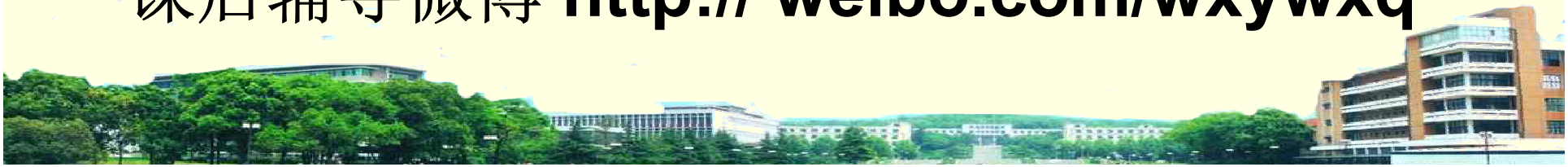
# 第4讲 关联分析

汪晓银 教授

课后辅导微博 **http:// weibo.com/wxywxq**



多个变量或者由多个变量划分所得的两组变量之间的线性相关关系，几乎普遍地存在于科学试验的一切领域之中。

本章将专门讲述多元线性相关的内容、线性相关系数的计算及其应用。

# 4.1 两个变量的相关性

## 4.1.1 简单线性相关

在一个涉及到多个变量 $x_1$，$x_2$，…，$x_p$ 或者还有 y 的问题中，任意两个变量所取的值按照以下公式

$$r_{j_1 j_2} = r_{j_2 j_1} = \frac{l_{j_1 j_2}}{\sqrt{l_{j_1 j_1} l_{j_2 j_2}}}, j_1, j_2 = 1, 2, \cdots, p \qquad j_1 \neq j_2$$

$$r_{jy} = \frac{l_{jy}}{\sqrt{l_{jj} l_{yy}}}, \quad j = 1, 2, ..., p$$

所算出的相关系数称为简单相关系数，$r_{j_1 j_2}$ 称为 $x_{j_1}$ 与 $x_{j_2}$ 的简单相关系数，$r_{jy}$ 称为 $x_j$ 与 y 的简单相关系数。

# 4.1 两个变量的相关性

由 $|r|=\sqrt{\dfrac{F}{F+(n-2)}}$ 得到界值 $|r_\alpha|=\sqrt{\dfrac{F_\alpha(1,\ n\text{-}2)}{F_\alpha(1,\ n\text{-}2)+(n-2)}}$

可以对 **r** 做检验.

两个变量之间，用**简单相关系数**所表示的相关关系称为**简单线性相关**。

如果只了解变量 **y** 与 $x_1, x_2, \cdots, x_p$ 的简单相关关系，还不足以刻画出**y** 与 $x_1, x_2, \cdots, x_p$ 之间的内在联系。

作为简单相关系数的发展，以下讲述复相关系数和偏相关系数。

# 4.2 偏相关系数

在涉及**多个变量**的问题中，**任意两个变量**都可能存在着程度不同的线性相关关系。

某两个变量变化取值时，其他的变量也在变化取值并且任意两个变量变化所取的值，都可能受到其他变量变化取值的影响。

因此，**两个变量之间的简单相关系数**往往不能反映这

# 4.2 偏相关系数

- 两个变量之间真实的线性相关关系，有必要在其他变量都保持不变的情况下计算某两个变量的相关系数。并且，为了与简单相关系数有所区别，**在其他变量都保持不变的情况下，某两个变量的相关系数称为偏相关系数。**

# 4.2 偏相关系数

这里所说的“保持不变”，含意是用统计学的方法消去其他变量变化取值的影响。

在多个变量分别以 $x_1$，$x_2$，…，$x_p$ 表示的问题中，定义其中的变量 $x_{j_1}$ 与 $x_{j_2}$ 的偏相关系数等于变量 $x_{j_1}-\hat{x}_{j_1}$ 与 $x_{j_2}-\hat{x}_{j_2}$ 的简单相关系数，而 $\hat{x}_{j_1}$ 及 $\hat{x}_{j_2}$ 分别是由 $x_{j_1}$ 及 $x_{j_2}$ 关于其他变量的线性回归方程所得的回归估计值。

# 4.2 偏相关系数

变量 $x_{j_1}$ 与 $x_{j_2}$ 的偏相关系数记为 $r_{j_1 j_2 \cdot}$ 或 $r_{j_1 j_2 \cdot 12\cdots(j_1-1)(j_1+1)\cdots(j_2-1)(j_2+1)\cdots p}$，式中的 $j_1 < j_2$。

根据上述定义，可以证明偏相关系数

$$r_{j_1 j_2 \cdot} = \frac{-c_{j_1 j_2}}{\sqrt{c_{j_1 j_1} c_{j_2 j_2}}},$$

式中的 $c_{j_1 j_2}$ 与 $c_{j_1 j_1}$ 及 $c_{j_2 j_2}$ 分别是矩阵 $\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$ 的逆矩阵中第 $j_1$ 行第 $j_2$ 列与第 $j_1$ 行第 $j_1$ 列及第 $j_2$ 行第 $j_2$ 列的元素。

# 4.2 偏相关系数

偏相关系数 $r_{jy.}$ 的显著性检验与多元线性回归方程

$$\hat{y} = b_0 + \sum_j b_j x_j$$

中变量 $x_j$ 的回归系数 $b_j$ 的显著性检验一致。

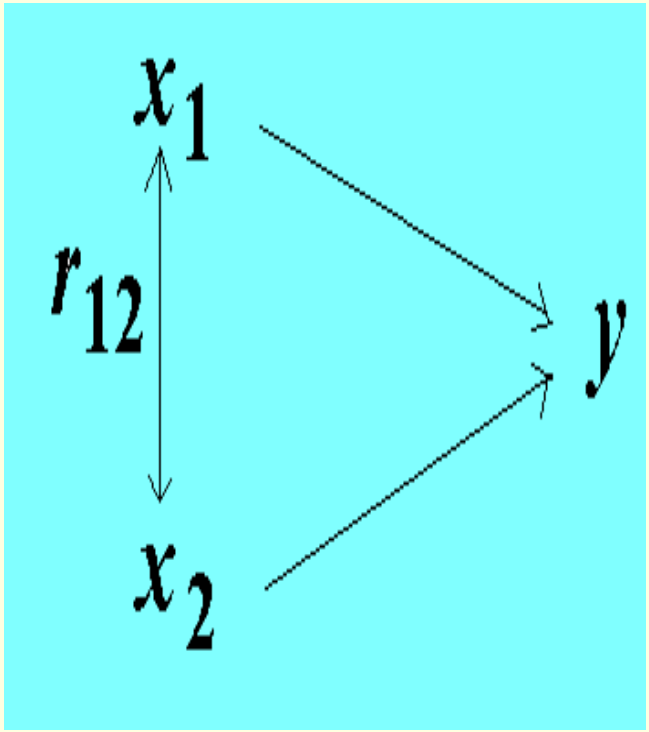
因此，偏相关系数 $r_{jy.}$ 作显著性检验时，可先算出

$$F = \frac{r_{jy.}^2}{(1 - r_{jy.}^2)/(n - p - 1)}$$

后同临界值 $F_\alpha(1, n-p-1)$ 进行比较，显著性检验的结果与多元线性回归方程 $\hat{y} = b_0 + \sum_j b_j x_j$ 中变量 $x_j$ 的回归系数 $b_j$ 作显著性检验的结果一致。

# 4.3 通径系数

设有因变量 $y$ 及自变量 $x_1$ 与 $x_2$ . 记 $x_1$ 与 $x_2$ 的简单相关系数 为 $r_{12}$ 且不为零. 如果将$x_1$ 与 $x_2$ 对$y$的影响图解为下图：



# 4.3 通径系数

称 $x_i$ 指向 $y$ 的连接线 $x_i \to y$ 为直接通径 ；

称 $x_i \to x_j \to y$ 为间接通径 。

在直接通径上，

若 $x_j$ 的取值增加一个标准差单位时，$y$ 将要改变的标准差单位数 $p_j$ 称为通径 $x_j \to y$ 的系数。

# 4.3 通径系数

通径系数 $p_j$ 可以看作是 $x_j$ 对 $y$ 的标准效应。当 $x_j$ 增加时，若 $y$ 增加，则 $p_j > 0$；若 $y$ 反而减少，则 $p_j < 0$，而 $p_j$ 的绝对值则反映 $x_j$ 对 $y$ 的标准影响力。可根据 $p_j$ 的绝对值确定 $x_j$ 对于改变 $y$ 的取值的相对重要性。在讲标准回归方程时已经说明：通径系数 $p_j$ 就是 $x_j$ 的标准回归系数。

# 4.3 通径系数

定义间接通径 $x_{j_1} \to x_{j_1} \to y$ 的系数

$$p_{j_1 \to j_2 \to y} = r_{j_1 j_2} p_{j_2},$$

间接通径 $x_{j_2} \to x_{j_1} \to y$ 的系数

$$p_{j_2 \to j_1 \to y} = r_{j_1 j_2} p_{j_1}。$$

# 4.3 通径系数

由通径系数定义，先写出标准回归方程的正规方程组

$$\begin{cases} d_1 r_{11} + d_2 r_{12} + \cdots + d_p r_{1p} = r_{1y}, \\ d_1 r_{21} + d_2 r_{22} + \cdots + d_p r_{2p} = r_{2y}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ d_1 r_{p1} + d_2 r_{p2} + \cdots + d_p r_{pp} = r_{py}, \end{cases}$$

简单相关系数的分解

# 4.3 通径系数

**例** 自变量 $x_1$， $x_2$，$x_3$， $x_4$ 与因变量 $y$ 的直接通径系数及间接通径系数.

解 作通径分析前，先建立因变量$Y$的**四元标准线性回归方程并作回归系数的显著性检验**，发现变量$x_4$的回归系数不显著，故在以下的通径分析中不考虑 $x_4$。

# 4.3 通径系数

得到标准化回归方程的正规方程如下：

$$\begin{cases} p_1 - 0.135742\,p_2 + 0.5007305\,p_3 = 0.8973138 \\ -0.135742\,p_1 + p_2 - 0.148887\,p_3 = 0.0461919 \\ 0.5007305\,p_1 - 0.148887\,p_2 + p_3 = 0.6889796, \end{cases}$$

以及直接通径系数为：

$p_1$=0.753,　$p_2$=0.199,　$p_3$=0.341.

# 4.4 典型相关系数

现实生活中两组变量间的相关关系的问题很多，例如家庭的特征（如户主的年龄、家庭的年收入、户主的受教育程度等）与消费模式（如每年去餐馆就餐的频率、每年外出看电影的频率等）等等。为此，**1936**年由**Hulling**提出了典型相关分析，揭示了两组多元随机变量之间的关系。

# 4.4 典型相关系数

## 典型相关分析基本思想

通常情况下，为了研究两组变量

$$(x_1, x_2, \cdots, x_p)\ (y_1, y_2, \cdots, y_q)$$

的相关关系，可以用最原始的方法，分别计算两组变量之间的全部相关系数，一共有**pq**个简单相关系数，这样又烦琐又不能抓住问题的本质。如果分别找出两组变量的各自的某个线性组合，讨论线性组合之间的相关关系，则更简捷。

# 4.4 典型相关系数

首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，

$$\begin{cases} u_1 = a_{11}x_1 + a_{21}x_2 + \cdots + a_{p1}x_p \\ v_1 = b_{11}y_1 + b_{21}y_2 + \cdots + b_{q1}y_q \end{cases}$$

然后再在每组变量中再找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性，即$\mathbf{u}_2$和$\mathbf{v}_2$与$\mathbf{u}_1$和$\mathbf{v}_1$相互独立，但$\mathbf{u}_2$和$\mathbf{v}_2$相关，

$$\begin{cases} u_2 = a_{12}x_1 + a_{22}x_2 + \cdots + a_{p2}x_p \\ v_2 = b_{12}y_1 + b_{22}y_2 + \cdots + b_{q2}y_q \end{cases}$$

如此下去，直至两组变量的相关性被提取完为止。

# 4.4 典型相关系数

对系数$a_{jk}$和$b_{jk}$的要求是：

(1)使各个$u_k$和$v_k$的均值为0,标准差为1;

(2)使任意两个$u_k$彼此独立或不相关,使任意两个$v_k$彼此独立或不相关,使$u_{k_1}$和$v_{k_2}$当$k_1 \neq k_2$时彼此独立或不相关 。

(3)使$u_k$和$v_k$的相关系数$r_k$满足关系式

$$1 \geq r_1 \geq r_2 \geq \cdots \geq r_p \geq 0。$$

称$u_k$和$v_k$为一对典型变量,称$r_k$为典型相关系数.

# 4.4 典型相关系数

在理论上,典型变量的对数和相应的典型相关系数的个数可以等于两组变量中数目较少的那一组变量的个数。其中, $u_1$ 和 $v_1$ 的相关系数 $r_1$ 反映的相关成分最多, 称为第一对典型变量; $u_2$ 和 $v_2$ 的相关系数 $r_2$ 反映的相关成分次多 ,称为第二对典型变量 ;…… $u_p$ 和 $v_p$ 的相关系数 $r_p$ 反映的相关成分最少,称为最后一对典型变量。

# 4.4 典型相关系数

**在应用时,只保留少数几对典型变量.**

**确定保留对数的依据:**

**(1)**对典型相关系数作显著性检验,看显著性检验的结果;

**(2)**结合应用看典型变量和典型相关系数的实际解释.

# 4.4 典型相关系数

## 典型相关分析的原理

若记 $a_k=\begin{pmatrix}a_{1k}\\a_{2k}\\\vdots\\a_{pk}\end{pmatrix}, x=\begin{pmatrix}x'_1\\x'_2\\\vdots\\x'_p\end{pmatrix}, b_k=\begin{pmatrix}b_{1k}\\b_{2k}\\\vdots\\b_{qk}\end{pmatrix}, y=\begin{pmatrix}y'_1\\y'_2\\\vdots\\y'_q\end{pmatrix},$

则有 $u_k=a'_k x, v_k=b'_k y$。

# 4.4 典型相关系数

相关系数矩阵 $\boldsymbol{R}_{xx}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix}$,

$\boldsymbol{R}_{yy}=\begin{pmatrix} r_{(p+1)(p+1)} & r_{(p+1)(p+2)} & \cdots & r_{(p+1)(p+q)} \\ r_{(p+2)(p+1)} & r_{(p+2)(p+2)} & \cdots & r_{(p+2)(p+q)} \\ \vdots & \vdots & & \vdots \\ r_{(p+q)(p+1)} & r_{(p+q)(p+2)} & \cdots & r_{(p+q)(p+q)} \end{pmatrix}$,

# 4.4 典型相关系数

$$R_{xy}=R'_{yx}=\begin{pmatrix} r_{1(p+1)} & r_{1(p+2)} & \cdots & r_{1(p+q)} \\ r_{2(p+1)} & r_{2(p+2)} & \cdots & r_{2(p+q)} \\ \vdots & \vdots & & \vdots \\ r_{p(p+1)} & r_{p(p+2)} & \cdots & r_{p(p+q)} \end{pmatrix},$$

由 $u_k=a'_k x, v_k=b'_k y$ 可得 $u_k v_k=u_k v'_k=a'_k x y' b_k$,

同理可得 $u_k^2=a'_k x x' a_k, v_k^2=b'_k y y' b_k$。

# 4.4 典型相关系数

由$E(x)=0, E(y)=0$可得 $E(u_k)=0, E(v_k)=0$。

由$E(xx')=R_{xx}, E(yy')=R_{yy}, E(xy')=R_{xy}$可得

$$D(u_k)=E(u_k^2)=a_k'E(xx')a_k=a_k'R_{xx}a_k=1,$$

$$D(v_k)=E(v_k^2)=b_k'E(yy')b_k=b_k'R_{yy}b_k=1,$$

$$r_k=\frac{\mathrm{cov}(u_k,v_k)}{\sqrt{D(u_k)D(v_k)}}=E(u_kv_k)$$

$$=a_k'E(xy')b_k=a_k'R_{xy}b_k=b_k'R_{xy}'a_k。$$

# 4.4 典型相关系数

要使$r_k$最大,并且满足条件 $D(u_k)=1$和$D(v_k)=1$可用拉格朗日乘数法 ,引入拉格朗日乘数 $\lambda$、$\mu$及

$$\varphi = E(u_k v_k) - \frac{\lambda}{2}[D(u_k)-1] - \frac{\mu}{2}[D(v_k)-1],$$

由$\frac{\partial \varphi}{\partial a_k}=0$及$\frac{\partial \varphi}{\partial b_k}=0$得到方程组

$$R_{xy}b_k - \lambda R_{xx}a_k = 0, R'_{xy}a_k - \mu R_{yy}b_k = 0。$$

# 4.4 典型相关系数

***Now prove that*** $\lambda=\mu=r_k$。

由$D(u_k)=E(u_k^2)=a_k'E(xx')a_k=a_k'R_{xx}a_k=1$,

$$D(v_k)=E(v_k^2)=b_k'E(yy')b_k=b_k'R_{yy}b_k=1,$$

由$R_{xy}b_k-\lambda R_{xx}a_k=0, R'_{xy}a_k-\mu R_{yy}b_k=0$得到

$$a_k'R_{xy}b_k-\lambda a_k'R_{xx}a_k=a_k'R_{xy}b_k-\lambda=0,$$

$$b_k'R'_{xy}a_k-\mu b_k'R_{yy}b_k=b_k'R'_{xy}a_k-\mu=0。$$

$$\lambda=a_k'R_{xy}b_k=b_k'R'_{xy}a_k=\mu=r_k。$$

# 4.4 典型相关系数

以下证明：$|R_{xx}^{-1}R_{xy}R_{yy}^{-1}R'_{xy}-\lambda^2 I|=0$。

这是计算 $\lambda=\mu=r_k$ 的方程组。

由 $R_{xy}b_k-\lambda R_{xx}a_k=0, R'_{xy}a_k-\mu R_{yy}b_k=0$ 得到

$$R_{xy}R_{yy}^{-1}R'_{xy}a_k-\mu R_{xy}R_{yy}^{-1}R_{yy}b_k$$

$$=R_{xy}R_{yy}^{-1}R'_{xy}a_k-\lambda R_{xy}b_k=0,$$

# 4.4 典型相关系数

由 $\boldsymbol{R}_{xy}\boldsymbol{b}_k-\lambda\boldsymbol{R}_{xx}\boldsymbol{a}_k=\boldsymbol{0},\boldsymbol{R}'_{xy}\boldsymbol{a}_k-\mu\boldsymbol{R}_{yy}\boldsymbol{b}_k=\boldsymbol{0}$ 得到

$$\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}'_{xy}\boldsymbol{a}_k-\lambda\boldsymbol{R}_{xy}\boldsymbol{b}_k=\boldsymbol{0}\text{ 及 }\boldsymbol{R}_{xy}\boldsymbol{b}_k=\lambda\boldsymbol{R}_{xx}\boldsymbol{a}_k,$$

$$\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}'_{xy}\boldsymbol{a}_k-\lambda^2\boldsymbol{R}_{xx}\boldsymbol{a}_k=\boldsymbol{0},$$

$$\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}'_{xy}\boldsymbol{a}_k-\lambda^2\boldsymbol{I}\boldsymbol{a}_k=\boldsymbol{0},$$

$$(\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}'_{xy}-\lambda^2\boldsymbol{I})\boldsymbol{a}_k=\boldsymbol{0},$$

# 4.4 典型相关系数

请对照线性代数中的 $Ax=\lambda x,(A-\lambda I)x=0,$

$\lambda$ 为 $A$ 的特征根，$x$ 为特征向量，

$|A-\lambda I|=0$ 为特征方程，因此这里的

$$|R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{xy}'-\lambda^2 I|=0。$$

## 4.4.4 典型相关系数的显著性检验

$$H_0: r_k=0 \leftrightarrow H_1: r_k \neq 0$$

可对典型相关系数的显著性作 $\chi^2$ 检验，所用的统计量为

# 4.4 典型相关系数

$$\chi^2 = -[n-k-\frac{1}{2}(p+q+1)]\ln\Lambda_k,$$

$$\text{where } \Lambda_k = (1-r_k^2)(1-r_{(k+1)}^2)\cdots(1-r_p^2),$$

所用的统计量服从 $\chi^2((p-k+1)(q-k+1))$分布。

然后根据典则相关系数的显著性，确定典型变量应保留的对数.

# 4.4 典型相关系数

例：蔬菜产出水平主要体现在蔬菜总产量(Y1)、人均蔬菜占有量(Y2)、蔬菜总产增长速度(Y3)三个方面，并称作因变量组（简称“产出组”）。问题：因变量组与自变量X1(市场经济综合因素)、X2(劳动力动力因素) 、X3(气候因素)(简称“影响组”)的关系如何？

```
data ex;
input y1-y3  x1-x3 @@;
cards;
/*数据省略*/
proc cancorr data=ex all;var y1-y3; with  x1-x3;
run;
```

# 4.4 典型相关系数

程序运行结果如下：

Canonical Correlation Analysis

| | Canonical Correlation | Adjusted Canonical Correlation | Approximate Standard Error | Squared Canonical Correlation |
|---|---|---|---|---|
| 1 | 0.982193 | 0.977396 | 0.010189 | 0.964703 |
| 2 | 0.810462 | 0.784388 | 0.099059 | 0.656848 |
| 3 | 0.439231 | . | 0.232983 | 0.192924 |

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

| | Eigenvalue | Difference | Proportion | Cumulative | Likelihood Ratio | Approximate F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 27.3309 | 25.4168 | 0.9270 | 0.9270 | 0.00977553 | 10.88 | 9 | 17.187 | <.0001 |
| 2 | 1.9142 | 1.6751 | 0.0649 | 0.9919 | 0.27694975 | 3.60 | 4 | 16 | 0.0282 |
| 3 | 0.2390 | | 0.0081 | 1.0000 | 0.80707608 | 2.15 | 1 | 9 | 0.1765 |

# 4.4 典型相关系数

整理得到蔬菜产出水平与影响因素的三个自变量的典型相关系数及特征值

| 序号 | 典型相关系数 | 标准误差 | 特征值 | 方差比率 | 累计方差比率 |
|---|---|---|---|---|---|
| 1 | 0.982193 | 0.010189 | 27.3309 | 0.9270 | 0.9270 |
| 2 | 0.810462 | 0.099059 | 1.9142 | 0.0649 | 0.9919 |
| 3 | 0.439231 | 0.232983 | 0.2390 | 0.0081 | 1.0000 |

结果表明：前两个典型相关系数较高，表明相应典型变量之间密切相关。

# 4.4 典型相关系数

| 序号 | F计算值 | 自由度 | F检验的显著性概率 |
|---|---|---|---|
| 1 | 10.88 | 9 | 0.0001 |
| 2 | 4.50 | 4 | 0.0282 |
| 3 | 2.15 | 1 | 0.1765 |

结果表明：只有前两对典型变量通过了统计量检验，表明相应典型变量之间相关关系显著，能够用三个自变量影响变量来解释产出变量。

# 4.4 典型相关系数

## 冗余度分析的结果

Canonical Redundancy Analysis

| Canonical Variable Number | Standardized Variance of the VAR Variables Explained by Their Own Canonical Variables: Proportion | Cumulative Proportion | Canonical R-Square | Standardized Variance of the VAR Variables Explained by The Opposite Canonical Variables: Proportion | Cumulative Proportion |
|---|---|---|---|---|---|
| 1 | 0.6481 | 0.6481 | 0.9647 | 0.6253 | 0.6253 |
| 2 | 0.2054 | 0.8535 | 0.6568 | 0.1349 | 0.7602 |
| 3 | 0.1465 | 1.0000 | 0.1929 | 0.0283 | 0.7884 |

| Canonical Variable Number | Standardized Variance of the WITH Variables Explained by Their Own Canonical Variables: Proportion | Cumulative Proportion | Canonical R-Square | Standardized Variance of the WITH Variables Explained by The Opposite Canonical Variables: Proportion | Cumulative Proportion |
|---|---|---|---|---|---|
| 1 | 0.3335 | 0.3335 | 0.9647 | 0.3217 | 0.3217 |
| 2 | 0.3338 | 0.6672 | 0.6568 | 0.2192 | 0.5409 |
| 3 | 0.3328 | 1.0000 | 0.1929 | 0.0642 | 0.6051 |

# 4.4 典型相关系数

典型变量的解释能力

| 序号 | 产出组与影响组典型相关系数平方 | 对产出组解释能力 | 产出组方差被影响组典型变量解释比例 | 对影响组解释能力 | 影响组方差被产出租典型变量解释比例 |
|---|---|---|---|---|---|
| 1 | 0.9647 | 0.6481 | 0.6253 | 0.3335 | 0.3217 |
| 2 | 0.6568 | 0.2054 | 0.1349 | 0.3338 | 0.2192 |
| 3 | 0.1929 | 0.1465 | 0.0283 | 0.3328 | 0.0642 |

①前两对典型变量的解释能力均较强；②第一、第二对典型变量具有较高的解释百分比，典型相关系数的平方表明，产出变量中分别有**94.47%**和**65.68%**的信息可以由相应的影响变量予以解释；③前两对典型变量的重叠系数较大，产出组的方差被影响组典型变量解释的比例分别为**62.53%**、**14.49%**。由于第三对典型变量在上述②、③项指标中的数值均较小，且未能通过F检验．因此舍弃第三对典型变量，只选前两对典型变量进行分析。

# 4.4 典型相关系数

典型相关模型结果如下：

The CANCORR Procedure

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

| | V1 | V2 | V3 |
|---|---|---|---|
| y1 | 6.1649 | 14.7443 | -19.9180 |
| y2 | -5.2034 | -15.0750 | 19.9861 |
| y3 | 0.0696 | 0.9105 | 0.4820 |

Standardized Canonical Coefficients for the WITH Variables

| | W1 | W2 | W3 |
|---|---|---|---|
| x1 | 0.9953 | -0.0132 | -0.0962 |
| x2 | -0.0054 | 0.9591 | -0.2831 |
| x3 | -0.0948 | -0.2804 | -0.9552 |

| 序号 | 典型相关模型 |
|---|---|
| 1 | $v1=4.1649\ Y_1-5.2034\ Y_2+0.0696\ Y_3$<br>$w1=0.9953X_1-0.0054\ X_2-0.0948X_3$ |
| 2 | $v2=14.7443Y_1-15.0750Y_2+0.9105Y_3$<br>$w2=-0.0132\ X_1+0.9591\ X_2-0.2804\ X_3$ |

# 4.4 典型相关系数

结果分析：自变量**X1**即市场经济综合因素对中国蔬菜产出水平起根本性作用。市场经济综合因素与蔬菜总产出的关系体现在第一对典型变量**v1**和**w1**中，**v1**是中国蔬菜产出水平各指标的线性组合，其中，蔬菜总产出（**Y1**）的载荷为**4.164**，是各产出水平指标中最大的。**w1**是影响因素指标的线性组合，其中市场经济综合因素**(X1)**的载荷为**0.9953**，远远超过**w1**内其它指标的数值。考虑到第一对典型相关变量的相关系数几乎接近于**1**，可以认为，市场经济综合因素对蔬菜总产出水平起根本性作用。自变量**X2**即劳动力动力因素是决定人均蔬菜占有量的关键因素。

# 4.4 典型相关系数

- 第二对典型变量中．人均蔬菜占有量(Y2)在典型变量v2中的载荷为－15.075，是各产出水平指标中最大的，而自变量X2则在典型变量w2中载荷最大，为0.9591。这一对典型相关变量的相关系数非常之高，表明自变量X2对劳动力动力因素起关键作用.
- 在第二对典型变量中，Y1与劳动力动力因素关系也非常密切。因为在第二对典型变量中，Y1在v2中的载荷14.7443，与Y2差距并不明显。由此可以分析的处，用Y1作为产出水平的代表，X1、X2、X3作为影响变量建立因果拟合模型效果是最好的。

# 4.5 独立性检验

- 设一个总体中的**n**个元素可以按两种标志进行分类，并已知按第一种标志划分为**r**个类，按第二种标志划分为**s**个类，观测到第**ij**类中元素的个数为$n_{ij}$ **(i＝1至r、j＝1至s)**，

$$且\sum_{i} n_{ij} = n_{\cdot j}，\sum_{j} n_{ij} = n_{i\cdot}，$$

# 4.5 独立性检验

- 要检验这两种分类标志是否相互独立。设**H0**为这两种分类标志相互独立，则检验**H0**相当于检验一个两维的离散型分布律是否等于两个边缘分布律的乘积，即

$$p_{ij} = p_{i\cdot} \cdot p_{\cdot j},$$

式中的$p_{ij}$是按两种标志分类时各类中元素出现的概率，而$p_{i\cdot}$和$p_{\cdot j}$则分别是按第一种和第二种标志分类时各类中元素出现的概率。作检验时先用极大似然法求出

$$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n}\text{和}\hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n},$$

再计算统计量 $\chi^2=\sum_i\sum_j\frac{(n_{ij}-n\hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{n\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$ 的观测值，

当$\chi^2>\chi^2_{1-\alpha}(f)$时放弃$H_0$，式中的$f=(r-1)(s-1)$为自由度，$\alpha$ 为显著性水平。

SAS 软件中的 freq 模块可以直接进行独立性检验。

# 4.5 独立性检验

**例** 抽取两批各500人分别使用及不使用某种预防感冒的措施，分类统计的结果如下：

| 预防措施 | 未患感冒 | 患感冒一次 | 患感冒一次以上 |
|---|---|---|---|
| 未使用 | 224人 | 136人 | 140人 |
| 使用 | 252人 | 145人 | 103人 |

试检验这种预防感冒的措施是无效的。

# 4.5 独立性检验

- 解 原假设$H_0$为是否使用预防措施与患感冒情况相互独立。
- 编写程序如下：
- **data ex;do a=1 to 2;do b=1 to 3; /*两行三列*/**
- **input f @@;output;end;end; /*输入样本数*/**
- **cards;**
- **224 136 140 252 145 103**
- **;**
- **proc freq;weight f;**
- **tables a*b/chisq;run; /*chisq表示检验*/**

The FREQ Procedure

Table of a by b

| a | b | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | Total |
| 1 | 224 | 136 | 140 | 500 |
| | 22.40 | 13.60 | 14.00 | 50.00 |
| | 44.80 | 27.20 | 28.00 | |
| | 47.06 | 48.40 | 57.61 | |
| 2 | 252 | 145 | 103 | 500 |
| | 25.20 | 14.50 | 10.30 | 50.00 |
| | 50.40 | 29.00 | 20.60 | |
| | 52.94 | 51.60 | 42.39 | |
| Total | 476 | 281 | 243 | 1000 |
| | 47.60 | 28.10 | 24.30 | 100.00 |

Frequency
Percent
Row Pct
Col Pct

# 4.5 独立性检验

Statistics for Table of a by b

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 7.5691 | 0.0227 |
| Likelihood Ratio Chi-Square | 2 | 7.5920 | 0.0225 |
| Mantel-Haenszel Chi-Square | 1 | 6.3498 | 0.0117 |
| Phi Coefficient | | 0.0870 | |
| Contingency Coefficient | | 0.0867 | |
| Cramer's V | | 0.0870 | |

Sample Size = 1000

# 4.5 独立性检验

- 结果有两部分：
- 第一部分是联立表的计算过程，包含四个统计量，分别是：每组频数（Frequency），按百分数计算的频率（Percent），占该行的百分比（Row Pct），占该列的百分比（Col Pct）。例如，第一行第二列的内容按顺序为：未使用预防措施患感冒一次的人数为136人，占总人数的14.50%，占这行的27.20%，占这一列的48.40%。
- 第二部分Statistics for Table of a by b才是检验结果。其中主要看Chi-Square这一行，其自由度为2，χ2（Chi-Square value）=7.5691，且Prob值为0.0227<0.05，故落在拒绝域，接受备择假设，即预防与感冒这两因素之间不是相互独立的，即措施对患感冒是有关系的。