Variance Components Estimation in Nested Designs

Miguel Fonseca[1], Vera de Jesus[2], João Tiago Mexia[3]

[1]New University of Lisbon – Faculty of Sciences and Technology, e-mail: fmig@fct.unl.pt
[2]Setúbal Polytechnic Institute – College of Business Administration, e-mail: vjesus@esce.ips.pt
[3]New University of Lisbon – Faculty of Sciences and Technology, e-mail: jtm@fct.unl.pt

SUMMARY

Estimation of error components in nested complete designs with any number of levels is presented, with an analysis of the negative estimators problem. It is also shown how it is possible to extend the results to the unbalanced case and an application to viniculture is presented.

1. Introduction

We consider the complete nested designs with n stages in the balanced case, in which each stage contributes with one experimental error component. It is assumed that the samples are random and that the last stage is constituted by the observations. The negative estimators problem is analyzed in order to provide a correct interpretation of the results.

An extension to unbalanced complete nested designs will be presented.

Lastly, a practical application of the obtained results is performed.

2. Complete Nested Designs - Balanced Case

In order to obtain a balanced nested design (see Scheffé, 1959) it is assumed that each of the J_0 is divided into J_1 sub-samples, and so on. In i-th stage, $i = 2,...,n-1$, each sub-sample of order i will be divided into J_i sub-samples of order $i+1$, $i = 2,...,n-1$. In the n-th stage there will be $\prod_{i=0}^{n-1} J_i$ sub-samples of order n, each one with J_n observations.

The assumed model for the observations is

$$Y_{j_0,j_1,\dots,j_n} = \mu + \sum_{t=0}^{n} e_{j_0,j_1,\dots,j_t};\ \ j_i = 1,\dots,J_i;\ \ i = 0,\dots,n, \tag{1}$$

where μ is the observational mean value, and the random variables $e_{j_0,j_1,\dots,j_i}$, $j_i = 1,\dots,J_i$, $i = 0,\dots,n$, represent the error components introduced by nesting. These variables are normal, independent, with null mean value and variance σ_i^2, $i = 0,1,\dots,n$.

Representing by $L_t = \prod_{i=t+1}^{n} J_i$ the number of observations in stage t, the average for the level t sub-samples and the overall mean are given by:

$$\begin{cases} \overline{Y}_{j_0,j_1,\dots,j_t} = \dfrac{1}{L_t}\displaystyle\sum_{j_{t+1}=1}^{J_{t+1}} \cdots \sum_{j_n=1}^{J_n} Y_{j_0,j_1,\dots,j_t} \\ \overline{Y} = \dfrac{1}{N}\displaystyle\sum_{j_0=1}^{J_0} \cdots \sum_{j_n=1}^{J_n} Y_{j_0,j_1,\dots,j_n} \end{cases}, \tag{2}$$

with N the total number of observations.

Representing the sum of squares of observations by S_n and the product of the number of observations on level t by the sum of squares of the level t averages,

$$\begin{cases} S_n = \displaystyle\sum_{j_0=1}^{J_0} \cdots \sum_{j_n=1}^{J_n} Y^2_{j_0,j_1,\dots,j_n} \\ S_t = L_t \displaystyle\sum_{j_0=1}^{J_0} \cdots \sum_{j_t=1}^{J_t} \overline{Y}^2_{j_0,j_1,\dots,j_t} \quad t = 0,..,n-1 \end{cases}, \tag{3}$$

and making $s_t = S_t - S_{t-1}$, $t = 0,\dots,n$, since $\sum_{j_n=1}^{J_n}\left(Y_{j_0,j_1,\dots,j_n} - \overline{Y}_{j_0,j_1,\dots,j_{n-1}}\right)^2$ is the product of σ_n^2 a chi-square with $(J_n - 1)$ degrees of freedom and the reproducibility of chi-squares, s_n will be the product of σ_n^2 by a chi-square with $g_n = \left(\prod_{h=0}^{n-1} J_h\right)(J_n - 1)$ degrees of freedom.

Representing the average for the level t by

$$\begin{aligned}&\overline{Y}_{j_0,j_1,\dots,j_t}=\mu+\sum_{h=0}^{t}e_{j_0,\dots,j_h}+\sum_{h=t+1}^{t}e^{(q)}_{j_0,\dots,j_t};\\&j_h=1,\dots,\ J_h;\ h=0,\dots,t;\ t=0,\dots,n-1,\end{aligned}\tag{4}$$

where $e^{(q)}_{j_0,\dots,j_t}$ is the average of the level t error components, follows the variance

$$V\left[\overline{Y}_{j_0,j_1,\dots,j_t}\right]=\sum_{h=0}^{t}\sigma_h^2+\sum_{q=t+1}^{n}\frac{L_q}{L_t}\sigma_q^2=\sigma_t^2+\sum_{q=t+1}^{n}\overline{\sigma}^2_{q,t};\ t=0,\dots,n.\tag{5}$$

Since

$$\begin{cases}\sum_{j_t=1}^{J_t}\left(\overline{Y}_{j_0,\dots,j_t}-\overline{Y}_{j_0,\dots,j_{t-1}}\right)^2\sim\overline{\sigma}^2_t\chi^2_{(J_t-1)}\\ \overline{\sigma}^2_t=\sigma_t^2+\sum_{q=t+1}^{n}\overline{\sigma}^2_{q,t}\end{cases},\tag{6}$$

and given the reproducibility of chi-squares:

$$s_t=L_t\sum_{j_0=1}^{J_0}\cdots\sum_{j_{t-1}=1}^{J_{t-1}}\sum_{j_t=1}^{J_t}\left(\overline{Y}_{j_0,\dots,j_t}-\overline{Y}_{j_0,\dots,j_{t-1}}\right)^2\sim\alpha_t\chi^2_{(g_t)},\tag{7}$$

where

$$\alpha_t=\sum_{h_t}^{n}L_h\sigma_h^2\ \text{ and }\ g_t=\left(\prod_{h=0}^{t-1}J_h\right)(J_t-1);\ t=0,\dots,n.\tag{8}$$

Unbiased estimators are then obtained:

$$\hat{\sigma}_n^2=\frac{s_n}{g_n}\ \text{ and }\ \hat{\alpha}_t=\frac{s_t}{g_t};\ t=0,\dots,n-1,\tag{9}$$

for σ_n^2 and α_t, $t=0,\dots,n-1$, respectively. Because $\sigma_t^2=\frac{\alpha_t-\alpha_{t+1}}{L_t}$, $t=0,\dots,n-1$, follows the unbiased estimator:

$$\hat{\sigma}_t^2=\frac{\hat{\alpha}_t-\hat{\alpha}_{t+1}}{L_t}.\tag{10}$$

3. Estimator Variance and Negative Estimators

As shown in applications to come, negative estimates can be obtained for σ_t^2, $t=0,\dots,n-1$.

One approach to this problem is to take restricted estimators (see Lehmann, Casella, 1998), but this only works with large samples. Another approach is to take the negativeness of the estimator as an indicator of nullity for the variance component.

In order to evaluate the probabilities associated with such estimates, set

$$\alpha_n=\sigma_n^2 \text{ and } s_t \sim \alpha_t\chi^2_{(g_t)};t=0,\dots,n-1. \tag{11}$$

Then,

$$\begin{cases} V\left[\hat{\sigma}_n^2\right]=V\left[\hat{\alpha}_n\right]=\dfrac{2\sigma_n^4}{g_n} \\ V\left[\hat{\sigma}_t^2\right]=\dfrac{2}{L_t}\left(\dfrac{\alpha_t}{g_t}+\dfrac{\alpha_{t+1}}{g_{t+1}}\right); \quad t=0,\dots,n-1 \end{cases}. \tag{12}$$

As for the probability of having negative estimates, from the expression of σ_t^2 it is easy to see that

$$P\left[\hat{\sigma}_t^2<0\right]=P\left[\hat{\alpha}_t<\hat{\alpha}_{t+1}\right]=P\left[F_t<\frac{\alpha_{t+1}}{\alpha_{t+1}+L_t\sigma_t^2}\right];\ t=0,\dots,n-1, \tag{13}$$

where F_t has an F distribution with g_t and g_{t+1} degrees of freedom.

From the expression of α_t, it is also possible observe that the negativeness of $\hat{\sigma}_t^2$ indicates that σ_t^2 is dominated by $\sigma_{t+1}^2,\dots,\sigma_n^2$.

4. Complete Nested Designs - Unbalanced Case ($n=2$ levels)

Consider the linear model

$$Y_{j_0,j_1,j_2}=\mu+e_{j_0}+e_{j_0,j_1}+e_{j_0,j_1,j_2}, \tag{14}$$

where μ is the observation's mean value, e_{j_0} the sampling error, e_{j_0,j_1} the sub-sampling error and e_{j_0,j_1,j_2} the error between observations.

The sub-sample mean will be given by

$$\overline{Y}_{j_0,j_1} = \mu + e_{j_0} + e_{j_0,j_1} + e^{(2)}_{j_0,j_1}, \tag{15}$$

with $e^{(2)}_{j_0,j_1}$ the average of the error components.

$\overline{Y}_{j_0,j_1}$ will have mean value μ and variance $\sigma_0^2 + \sigma_1^2 + \frac{\sigma_2^2}{n_{j_0,j_1}}$, where n_{j_0,j_1} is the sub-sample size.

Since $\sum_{j_2=1}^{n_{j_0,j_1}} \left(Y_{j_0,j_1,j_2} - \overline{Y}_{j_0,j_1}\right)^2$ is the product of σ_2^2 by a chi-square with n_{j_0,j_1} degrees of freedom, given the reproducibility of chi-squares,

$$s_2 = \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} \sum_{j_2=1}^{n_{j_0,j_1}} \sum_{j_2=1}^{n_{j_0,j_1}} \left(Y_{j_0,j_1,j_2} - \overline{Y}_{j_0,j_1}\right)^2 \tag{16}$$

will be the product of σ_2^2 by a chi-square with $g = \sum_{j_0}^{J_0} \sum_{j_1}^{n_{j0}} (n_{j_0,j_1} - 1) = n - n^{(1)}$ degrees of freedom, with $n = \sum_{j_0}^{J_0} \sum_{j_1}^{n_{j0}} n_{j_0,j_1}$ and $n^{(1)} = \sum_{j_0}^{J_0} n_{j_0}$, and therefore,

$$E[s_2] = g\sigma_2^2, \tag{17}$$

thus the unbiased estimator for σ_2^2:

$$\hat{\sigma}_2^2 = \frac{s_2}{g}. \tag{18}$$

Representing the number of observation taken in sample j_0 by

$$m_{j_0} = \sum_{j_1}^{n_{j0}} n_{j_0,j_1};\ j_0 = 1,\dots,J_0 \tag{19}$$

and the fraction of the observations from sample j_0 in sub-sample (j_0, j_1) by

$$h_{j_0,j_1} = \frac{n_{j_0,j_1}}{m_{j_0}};\ j_1 = 1,\dots,n_{j_0};\ j_0 = 1,\dots,J_0, \tag{20}$$

the averages for sample j_0 will be

$$\overline{Y}_{j_0} = \sum_{j_1=1}^{n_{j0}} h_{j_0,j_1} \overline{Y}_{j_0,j_1};\ j_0 = 1,\dots,J_0; \tag{21}$$

$$e_{j_0,\cdot} = \sum_{j_1=1}^{n_{j0}} h_{j_0,j_1} e_{j_0,j_1};\ j_0 = 1,\dots,J_0; \tag{22}$$

$$e_{j_0,\cdot}^{(2)} = \sum_{j_1=1}^{n_{j0}} h_{j_0,j_1} e_{j_0,j_1}^{(2)};\ j_0 = 1,\dots,J_0. \tag{23}$$

With

$$\overline{Y}_{j_0,j_1} - \overline{Y}_{j_0} = (e_{j_0,j_1} - e_{j_0,\cdot}) + (e_{j_0,j_1}^{(2)} - e_{j_0,\cdot}^{(2)}) \tag{24}$$

comes that

$$E\left[(\overline{Y}_{j_0,j_1} - \overline{Y}_{j_0})^2\right] = V\left[e_{j_0,j_1} - e_{j_0,\cdot}\right] + V\left[e_{j_0,j_1}^{(2)} - e_{j_0,\cdot}^{(2)}\right]; j_0 = 1,\dots,J_0. \tag{25}$$

Denoting the sum of squares of deviations by

$$s_1 = \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} (\overline{Y}_{j_0,j_1} - \overline{Y}_{j_0})^2, \tag{26}$$

with

$$\begin{cases} k_1 = \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} (1 - h_{j_0,j_1})^2 + (n_{j_0} - 1) \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} h_{j_0,j_1}^2 \\ k_2 = \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} \frac{(1 - h_{j_0,j_1})^2}{n_{j_0,j_1}} + (n_{j_0} - 1) \sum_{j_0=1}^{J_0} \sum_{j_1=1}^{n_{j0}} \frac{h_{j_0,j_1}^2}{n_{j_0,j_1}} \end{cases}, \tag{27}$$

follows the mean value

$$E[s_1] = k_1 \sigma_1^2 + k_2 \sigma_2^2 \tag{28}$$

and the unbiased estimator of σ_1^2:

$$\hat{\sigma}_1^2 = \frac{s_1 - \frac{k_2}{g} s_2}{k_1}. \tag{29}$$

With

$$q_{j_0} = \frac{m_{j_0}}{n}; j_0 = 1,\dots,J_0 \tag{30}$$

the fraction of the observations corresponding to level j_0 the general mean and error mean are:

$$\overline{Y} = \sum_{j_0=1}^{J_0} q_{j_0} \overline{Y}_{j_0}, \tag{31}$$

$$e_{.} = \sum_{j_0=1}^{J_0} q_{j_0} e_{j_0}, \tag{32}$$

$$e_{.,\cdot} = \sum_{j_0=1}^{J_0} q_{j_0} e_{j_0,\cdot}, \tag{33}$$

$$e_{.,\cdot}^{(2)} = \sum_{j_0=1}^{J_0} q_{j_0} e_{j_0,\cdot}^{(2)}. \tag{34}$$

Since

$$\overline{Y}_{j_0} - \overline{Y} = (e_{j_0} - e_{.}) + (e_{j_0,\cdot} - e_{.,\cdot}) + (e_{j_0,\cdot}^{(2)} - e_{.,\cdot}^{(2)}), \tag{35}$$

it follows that

$$E\left[\sum_{j_0=1}^{J_0} (\overline{Y}_{j_0} - \overline{Y})^2\right] = V\left[\sum_{j_0=1}^{J_0} e_{j_0} - e_{.}\right] + V\left[\sum_{j_0=1}^{J_0} e_{j_0,\cdot} - e_{j_0,\cdot}\right] + V\left[\sum_{j_0=1}^{J_0} e_{j_0,\cdot}^{(2)} - e_{j_0,\cdot}^{(2)}\right]. \tag{36}$$

With

$$\begin{cases} k_0 = \sum_{j_0=1}^{J_0} (1-q_{j_0})^2 + (J_0-1)q_{j_0}^2 \\ k_1 = \sum_{j_0=1}^{J_0} \dfrac{(1-q_{j_0})^2 + (J_0-1)q_{j_0}^2}{m_{j_0}^2} \sum_{j_1=1}^{n_{j_0}} n_{j_0,j_1}, \\ k_2 = \sum_{j_0=1}^{J_0} \dfrac{1}{m_{j_0}} \left((1-q_{j_0})^2 + (J_0-1)q_{j_0}^2\right) \end{cases} \tag{37}$$

the mean value

$$E\left[\sum_{j_0=1}^{J_0} (\overline{Y}_{j_0} - \overline{Y})^2\right] = k_0\sigma_0^2 + k_1\sigma_1^2 + k_2\sigma_2^2 \tag{38}$$

is obtained, as well as the unbiased estimator

$$\hat{\sigma}_0^2 = \frac{\sum_{j_0=1}^{J_0} (\overline{Y}_{j_0} - \overline{Y})^2 - k_1\hat{\sigma}_1^2 - k_2\hat{\sigma}_2^2}{k_0}. \tag{39}$$

5. Application

In the following application, real data from a company that produces certified classes of grapevines was used.

Four castes were used: Aragonęs, Trincadeira, Touriga Nacional and Arinto, and for each one two different clones with different number of observations.

All observations were registered in the same farm and in the same year.

The obtain results are presented in table 1.

Table 1. Data on castes

Cast	Clone	N (n_{j_0,j_1})	Average $\left(\overline{Y}_{j0,j1}\right)$	Standard Deviation $\left(\hat{\sigma}_{j0,j1}\right)$
Aragones (C1)	234	24	5338	1938
	238	9	3889	2073
Trincadeira (C2)	46	25	7120	2237
	47	24	3100	1239
Touriga Nacional (C3)	378	20	5400	2505
	379	9	3922	2207
Arinto (C4)	536	22	1832	1011
	538	15	4393	2116

Assuming the model

$$Y_{j_0,j_1,j_2} = \mu + e_{j_0} + e_{j_0,j_1} + e_{j_0,j_1,j_2}, \tag{40}$$

follows the sum of squares

$$s_2 = \sum_{j_0=1}^{4}\sum_{j_1=1}^{2}\sum_{j_2=1}^{n_{j_0,j_1}} (Y_{j_0,j_1,j_2} - \overline{Y}_{j_0,j_1})^2 = \sum_{j_0=1}^{4}\sum_{j_1=1}^{2} (n_{j_0,j_1} - 1)\hat{\sigma}^2_{j_0,j_1} \tag{41}$$
$$= 518554558.$$

Also,

$$g = n - n^{(1)} = \sum_{j_0=1}^{4}\sum_{j_1=1}^{2} n_{j_0,j_1} - \sum_{j_0=1}^{4} n_{j_0} = 140, \tag{42}$$

getting the unbiased estimator for the between observations variance component:

$$\hat{\sigma}_2^2 = \frac{s_2}{g} = 3703961.1. \tag{43}$$

With h_{j_0,j_1} the fraction of observations corresponding to clones, the averages for castes will be

$$\overline{Y}_{j_0} = \sum_{j_1=1}^{2} h_{j_0,j_1} \overline{Y}_{j_0,j_1} = \begin{cases} \overline{Y}_1 = 4942.818 \\ \overline{Y}_2 = 5151.02 \\ \overline{Y}_3 = 4941.31 \\ \overline{Y}_4 = 2870.838 \end{cases}, \tag{44}$$

along with the deviations sum of squares

$$s_1 = \sum_{j_0=1}^{4} \sum_{j_1=1}^{2} (\overline{Y}_{j_0,j_1} - \overline{Y}_{j_0})^2 = 13993741.6, \tag{45}$$

with

$$k_1 = \sum_{j_0=1}^{4} k_{j_0,1} = 4.387 \text{ and } k_2 = \sum_{j_0=1}^{4} k_{j_0,2} = 0.283, \tag{46}$$

getting the unbiased estimator for clones

$$\hat{\sigma}_1^2 = \frac{s_1 - \frac{k_2}{g} s_2}{k_1} = 2950882.289. \tag{47}$$

For the general mean

$$\overline{Y} = \sum_{j_0=1}^{4} q_{j_0} \overline{Y}_{j_0} = 4493.459, \tag{48}$$

with

$$k_0 = 3.514,\ k_1 = 1.666 \text{ and } k_2 = 0.085, \tag{49}$$

the unbiased estimator for castes

$$\hat{\sigma}_0^2 = \frac{\sum_{j_0=1}^{4} (\overline{Y}_{j_0} - \overline{Y})^2 - k_1 \hat{\sigma}_1^2 - k_2 \hat{\sigma}_2^2}{k_0} = -580595.38 \quad (50)$$

is derived.

As seen before, the negative value of $\hat{\sigma}_0^2$ indicates that the variation between castes is dominated by the variation between clones, which can in part be explained by the aging of the castes through clone separation. Applying Khuri's method for random models with unbalanced cell frequencies in the last stage (see Khuri et al. 1997, chapter 5), the estimated probability in (13) is 0.99, indicating the predominance of the other variance components over σ_0^2.

REFERENCES

Khuri A.I., Mathew T., Sinha B.K., (1997). *Statistical Tests for Mixed Linear Models.* Wiley

Lehmann E.L., Casella G. (1998). *Theory of Point Estimation. Second Edition.* Springer

Scheffé H. (1959). *The Analysis of Variance.* John Wiley & Sons