不全相等. 当n1=n2=...=nr

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

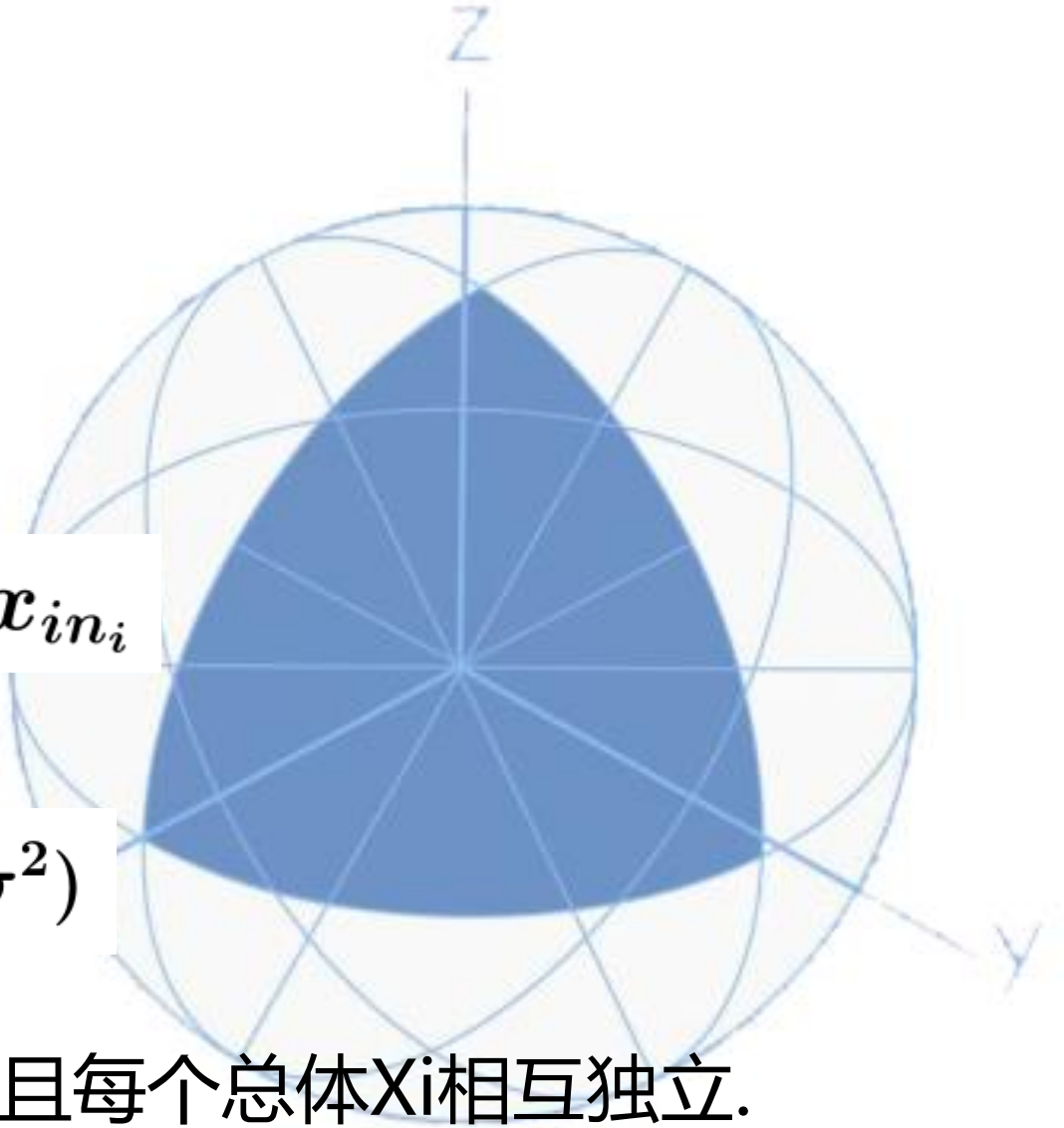
将水平Ai下的实验结果

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

看作来自第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Xi相互独立.



考虑线性统计模型

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

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

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

这里μ表示总平均，α_i表示水平Ai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

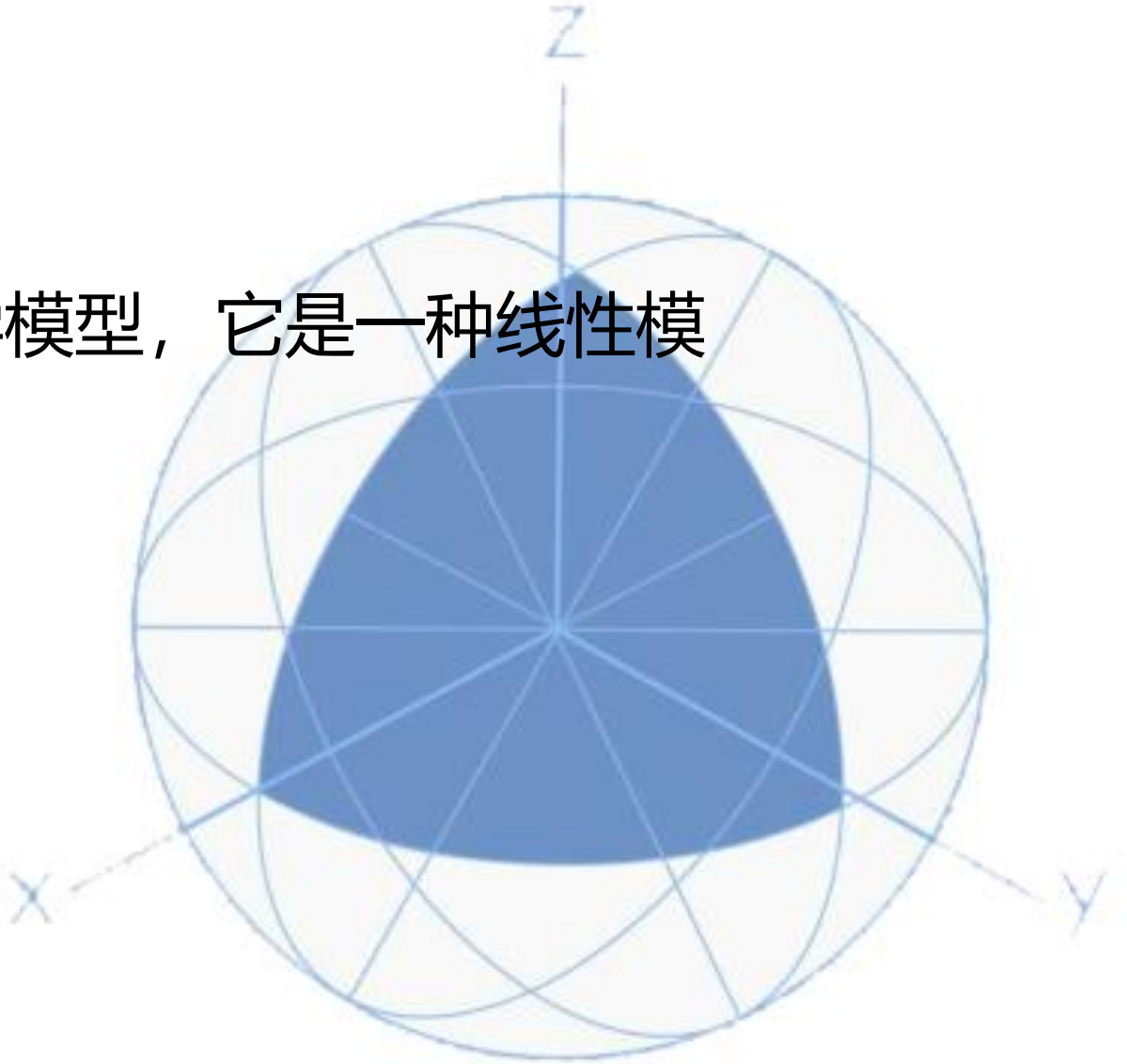
$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \ vs$$

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

统计模型等价为

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

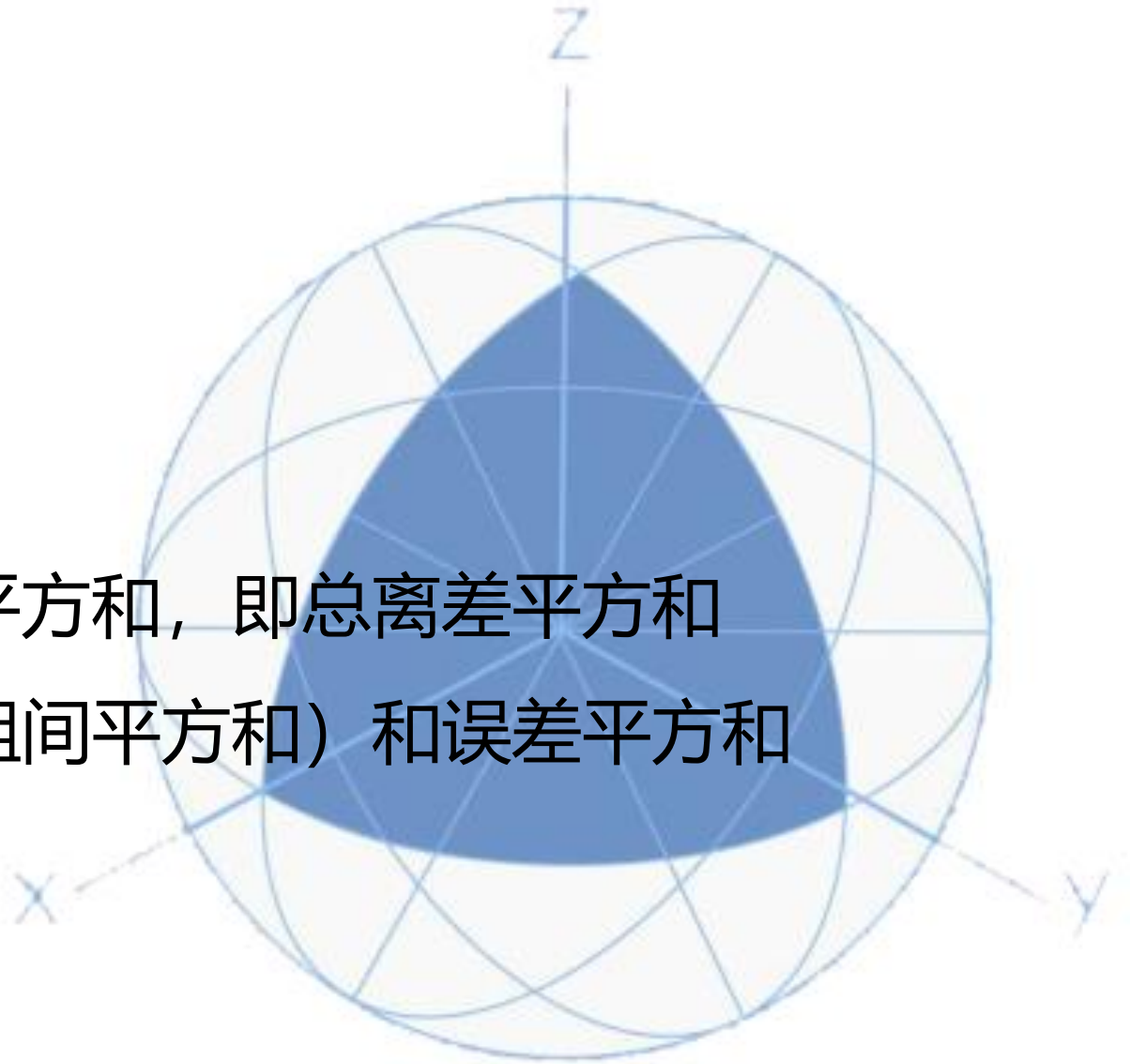
此模型称为单因素方差分析的数学模型，它是一种线性模型.



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

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

误差平方和描述的是随机误差的影响.

效应平方和

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

反映的是r个总体之间的差异.

另外，我们很容易通过计算得到以下平方和分解式

$$ST=SE+SA$$

由偏差平方和的构造可以看出，利用偏差平方和来作为数据变异性的一个度量，直观来看很合理.但在同样的波动程度下，测定数据越多，计算出的偏差平方和就越大.因此，仅用偏差平方和来反映数据的各种变异显然不够，还应当考虑测定值个数对偏差平方和的贡献，这边是我们要说的相对偏差平方和.

2、自由度

通常，随机变数的自由度由数据个数n及数据所受的线性约束方程个数m所决定. 即当n个随机变数x1,x2,...,xn受到且仅受到m个独立方程的约束时，则这n个数据的平方和自由度为v=n-m.

由上述定义，我们可以得到三个偏差平方和自由度

$$vT=n-1,\quad vA=r-1,\quad vE=n-r$$

且显然满足关系式vT=vA+vE. 这称为自由度的可分解性

而

$$\text{相对偏差平方和} = \frac{\text{偏差平方和}}{\text{偏差平方和自由度}}$$

记相对总离差平方和为VT，相对因素A效应平方和为VA，相对误差平方和为VE，则有

$$V_T = \frac{S_T}{v_T}, V_A = \frac{S_A}{v_A}, V_E = \frac{S_E}{v_E}$$

VA，VE分别称为因素A均方差和误差均方差，简称均方或方差.

3、统计量

由统计量的无偏估计相关内容，容易得到，在原假设H0下，

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

均是σ^2的无偏估计，即VA, VE均是σ^2的无偏估计.

且 $S_A/\sigma^2 \sim \chi^2(r-1)$, $S_E/\sigma^2 \sim \chi^2(n-r)$.

SA，SE相互独立，由F分布的定义，可构造H0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H0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VA比误差均方差VE大的很多，即F值比“1”大的很多，则与原假设H0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H0，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			

例：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，你能从这些数据推断他们的生产率有无显著差异吗?

	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
平均产量	269	292.25	264.75	280.5	240

运用Matlab编程，用anova1做单因素方差分析，求解得到p=0.1109>α=0.05，故接受H0，即五名工人的生产率没有显著差异. matlab程序如下：

```
s=[256    254    250    248    236
   242    330    277    280    252
   280    290    230    305    220
   298    295    302    289    252];
p=anova1(s)
```

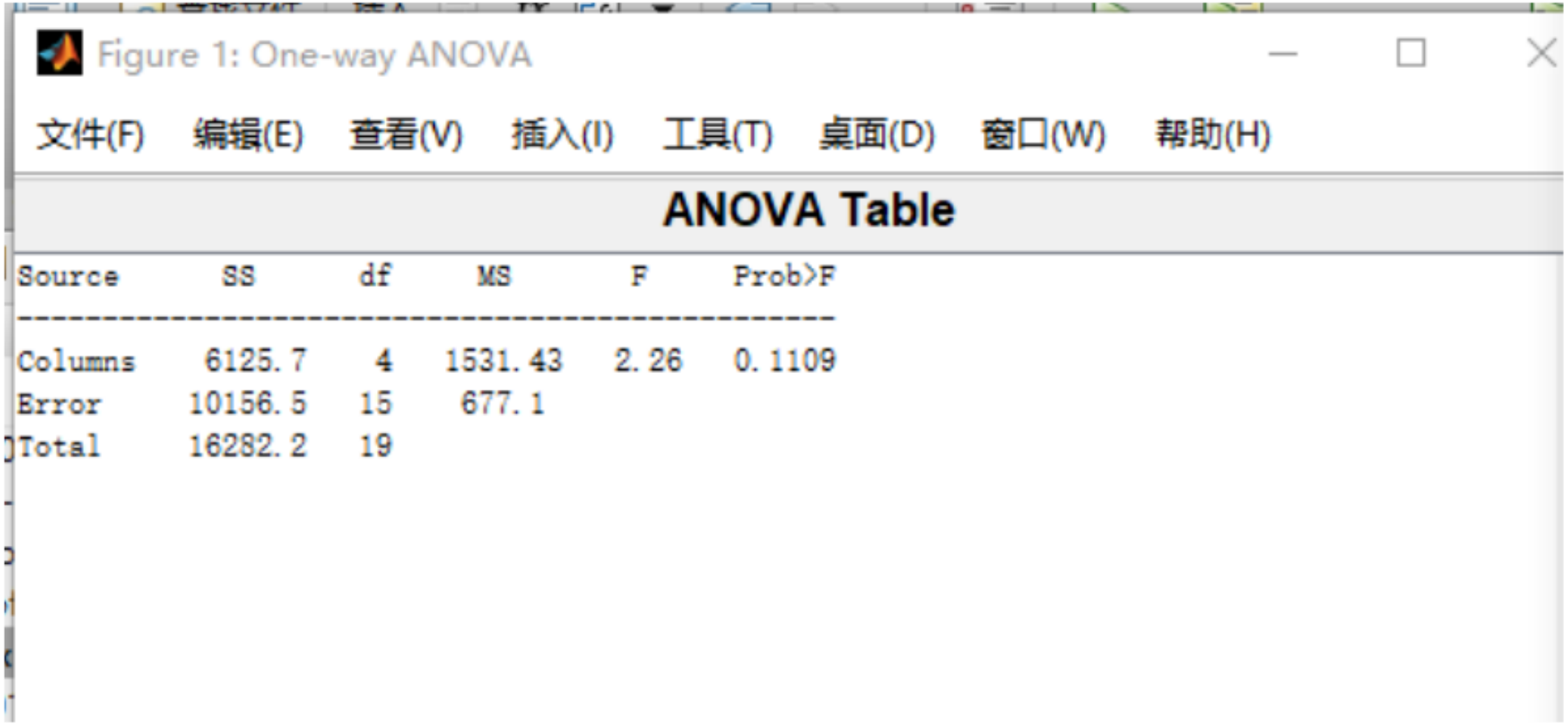


Figure 1: One-way ANOVA
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
ANOVA Table
Source SS df MS F Prob>F
Columns 6125.7 4 1531.43 2.26 0.1109
Error 10156.5 15 677.1
Total 16282.2 19

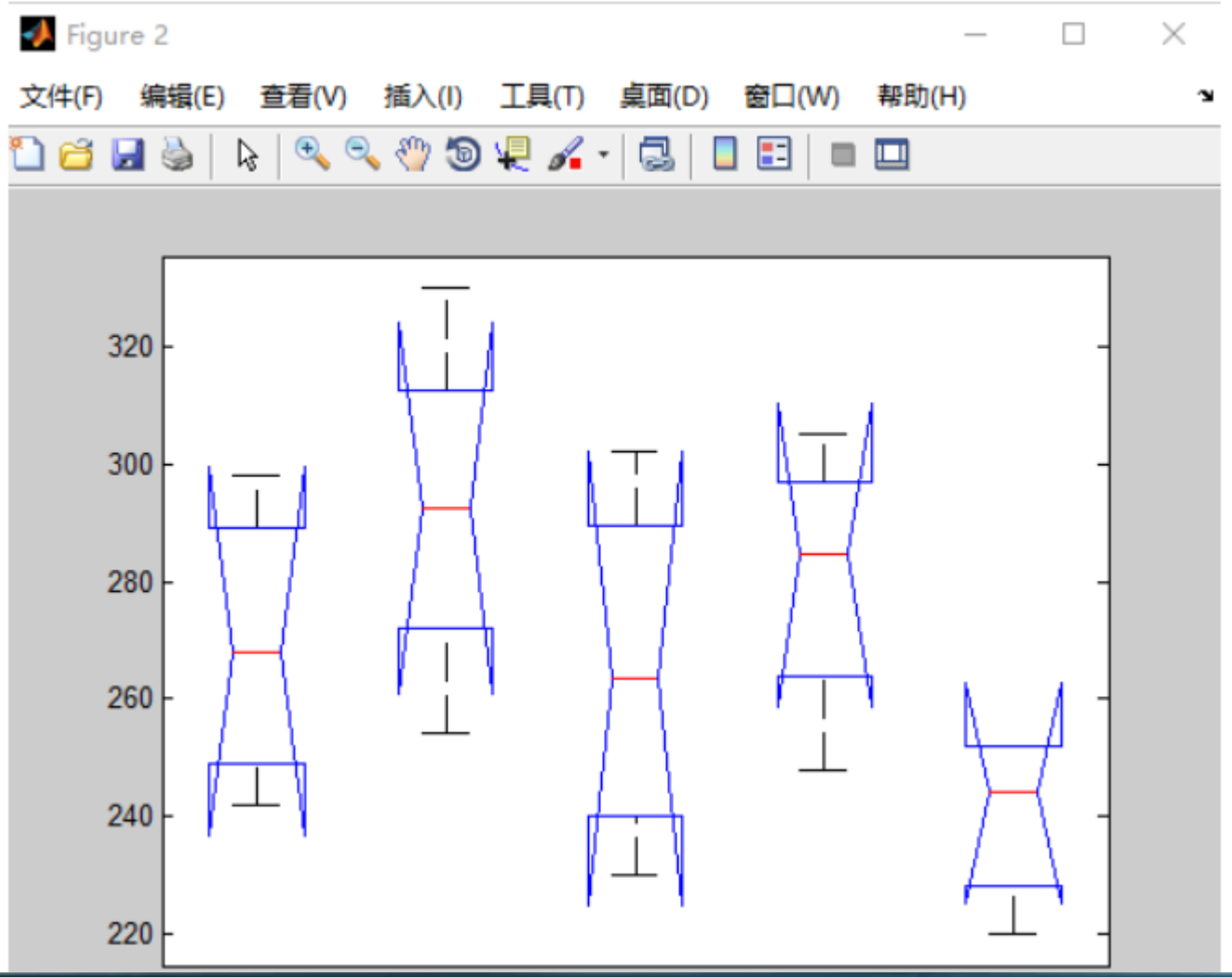
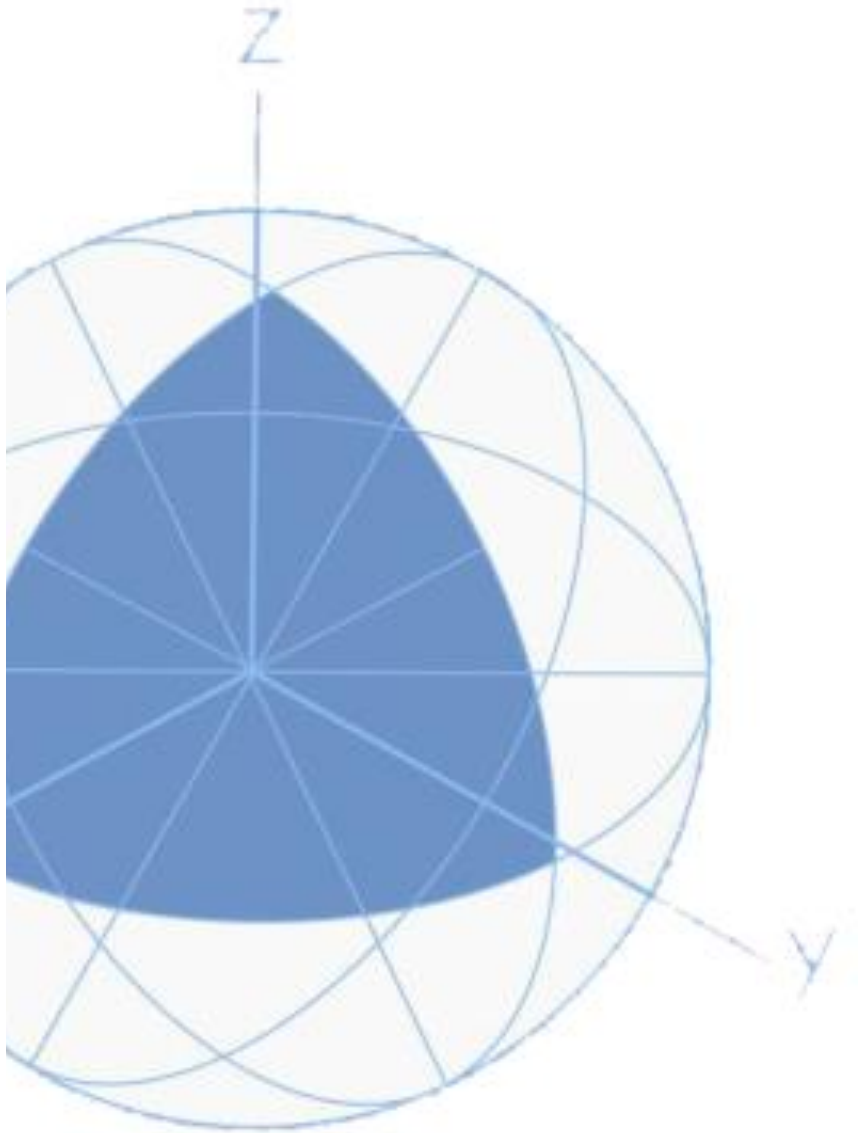


Figure 2
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
320
300
280
260
240
220


Z
Y

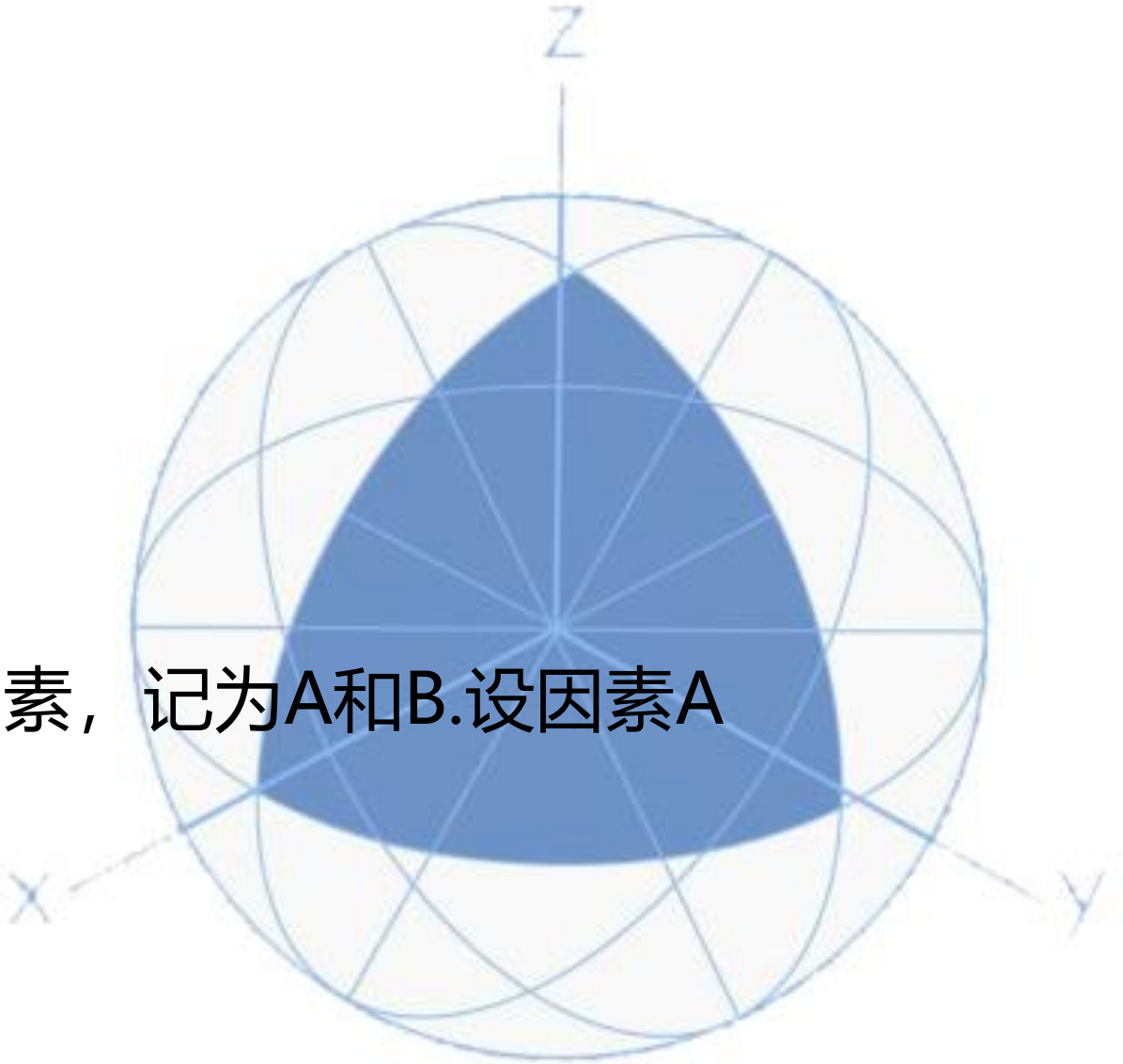
6.3.2 双因素方差分析

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：A1,A2,...,Ar,

因素B有s个水平：B1,B2,...，Bs，



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.

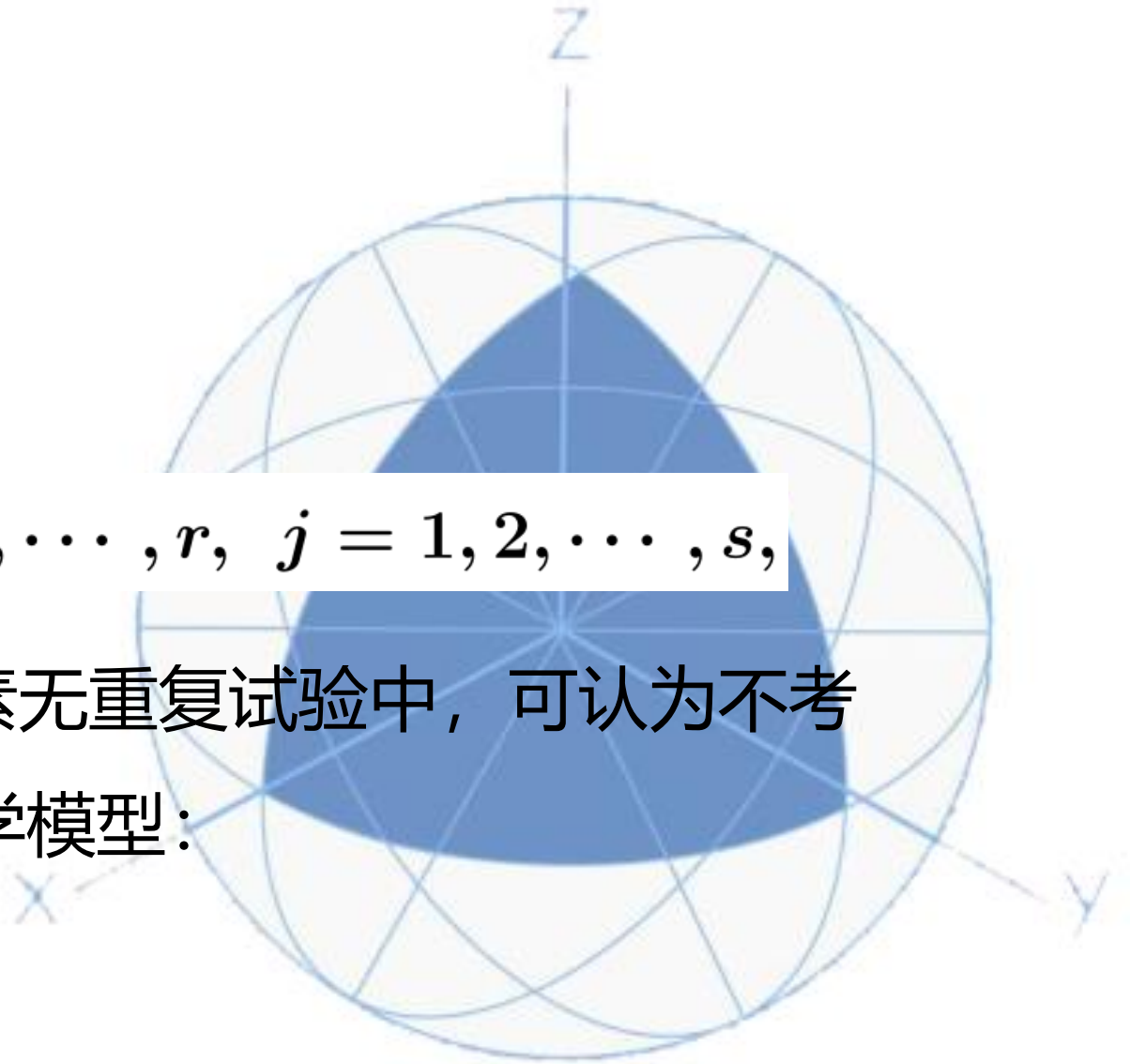
	B_1	B_2	$\cdots$	B_s
A_1	x_{11}	x_{12}	$\cdots$	x_{1s}
A_2	x_{21}	x_{22}	$\cdots$	x_{2s}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	x_{r1}	x_{r2}	$\cdots$	x_{rs}

2、方差分析

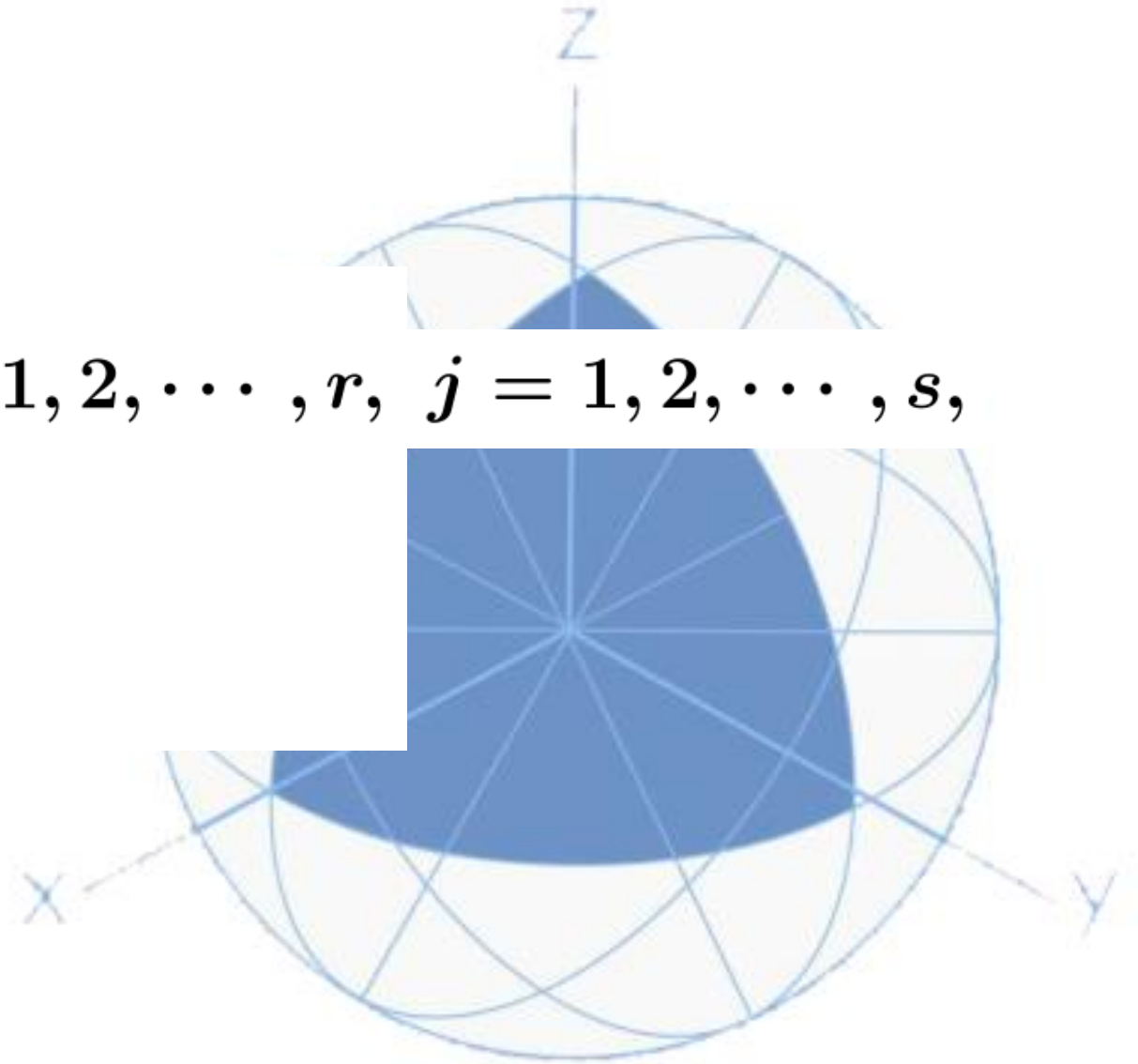
假定

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

且各xij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：



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

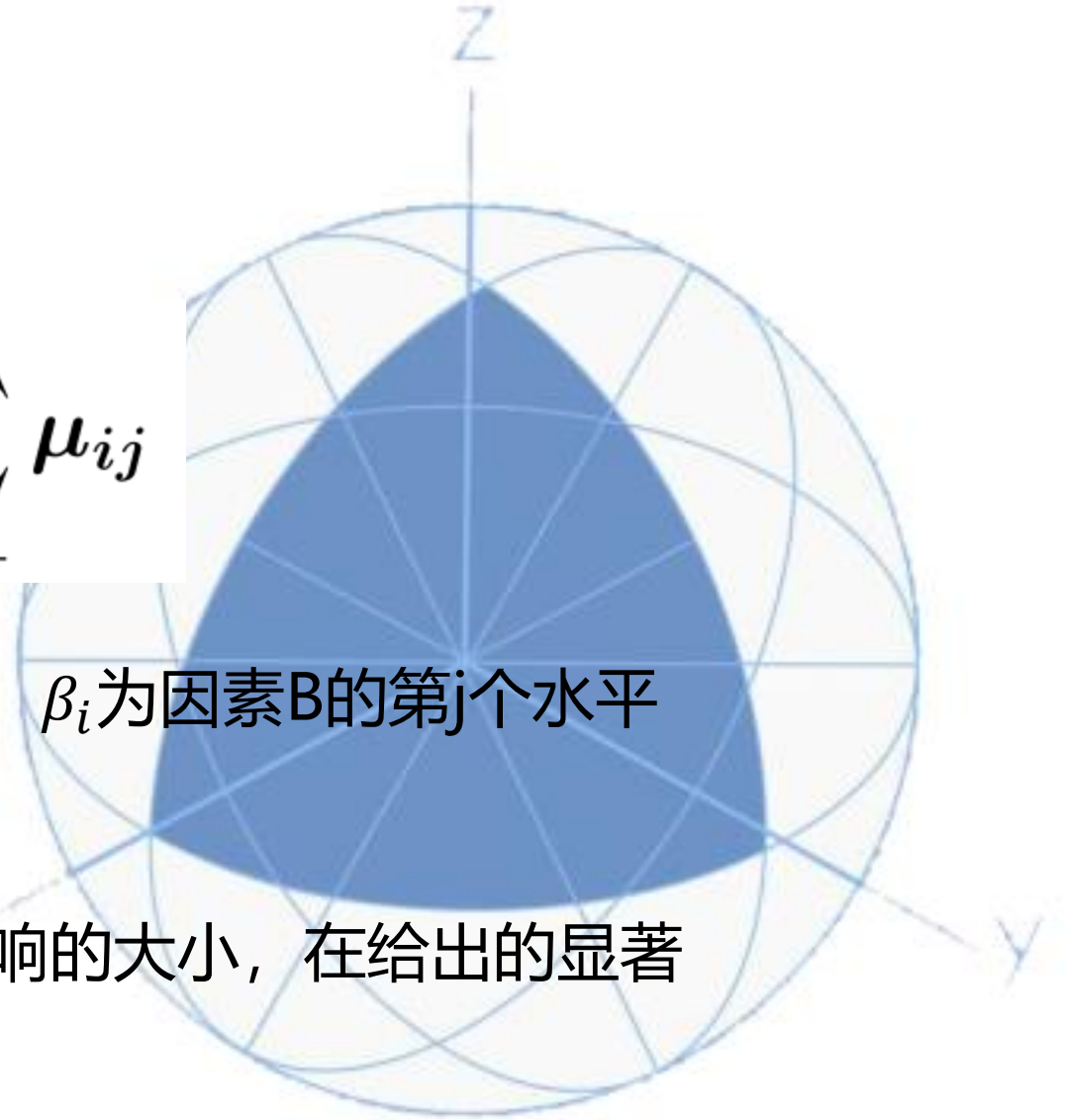


其中

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

为总平均，α_i为因素A的第i个水平的效应，β_i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：

ST=SE+SA+SB

ST为总离差平方和，SA为因素A效应平方和，SB为因素B效应平方和，SE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，SA与SE相互独立，SB与SE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

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

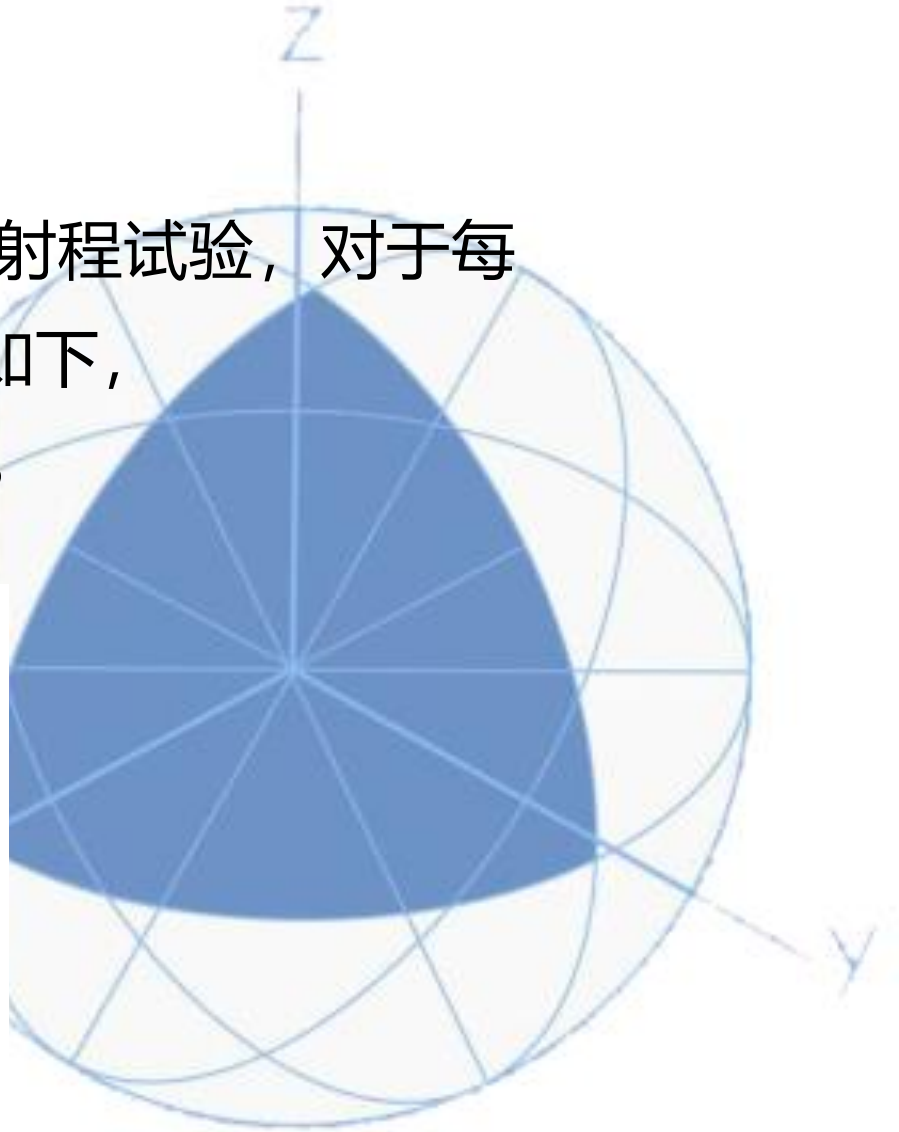
以FA、FB分别作为H01、H02的检验统计量，将结果列成方差分析表：

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			

例：一种火箭使用了四种燃料，三种推进器，进行射程试验，对于每种燃料与每种推进器的组合作一次试验，得到数据如下，

问各种燃料之间以及各种推进器之间有无显著差异?

	B_1	B_2	B_3
A_1	58.2	56.2	65.3
A_2	49.1	54.1	51.6
A_3	60.1	70.9	39.2
A_4	75.8	58.2	48.7



记燃料为因素A，它有4个水平，推进器为因素B，它有3个水平，我们在显著性水平$\alpha = 0.05$下检验

$$H_{01}: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0,$$

$$H_{02}: \beta_1 = \beta_2 = \beta_3 = 0.$$

利用Matlab实用统计工具箱中的anova2,程序代码如下：

```
w=[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(w,1)
```

二、双因素等重复试验方差分析

设因素A有r个水平：A1,A2,...,Ar, 因素B有s个水平：B1,B2,...,Bs,

在每种水平组合(Ai, Bj)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 $x_{ijk} \sim N(\mu_{ij}, \sigma^2)$， i=1,2,...,r, j=1,2,...,s, k=1,2,...t,

且各x_ijk相互独立. 建立如下数学模型

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

其中α_i为因素A的第i个水平的效应， β_i为因素B的第j个水平的效应. δ_{ij}表示Ai和Bj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,...,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,...,s,

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

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

ST为总离差平方和，SA为因素A效应平方和，SB为因素B效应平方和，SE 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FA, FB, F_{AxB}作为H01, H02, H03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：一种火箭使用了四种燃料，三种推进器，进行射程试验，每种燃料与每种推进器的组合各发射火箭2 次，得到如下结果：

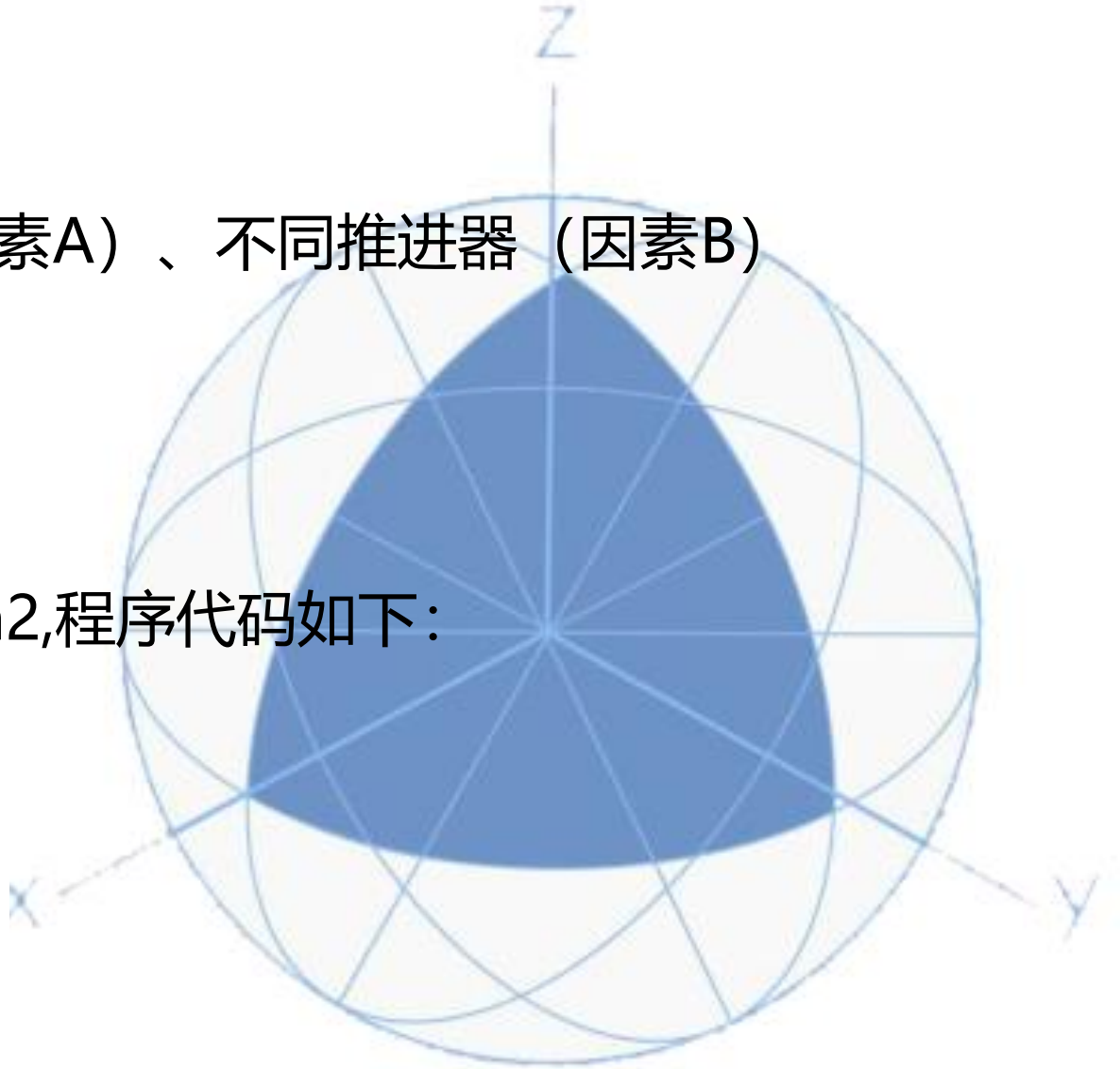
	B_1	B_2	B_3
A_1	58.2,52.6	56.2,41.2	65.3,60.8
A_2	49.1,42.8	54.1,50.5	51.6,48.4
A_3	60.1,58.3	70.9,73.2	39.2,40.7
A_4	75.8,71.5	58.2,51.0	48.7,41.4

试问在水平0.05下，检验不同燃料（因素A）、不同推进器（因素B）下的射程是否有显著差异?

交互作用是否显著?

利用Matlab实用统计工具箱中的anova2,程序代码如下:

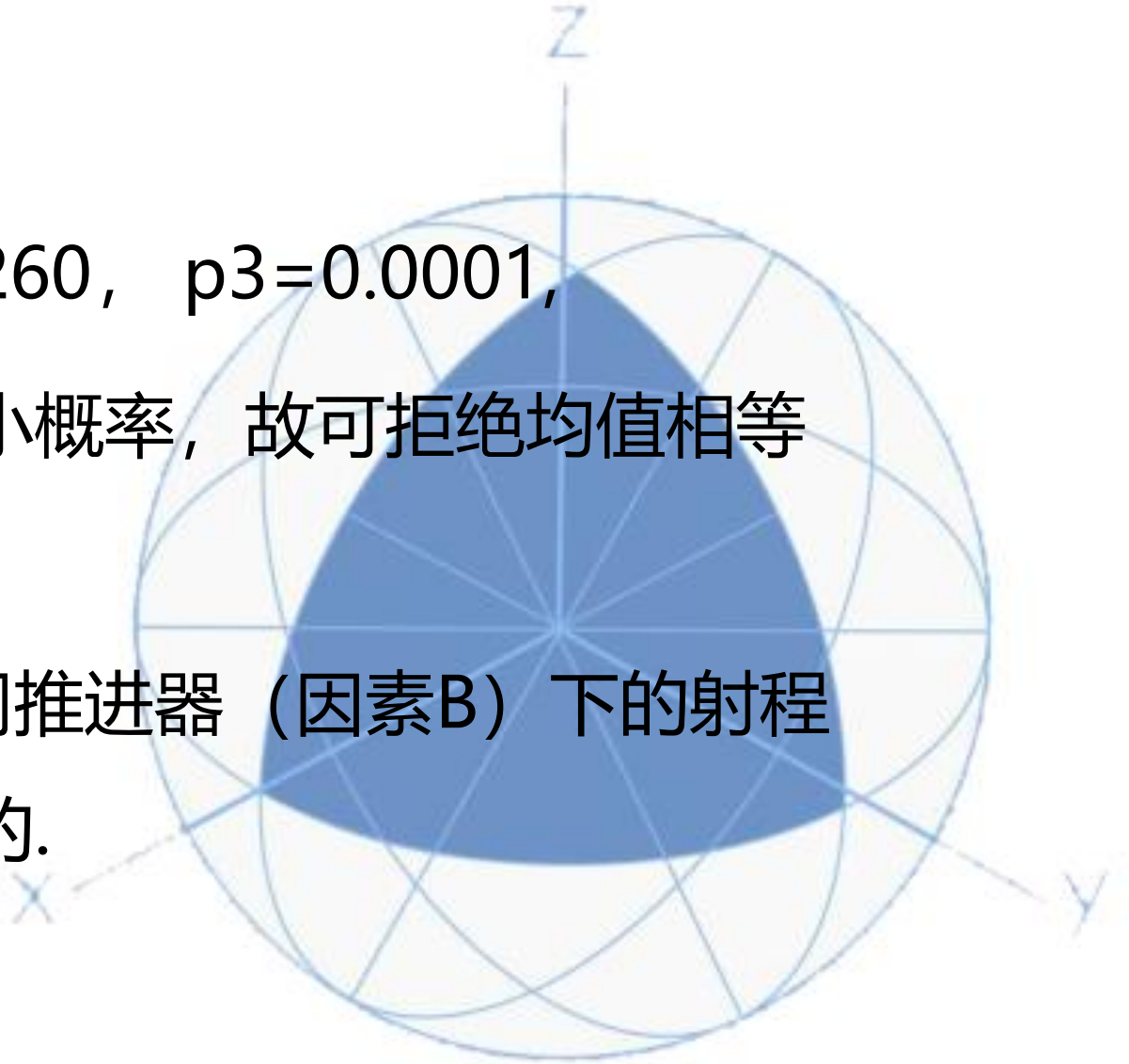
```
w=[58.2  56.2  65.3;52.6  41.2  60.8
    49.1  54.1  51.6;42.8  50.5  48.4
    60.1  70.9  39.2;58.3  73.2  40.7
    75.8  58.2  48.7;71.5  51.0  41.4];
p=anova2(w,2)
```



可求解得到p1=0.0035, p2=0.0260， p3=0.0001,

表明各试验均值相等的概率都为小概率，故可拒绝均值相等假设.

即认为不同燃料（因素A）、不同推进器（因素B）下的射程有显著差异，交互作用也是显著的.



6.3.3方差分析的模型检验

1、正态性

在讲方差分析原理时，我们做了一个这样的假设：每种水平或水平组合下的试验数据来自于方差相同的正态总体.这个总前提是不可忽视的.因为在此前提下，我们才提出了方差分析的检验假设：各正态总体的均值相同.也是在这个检验假设的基础上才有各因素均方差服从卡方分布的讨论，以及构造出服从F分布的统计量. 因此，如果没有这个总前提，即数据正态性，方差分析就成为空谈，统计和推断也就失去了意义.

当数据本身就不服从正态分布时，方差分析的过程就不可信，结论很可能是完全错误的，因此在必要的时候，可对数据的正态性进行检验.如可利用残差平方图、残差的正态概率图等等.

2、方差齐性

显然，方差齐性是方差分析的又一个前提条件，也就是我们说的样本来自于方差相同的正态总体.总体的方差可以不知，但只要能确定他们相等，就可以用方差分析的方法检验各总体的均值差异性.因此，必要时可在方差分析前用简便的方法，现对数据的方差齐性做个检验.常见方法有Bartlett法、极差比值法和平均极差法等.

3、多重比较

当对问题应用方差分析时，有一些时候，只需要对因素是否显著作出判断，因此，当得出显著性结论时，问题解决了.但在另一些场合，当得出某个因素的影响是显著的时候，还要进一步讨论，以确定哪些水平之间有显著差异，哪些水平之间没有显著差异.我们称这种进一步比较同一因素下各水平之间差异的显著性方法为多重比较法.多重比较法的具体方法有很多种，如多重比较Scheffe 法（S法）、最小显著性差异（LSD）法、多重极差检验法和多重比较Tukey（T法）.

厦门大学
XIAMEN UNIVERSITY

THANK YOU