生物统计学（XI）
第十章：协方差分析

西安交通大学数学与统计学院

May, 2018

- 协方差分析是将方程分析与回归分析结合应用的一种综合统计分析方法，它将乘积和与平方和同时按照变异来源进行分析，利用协变量来降低试验误差，矫正处理平均数，实现统计控制
- 协方差分析还可分析不同变异来源的相关关系，并对缺失数据进行估计
- 在满足协变量 x 为固定变量，离回归方差同质，各个处理 (x,y) 的总体为线性且具有共同的回归系数 3 个基本假定的基础上，可对单因素、二因素及多因素分组资料进行协方差分析，计算变量各变异来源的平方和，乘积和与自由度，检验 x 与 y 之间是否存在直线回归关系，并对矫正后的 y 值进行差异显著性检验和多重比较

- 第六章介绍的方差分析方法，是对某一种性状的变量进行分析，根据变异来源将总变异的自由度和平方和分解为相应部分的自由度和平方和；
- 但在方差分析中，常会遇到所分析的变量本身就是一个受到另一个或多个自变量影响的依变量。
- 有时对这些自变量难以进行有效控制，但又要消除其对依变量的影响，以提高试验结果的可靠程度。
  - 在动物饲喂试验中，为比较不同饲料对动物增重的影响，应选用初始体重相同（或相近）的动物来试验
  - 初始体重影响可能会增大处理间（组间）差异，是不同饲料的差异不能被真正体现，也可能会增大处理内（组内）的差异，从而降低检验功效。
  - 客观条件的限制，使得无法找到足够数量的初始体重相近的动物进行试验，这是可用统计学的方法将这种由初始体重差异造成的影响降到最低
  - 如果初始体重对增重的影响可以通过回归分析来度量，可以用回归分析先对初始体重的影响进行矫正，然后再方差分析。

- 把回归分析与方差分析结合起来的分析方法，就称为协方差分析 (analysis of covariance)。
  - 协方差分析用于比较一个变量 y 在一个或几个因素不同水平上的差异。
  - 与方差分析不同的是，y 在受这些因素影响的同时，还收到另一个变量 x 的影响，而且 x 变量的取值难以人为控制，不能作为方差分析中的一个因素处理
  - 如果 x 与 y 之间可以建立回归关系，则可用回归分析的方法排除 x 对 y 的影响，然后利用方差分析的方法对各因素水平的影响作出统计推断。
  - 通常称 y 为依变量，x 为协变量（covariate），方差分析也可用于分析多组均数间的差异有无显著性意义，只是多考虑一个协变量的因素。
- 协方差分析可用于单因素试验资料，也可以用于二因素或多因素试验资料。

协方差分析作用主要有以下 3 个方面

(1) 降低试验误差，实现统计控制

- 提高试验的精确度与灵敏度，必须严格控制试验条件的均匀性，使各处理处于尽可能一致的条件下——试验控制（experimental control）
- 但在某些情况下，即使做出很大努力也难达到试验控制预期要求
  - 如研究植物生长调节剂对减少棉铃脱落的效应，要求各处理的单株有相同的蕾铃数，这是很难达到的
  - 如果单株蕾铃数（x）与脱落率（y）之间存在直线回归关系，则可以利用这种直线关系将各处理 y 的观测值都矫正到 x 相同时的结果，使得各处理 y 的比较能够在相同的 x 基础上进行，从而得到正确的结论，实现统计控制 (statistical control)
- 统计控制是试验控制的一种辅助手段，是用统计方法来矫正因自变量的不同而对依变量所产生的影响。经过矫正，使试验误差减少，对试验处理效应的估计更为准确

(2) 分析不同变异来源的相关关系

- 相关系数可以表示两个相关变量线性相关的性质与程度

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

- 将该式的分子分母同时除以 $(n-1)$，可得

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}\frac{\sum(y-\bar{y})^2}{n-1}}}$$

其中 $\frac{\sum(x-\bar{x})^2}{n-1}$ 为 x 的均方（mean square），记作 MS_x，它是变量 x 的总体方差 σ_x^2 的无偏估计量；$\frac{\sum(y-\bar{y})^2}{n-1}$ 为 y 的均方，记作 MS_y，它是变量 y 的总体方差 σ_y^2 的无偏估计量；$\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$ 为 x 与 y 的平均的离均差的乘积和，简称为均积（mean product），记作 MP_{xy}

$$MP_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1}$$

- 与均积相应的总体参数称为协方差 (covariance)，记作 COV_{xy} 或 σ_{xy}^2:

$$COV_{xy} = \frac{\sum(x-\mu_x)(y-\mu_y)}{N}$$

统计学上已经证明均积 MP_{xy} 是总体协方差 COV_{xy} 的无偏估计量

- 因此样本相关系数 r 可用均方 MS_x, MS_y 和均积 MP_{xy} 来表示

$$r = \frac{MP_{xy}}{\sqrt{MS_x \cdot MS_y}}$$

- 相应的总体相关系数 ρ 可用 x 与 y 的总体方差与总体协方差表示

$$\rho = \frac{COV_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{COV_{xy}}{\sigma_x \sigma_y}$$

(2) 分析不同变异来源的相关关系

- 在随机模型的方差分析中，根据方差和期望方差的关系，可以得到不同来源的方差组分的估计值；同样，在随机模型的协方差分析中，根据均积和期望均积的关系，可以得到不同变异来源的协方差组分的估计值
- 有了这些估计值，就可以进一步计算出两个变量 x 与 y 之间各个变异来源的相关系数，从而进行相应的总体相关分析

(3) 估计缺失数据

- 利用方差分析的方法对缺失数据进行估计，是建立在误差平方和最小的基础上的，但处理平方和却向上偏倚。
- 如果用协方差分析的方法估计缺失数据，则既可保证误差平方和最小，又能得到无偏的处理平方和。

- 均积和均方有相似的形式，同时二者也具有相似的性质。
- 在方差分析中，一个变量的总平方和与自由度可以按照变异来源进行分解，从而求得相应的均方
- 统计学已证明，两个变量的总乘积和与自由度也可按照变异来源进行分解从而获得相应的均积。
- 设有 k 组双变量资料，每组样本皆有 n 对 (x,y) 观测值，那么该资料共有 nk 对观测值，其数据模式如下表

组别	观测值						总和	平均
1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1n}	$T_{x_1.}$	$\bar{x}_{1.}$
	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n}	$T_{y_1.}$	$\bar{y}_{1.}$
2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2n}	$T_{x_2.}$	$\bar{x}_{2.}$
	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n}	$T_{y_2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{in}	$T_{x_i.}$	$\bar{x}_{i.}$
	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in}	$T_{y_i.}$	$\bar{y}_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	x_{k1}	x_{k2}	$\cdots$	x_{kj}	$\cdots$	x_{kn}	$T_{x_k.}$	$\bar{x}_{k.}$
	y_{k1}	y_{k2}	$\cdots$	y_{kj}	$\cdots$	y_{kn}	$T_{y_k.}$	$\bar{y}_{k.}$
总计							T_x	$\bar{x}$
							T_y	$\bar{y}$

x 和 y 的总平方和与总自由度皆可按方差分析分解为组间和组内两个部分，相应的总乘积和（total sum of products）也可分解为组间乘积和（sum of products between groups）和组内乘积和（sum of products within groups）两部分。总乘积和用 $SP_{总}$ 或 SP_T 表示，组间乘积和用 $SP_{组间}$ 或 SP_t 表示，组内乘积和用 $SP_{组内}$ 或 SP_e 表示，则有

$$SP_T = SP_t + SP_e$$

各部分的相应自由度为

$$nk - 1 = (k - 1) + k(n - 1)$$

乘积和的计算公式为

$$SP_T = \sum_i \sum_j (x_{ij} - \bar{x})(y_{ij} - \bar{y}) = \sum_i \sum_j x_{ij} y_{ij} - \frac{1}{nk} T_x T_y$$

$$SP_t = n\sum_i(\bar{x}_{i\cdot}-\bar{x})(\bar{y}_{i\cdot}-\bar{y}) = \frac{1}{n}\sum_i T_{x\cdot}T_{y\cdot} - \frac{1}{nk}T_xT_y$$

$$SP_e = \sum_i\sum_j(x_{ij}-\bar{x})(y_{ij}-\bar{y}) = \sum_i\sum_j x_{ij}y_{ij} - \frac{1}{n}\sum_i T_{x\cdot}T_{y\cdot}$$

那么我们有

$$\sum_i\sum_j(x_{ij}-\bar{x})(y_{ij}-\bar{y}) = n\sum(\bar{x}_{i\cdot}-\bar{x})(\bar{y}_{i\cdot}-\bar{y}) + \sum_i\sum_j(x_{ij}-\bar{x}_{i\cdot})(y_{ij}-\bar{y}_{i\cdot})$$

例 10.1

比较 3 种不同配合饲料的效应，现将 24 头猪随机分成 3 组进行不同饲料喂养试验，测定结果见下表，试分析 3 种配合饲料对猪的增重差异是否显著。

饲料	观测值								总和	平均	
A_1	x	18	16	11	14	14	13	17	17	120	15
	y	85	89	65	80	78	83	91	85	656	82
A_2	x	17	18	18	19	21	21	16	22	152	19
	y	95	100	94	98	104	97	90	106	784	98
A_3	x	18	23	23	20	24	25	25	26	184	23
	y	91	89	98	82	100	98	102	108	768	96
总计	x									456	19
	y									2208	92

先直接对增长 y 进行方差分析，结果见下表

变异来源	df	SS	s^2	F	$F_{0.05}$	$F_{0.01}$
饲料间	2	1216	608.00	11.34**	3.47	5.78
误差	21	1126	53.26			
总变异	23	2342				

- 方差分析结果表明，3 种配合饲料对猪的增重影响差异达到极显著水平
- 但我们不能轻易相信这一推断
- 由资料可知，各组猪的始重相差较大，而始重不同对猪的增重亦有影响，一般始重大的猪增重快，始重小的猪，增重慢
- 方差分析的方法忽略了始重不同的影响，将始重与饲料对增重的效应混在一起，不能反映饲料的真实效应
- 因此需用协方差分析的方法，矫正始重对增重的影响，获得真实的饲料效应。

下面结合本例详述协方差分析的方法步骤

一、计算变量各变异来源的平方和、乘积和与自由度

- 首先计算分析所需的 6 个数据，即

$$\sum x, \sum x^2, \sum y, \sum y^2, \sum xy, nk$$

- x 变量的平方和

$$\text{总变异：} SS_{T_x} = \sum x^2 - \frac{(\sum x)^2}{nk} = 380$$

$$\text{饲料间：} SS_{t_x} = \frac{\sum T_{x_{i\cdot}^2}}{n} - \frac{(\sum x)^2}{nk} = 256$$

$$\text{误差：} SS_{e_x} = SS_{T_x} - SS_{t_x} = 124$$

- y 变量的平方和

$$\text{总变异：} SS_{T_y} = \sum y^2 - \frac{(\sum y)^2}{nk} = 2342$$

$$\text{饲料间：} SS_{t_y} = \frac{\sum T_{y_{i\cdot}^2}}{n} - \frac{(\sum y)^2}{nk} = 1216$$

$$\text{误差：} SS_{e_y} = SS_{T_y} - SS_{t_y} = 1126$$

一、计算变量各变异来源的平方和、乘积和与自由度

- x 与 y 的乘积和

$$
\begin{aligned}
\text{总变异：} SP_T &= \sum xy - \frac{\sum x \sum y}{nk} = 749 \\
\text{饲料间：} SP_t &= \frac{\sum T_{x_{i\cdot}} T_{y_{i\cdot}}}{n} - \frac{\sum x \sum y}{nk} = 448 \\
\text{误差：} SP_e &= SS_T - SS_t = 301
\end{aligned}
$$

- 相应的自由度

$$
\begin{aligned}
\text{总变异：} df_T &= nk - 1 = 23 \\
\text{饲料间：} df_t &= k - 1 = 2 \\
\text{误差：} df_e &= k(n-1) = 21
\end{aligned}
$$

将结果列于下表

表: 表 10-4: 始重 x 与增重 y 协方差分析表

变异来源	df	SS_x	SS_y	SP_{xy}	$b_{e(y\|x)}$	矫正值（离回归部分）变异的分析 df	Q	s^2	F
总变异	23	380	2342	749		-	-	-	-
饲料组	2	256	1216	448		-	-	-	-
组内	21	124	1126	301	-	-	-	-	-
矫正组间变异						-	-	-	-

二、检验 x 和 y 是否存在直线回归关系

- 计算误差项（处理内项）的回归系数 $b_{e(y|x)}$，并对线性回归关系进行显著性检验，其目的是要从组内项变异种找出始重 x 与增重 y 之间是否存在真实的线性回归关系
- 在对回归系数进行显著性检验时，假设 $H_0 : \beta = 0; H_A : \beta \neq 0$，若接受 H_0，则二者之间回归关系不显著；说明增重 y 不受始重 x 的影响，即 y 与 x 无关，可以不用考虑始重 x，而直接对增重 y 进行方差分析；
- 若否定 H_0，则说明二者存在显著的直线回归关系，表明增重 y 受始重 x 的影响，应用线性回归关系来矫正 y 值以消除因 x 的不同而产生的影响，然后根据矫正后的 y 值进行方差分析。

（一）计算误差项回归系数、回归平方和、离回归平方和与相应的自由度

- 误差项回归系数

$$b_{e(y|x)} = \frac{SP_e}{SS_{e_x}} = \frac{301}{124} = 2.4274$$

- 误差项回归平方和与自由度

$$\begin{aligned} U_e &= \frac{SP_e^2}{SS_{e_x}} = \frac{301^2}{124} = 730.6532 \\ df_{e(U)} &= 1 \end{aligned}$$

- 误差项离回归平方和与相应的自由度

$$\begin{aligned} Q_e &= SS_{e_y} - U_e = 1126 - 730.6532 = 395.3468 \\ df_{e(Q)} &= df_e - df_{e(U)} = k(n-1) - 1 = 20 \end{aligned}$$

（二）检验误差项回归系数 b_e 的显著性（t 检验）

$$
\begin{aligned}
s_{e(y|x)} &= \sqrt{\frac{Q_e}{df_{e(Q)}}} = 4.4460 \\
s_{b_e} &= \frac{s_{e(y|x)}}{\sqrt{SS_{e_y}}} = 0.3993 \\
t &= \frac{b_e}{s_{b_e}} = 6.0791
\end{aligned}
$$

查附表 3，$t_{0.01(20)} = 2.845$，该 t 值达到极显著水平，应否定 H_0，接受 H_A，因此推断：y 依 x 有极显著的直线回归关系，即猪的增重确实受到始重的影响。

三、检验矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 间的差异显著性

- 误差项（组内项）回归关系显著，需用误差项回归系数对增重 y 进行矫正，消除始重对增重的影响，从而使各种不同的饲料效应应处于始重相同水平的基础上进行比较，也就是在消除协变量 x 的影响后，比较各处理矫正 y 值的差异显著性
- 检验矫正后 y 值的差异显著性，在进行平方和计算时，并不需要将各个 y 的矫正值求出后再进行计算
- 由回归分析可知，依变量 y 的平方和可分解为回归平方和与离回归平方和：前者是 y 受 x 影响而产生的变异部分，后者是 y 去除了 x 影响后剩余的变异部分
- 矫正增重 y 计算出的各项平方和，实际上就是去除了始重 x 影响的部分
- 统计已经证明：矫正后 y 的各项平方和及自由度等于其相应变异项的离回归平方和及自由度。
- 其具体分析过程
  - 对总变异项作回归分析，求得其离回归平方和 Q'_T 和自由度 $df'_{T(Q)}$，再由 $Q'_T - Q'_e$ 和 $df'_{T(Q)} - df'_{e(Q)}$ 即得矫正 y 值的平方和与自由度，进而可对矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 间的差异显著性进行 F 检验
  - 在对矫正 y 值进行方差分析时，并不需要对各处理的矫正 y 值进行计算

（一）计算矫正增重 y 的各项平方和与自由度

- 矫正 y 值的总平方和与自由度，即总离回归和与自由度

$$Q'_T = SS_{T_y} - \frac{SP_T^2}{SS_{T_x}} = 865.6816$$

$$df'_{T(Q)} = (nk-1) - 1 = nk - 2 = 22$$

- 矫正 y 值的误差项（组内项）平方和与自由度，即误差项离回归平方和和与自由度

$$Q'_e = Q_e = SS_{e_y} - \frac{SP_e^2}{SS_{e_x}} = 395.3468$$

$$df'_{e(Q)} = k(n-1) - 1 = nk - 2 = 20$$

- 矫正 y 值处理项（组间项）的平方和与自由度

$$Q'_t = Q'_T - Q'_e = 470.3348$$

$$df'_{t(Q)} = df'_{T(Q)} - df'_{e(Q)} = 22 - 20 = 2$$

（二）列出协方差分析表，进行矫正值的方差分析

- 将矫正增重 y 的各项平方和与自由度列入上表，就得到完整的协方差分析表

$$\text{矫正（饲料）组间方差} = \frac{Q'_t}{df'_{t(Q)}} = \frac{470.3348}{2} = 235.1674$$

$$\text{矫正（饲料）组内方差} = \frac{Q'_e}{df'_{e(Q)}} = \frac{395.3468}{20} = 19.7673$$

- 则有

$$F = \frac{\text{矫正（饲料）组间方差}}{\text{矫正（饲料）组内方差}} = \frac{235.1674}{19.7673} = 11.90$$

- 查 F 表，$F_{0.01(2,20)} = 5.85, F > F_{0.01(2,20)}$，F 值达到极显著水平。提高矫正消除始重影响后，各饲料组间矫正增重差异达到极显著水平，需进行多重比较。

四、矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 间的多重比较

（一）计算各矫正增重的平均数

- 误差项（组内项）回归系数 $b_{e(y|x)}$ 表示始重 x 对增重 y 影响的性质和程度，且不包含处理间差异的影响，因此可用 $b_{e(y|x)}$ 根据平均始重的不同来矫正每一处理的增重平均值
- 矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 的公式为

$$\bar{y}_{i(x=\bar{x})} = \bar{y}_i - b_{e(y|x)}(\bar{x}_i - \bar{x})$$

式中，$\bar{y}_{i(x=\bar{x})}$ 为第 i 处理 y 的矫正平均数；$\bar{y}_i$ 为第 i 处理实际观测值 y 的平均数；$\bar{x}_i$ 为自变量 x 第 i 处理的平均数；$\bar{x}$ 为自变量 x 的总平均数

- 将数据代入，计算得到各处理矫正平均增重如下：

$$\begin{aligned}
\bar{y}_{1(x=\bar{x})} &= \bar{y}_1 - \bar{y}_{1(x=\bar{x})}(\bar{x}_1 - \bar{x}) = 91.7096 \\
\bar{y}_{2(x=\bar{x})} &= \bar{y}_2 - \bar{y}_{2(x=\bar{x})}(\bar{x}_2 - \bar{x}) = 98.0000 \\
\bar{y}_{3(x=\bar{x})} &= \bar{y}_3 - \bar{y}_{3(x=\bar{x})}(\bar{x}_3 - \bar{x}) = 86.2904
\end{aligned}$$

- 由此可见，消除始重影响后，$\bar{y}_i$ 与 $\bar{y}_{i(x=\bar{x})}$ 不仅数值不同，而且次序也发生改变
- 若不进行矫正，则会得到饲料 A_3 优于饲料 A_1 的假象

（二）矫正平均数间的多重比较

1. t 检验

- 当矫正均数的误差项自由度小于 20，且变量 x 的变异较大时，可采用两两比较的 t 检验法

$$t = \frac{\bar{y}_{i(x=\bar{x})} - \bar{y}_{I(x=\bar{x})}}{s_D}$$

其中

$$s_D = s_{e(y|x)}\sqrt{\frac{1}{n_i} + \frac{1}{n_I} + \frac{(\bar{x}_i - \bar{x}_I)^2}{SS_{e_x}}}$$

式中 $s_{e(y|x)}$ 为误差项离回归方差；n_i, n_I 为比较的两个样本容量；$\bar{x}_i, \bar{x}_I$ 为两个相比较样本的 x 变量平均数；SS_{e_x} 为 x 变量的组内项平方和

- 如果两个样本容量相同，即 $n_i = n_I = n$ 时，

$$s_D = s_{e(y|x)}\sqrt{\frac{2}{n} + \frac{(\bar{x}_i - \bar{x}_I)^2}{SS_{e_x}}}$$

- 本例中，检验 A_1 与 A_2 的平均数差异显著性，已知 $\bar{y}_{1(x=\bar{x})} = 91.7096, \bar{y}_{2(x=\bar{x})} = 98.0000, s_{e(y|x)} = 19.7673, df_{e(Q)} = 20, n_1 = n_2 = 8, \bar{x}_= 15, \bar{x}_2 = 19, SS_{e_x} = 124$，可计算得到

$$s_D = \sqrt{19.7673 \times \left[\frac{2}{8} + \frac{(15-19)^2}{124}\right]} = 2.7372,$$

$$t = \frac{\bar{y}_{1(x=\bar{x})} - \bar{y}_{2(x=\bar{x})}}{s_D} = -2.298$$

- 由于 $|t| = 2.298 > t_{0.05(20)} = 2.086$ 达到显著水平。表明饲料 A_1 与 A_2 对猪增重效果有明显差别，即 A_2 优于 A_1，同理可进行其他比较

2. LSD 法

- 由于每次比较都要分别计算两个矫正平均数差数的标准误差 s_D，显然比较麻烦
- 当误差项自由度在 20 或者 20 以上，且 x 变量的变异较小时，可用一个平均数差数标准误差进行比较，此时

$$s_D = s_{e(y|x)}\sqrt{\left(\frac{1}{n_i}+\frac{1}{n_l}\right)\left[1+\frac{SS_{t_x}}{(k-1)SS_{e_x}}\right]}$$

其中，SS_{t_x} 为 x 变量组间平方和，k 为组数。

- 当 $n_i = n_l = n$ 时，我们有

$$s_D = s_{e(y|x)}\sqrt{\left[1+\frac{SS_{t_x}}{(k-1)SS_{e_x}}\right]\times\frac{2}{n}}$$

- 按公式 $LSD_\alpha = t_\alpha \cdot s_D$，计算出最小显著差数，并与各组均数矫正值的差数进行比较，即可检验它们的显著差异性。

本例中，其误差项自由度为 20，但不满足 “x 变量的变异较小” 这一条件，因此不宜采用此法进行多重比较:

$$
\begin{aligned}
s_D &= \sqrt{19.7673 \times \left[1 + \frac{256}{(3-1)\times 124}\right] \times \frac{2}{8}} = 3.1691 \\
LSD_{0.05} &= 2.086 \times s_D = 6.6107 \\
LSD_{0.01} &= 2.845 \times s_D = 9.0161
\end{aligned}
$$

表: 3 种饲料效应的差异显著性比较（LSD 法）

饲料	矫正 y 值平均数 $\bar{y}_{i(x=\bar{x}}$	差异显著性	
		$\bar{y}_{i(x=\bar{x}} - 86.2904$	$\bar{y}_{i(x=\bar{x}} - 91.7096$
A_2	98.000	11.7096 **	6.2904
A_1	91.7096	5.4192	
A_3	86.2904		

由于变量 x 的变异较大，而在检验时使用一个共同的平均数差数标准误差，因而出现了与 t 检验不一致的结果。用 t 检验时，A_1 与 A_2 之间存在显著差异，而用 LSD 法进行比较时，二者差异未达到显著水平。

若试验设有 k 个处理，每个处理 n 个类别（重复），则 nk 对观测值可以按两向进行分组，其数据模式如下表

表: 二因素试验资料的数据模式

组	类 1		···	类 j		···	类 n		总和		平均	
1	x_{11}	y_{11}	···	x_{1j}	y_{1j}	···	x_{1n}	y_{1n}	$T_{x_1.}$	$T_{y_1.}$	$\bar{x}_{1.}$	$\bar{y}_{1.}$
2	x_{21}	y_{21}	···	x_{2j}	y_{2j}	···	x_{2n}	y_{2n}	$T_{x_2.}$	$T_{y_2.}$	$\bar{x}_{2.}$	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	y_{i1}	···	x_{ij}	y_{ij}	···	x_{in}	y_{in}	$T_{x_i.}$	$T_{y_i.}$	$\bar{x}_{i.}$	$\bar{y}_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	x_{k1}	y_{k1}	···	x_{kj}	y_{kj}	···	x_{kn}	y_{kn}	$T_{x_k.}$	$T_{y_k.}$	$\bar{x}_{k.}$	$\bar{y}_{k.}$
总和	$T_{x.1}$	$T_{y.1}$	···	$T_{x.1=j}$	$T_{y.j}$	···	$T_{x.n}$	$T_{y.n}$	T_x	T_y		
平均	$\bar{x}_{.1}$	$\bar{y}_{.1}$	···	$\bar{x}_{.j}$	$\bar{y}_{.j}$	···	$\bar{x}_{.n}$	$\bar{y}_{.n}$			$\bar{x}$	$\bar{y}$

上表的总 SP 可分解为类间，组间和误差 3 个部分，其值为

$$\begin{aligned}
\text{总 SP} &= \sum_{1}^{nk}(x_{ij}-\bar{x})(y_{ij}-\bar{y}) = \sum_{1}^{nk} x_{ij}y_{ij} - \frac{T_x T_y}{nk} \\
\text{类间 SP} &= k\sum_{1}^{n}(\bar{x}_{\cdot j}-\bar{x})(\bar{y}_{\cdot j}-\bar{y}) = \frac{1}{k}\sum_{1}^{n} T_{x_{\cdot j}} T_{y_{\cdot j}} - \frac{T_x T_y}{nk} \\
\text{组间 SP} &= n\sum_{1}^{n}(\bar{x}_{i\cdot}-\bar{x})(\bar{y}_{i\cdot}-\bar{y}) = \frac{1}{k}\sum_{1}^{k} T_{x_{i\cdot}} T_{y_{i\cdot}} - \frac{T_x T_y}{nk} \\
\text{误差 SP} &= \text{总 SP - 类间 SP - 组间 SP}
\end{aligned}$$

各 SP 的相应自由度依次为 $nk-1, n-1, k-1, (n-1)(k-1)$
根据这些 SP 和第六章计算得到的 SS，就可以进行协方差分析。

施

肥期和施肥量对杂交水稻‘南优 3 号’结实率影响的部分结果，共 14 个处理，两个区组，随机区组设计。在试验过程中发现单位面积上的颖花数对结实率有明显的回归关系，因此将颖花数（x）和结实率（y）一起测定，试做协方差分析。

具体步骤如下

一 乘积和与自由度分解

二 检验 x 和 y 是否存在直线回归关系

三 检验矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 间的差异显著性

首先用二因素资料的通常方法计算 SS_x, SS_y，各 SP 计算如下

$$\text{总 SP} = \sum_{1}^{nk} x_{ij}y_{ij} - \frac{T_x T_y}{nk} = -73/5986$$

$$\text{区组 SP} = \frac{1}{k}\sum_{1}^{n} T_{x\cdot j}\, T_{y\cdot j} - \frac{T_x T_y}{nk} = -0.7907$$

$$\text{处理 SP} = \frac{1}{k}\sum_{1}^{k} T_{x i\cdot}\, T_{y i\cdot} - \frac{T_x T_y}{nk} = -66.3636$$

$$\text{误差 SP} = \text{总 SP - 类间 SP - 组间 SP} = -6.4443$$

表: 10-8: 杂交水稻试验的平方和与乘积和

变异来源	SS_x	SS_y	SP
总变异	7.7343	802.9643	-73.5986
区组间	0.0240	26.0357	-0.7907
处理间	6.8731	694.4643	-66.3636
误差	0.8372	82.4643	-6.4443

有了上述结果，可先对 x 与 y 分别进行方差分析，结果如下表

表: 10-9:x 与 y 的方差分析

变异来源	df	x 变量			y 变量			$F_{0.05}$
		SS	s^2	F	SS	s^2	F	
区组间	1	0.0240	0.0240	0.3727	26.0357	26.0357	4.1044	4.67
处理间	13	6.8731	0.5287	8.2096**	694.4643	53.4203	8.4214**	2.57
误差	13	0.8372	0.0644		82.4643	6.3434		

上表表明，不同区组的颖花数（x）和结实率（y）都没有显著差异，但不同施肥处理的 x 和 y 的差异均达到极显著水平。在单因素资料的协方差分析中，如果 y 和 x 无关，上述推断是正确的；如果 y 与 x 有关，则不一定正确。因此，因首先明确 y 与 x 是否有线性关系。

二、检验 x 和 y 是否存在直线回归关系

由表 10-8 的误差项可得线性回归的 U_e 与 Q_e：

$$U_e = \frac{SP_e^2}{SS_{e_x}} = 49.6046$$

$$Q_e = SS_{e_y} - U_e = 82.4643 - 49.6046 = 32.8597$$

将线性回归进行 F 检验，F=18.1151，达到极显著水平，因此应对 y 值进行矫正，并对矫正平均数进行差异显著性分析，才能明确不同区组或处理对于结实率的效应。

三、检验矫正平均数 $\bar{y}_{i(x=\bar{x})}$ 间的差异显著性

二因素资料分为区组和处理两项，因此，检验矫正平均数的差异显著性需一项一项的进行。

- 检验处理的矫正平均数 $\bar{y}_{i\cdot(x=\bar{x})}$ 的差异显著性时，需将处理间和误差项的 df，SS 和 SP 相加以代替总变异的 df，SS 和 SP；
- 检验区组的矫正平均数 $\bar{y}_{\cdot j(x=\bar{x})}$ 的差异显著性时，需将区组间和误差项的 df，SS 和 SP 相加以代替总变异的 df，SS 和 SP；
- 由此，可以分别对两项矫正平均数间的差异进行显著性检验
- 区组只是局部控制的一种手段，在结果分析上只需剔除其影响，而不必研究其效应，因此这里仅对处理项的矫正平均数 $\bar{y}_{i\cdot(x=\bar{x})}$ 之间的差异显著性进行检验

表: 表 10-11: 水稻研究资料处理间矫正平均数的显著性检验

变异来源	df	SS_x	SS_y	SP	b_e	矫正值（离回归）变异分析				
						df	Q	s^2	F	$F_{0.05}$
处理 + 误差	26	7.71	776.92	-72.80		25	89.40			
处理	13	6.87	694.46	-66.36						
误差	13	0.83	82.46	-6.44	-7.69	12	32.85	2.73		
矫正差异						13	56.54	4.34	1.58	2.66

前面分析得知，误差项线性回归达到极显著水平，所以可用上表的误差项回归系数 b_e 来对各处理的 $\bar{y}_i$ 进行矫正。$b_e = -7.69$ 表示颖花数每增加 1 万/平方米，结实率将下降 7.69% 。将 b_e 代入，有方程

$$\bar{y}_{i(x=\bar{x})} = \bar{y}_i + 7.6974(\bar{x}_i - 3.7714)$$

上式可将各处理的结实率都矫正到颖花数为每平方米 3.7714 万个的结实率，由此可计算各处理矫正平均数。

- 处理 1:$\bar{y}_{1(x=\bar{x})} = 59.5 + 7.6974 \times (4.455 - 3.7714) = 64.76\%$
- 处理 2:
- $\vdots$
- 计算出的 $\bar{y}_{i(x=\bar{x})}$ 已经与单位面积上颖花数的多少无关，故在相互比较时更为真实。
- 但，在未算出 $\bar{y}_{i\cdot(x=\bar{x})}$ 之前，我们已知其 F 值为达到显著水平，说明各处理矫正平均数并无显著差异，因而不需要再对各矫正平均数进行多重比较。

本试验的基本结论可总结如下：

- 不同的施肥期和施肥量，对水稻单位面积上的颖花数和结实率都有极显著的影响
- 但是，颖花数和结实率有极显著的线性回归关系，将各处理每平方米的颖花数都矫正到同一水平，则不同的结实率没有显著差异
- 因此，本例中不同施肥期和施肥量，对水稻的结实率只有间接效应，而没有显著的直接效应，即不同的施肥期和施肥量造成了单位面积上颖花的差异，进而造成了结实率的差异。

一、协方差分析的数学模型

表 10-1 数据资料协方差分析的线性数学模型为

$$y_{ij} = \mu_y + \alpha_i + \beta(x_{ij} - \mu_x) + \epsilon_{ij}$$

式中：μ_x, μ_y 分别是 y 和 x 的总体平均数；α_i 是第 i 个处理的效应，β 是 y 依 x 的总体回归系数。移项可得

$$\begin{aligned} y_{ij} - \alpha_i &= \mu_y + \beta(x_{ij} - \mu_x) + \epsilon_{ij} \\ y_{ij} - \beta(x_{ij} - \mu_x) &= \mu_y + \alpha_i + \epsilon_{ij} \end{aligned}$$

式中 $y'_{ij} = y_{ij} - \alpha_i$ 表示剔除了处理效应，即误差项，此时协方差分析即为 y'_{ij} 与 x_{ij} 的线性回归分析。若令 $y''_{ij} = y_{ij} - \beta(x_{ij} - \mu_x)$，则其表示对观测值进行了回归矫正，此时协方差分析即为 y''_{ij} 的方差分析，并且是消除了 x_{ij} 不一致对 y_{ij} 影响后的方差分析——协方差分析是回归分析与方差分析的结合

二、协方差分析的 3 个基本假定

- x 是固定的变量，因而处理效应 α_i 是固定模型
- ϵ_{ij} 是独立的（与处理效应无关），且服从 $\mathcal{N}(0, \sigma^2_{y|x}$，从样本来说，即各处理的离回归方差 $s_{y|x} = \frac{Q_i}{n-2}$ 没有显著出阿姨，即离回归方差同质
- 各处理的（x，y）总体都是线性的，且具有共同的回归系数 β，因而各处理总体的回归是一组平行的直线。对样本来说，各误差项的回归系数本身是显著的，但各回归系数 b_i 之间的差异不显著，即误差项的线性回归是显著的，而回归系数 b_i 的差异是不显著的。

- 应注意，我们介绍的协方差分析，并没有检验 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$，即没有检验各处理回归系数之间的差异显著性
- 这是因为我们的目的仅在于明确 y 与 x 是否有线性关系，以及在存在这一关系时矫正 y 的效果，所以省掉了回归系数 b_i 的齐次检验
- 同时，处理内回归显著亦包含了 b 间差异不显著的部分信息
- 但是，如果各 b 是否同质亦是研究目的，则该步骤不能省略。

Question?