Solution to exercise 24

Question 1:

Model I: Deterministic effect analysis of variance (ANOVA).

If the 5 localities are modeled as deterministic for example characterized as being the only ones we are interested in the following model could be chosen

$$Y_{i,j} = \mu + \alpha_i + Z_{i,j} \quad , \quad \sum_{i=1}^{5} \alpha_i = 0 \quad , \quad Z_{i,j} \in NID(0, \sigma_Z^2)$$

The notation $NID(\mu, \sigma^2)$ means "Normally and Independently Distributed" with mean μ and variance σ^2.

The investigation of whether the 5 localities are different or not with respect to content of fluoride is the same as testing the hypothesis $\alpha_1 = \cdots = \alpha_5 = 0$. This is done by means of a usual one-way analysis of variance for deterministic effects:

Source of variation	SSQ	d.f.	s^2	$E\{s^2\}$	F–value
Localities	5.764	5–1	1.441	σ_Z^2+const$\sum_i \alpha_i^2$	13.67
Residual variation	1.054	10	0.1054	σ_Z^2	
Total	6.818	14			

The critical value for the F–test quantity is $F > F(4, 10)_{0.95} = 3.48$. Thus a significant variation between localities is found. The α's must therefore be considered different.

The following estimates are then computed:

$\hat{\sigma}_Z^2 = 0.1054 = 0.325^2$, $\hat{\mu} = \bar{y}_{.,.} = 1.35$ and

Localities	1	2	3	4	5
$\hat{\alpha}_i = \bar{y}_{i,.} - \bar{y}_{.,.} =$	–0.48	0.82	–0.61	0.69	–0.41

Model II: Random effect ANOVA

If the 5 localities are modeled as random, characterized for example as being 5 more or less randomly selected among a large number of localities, the following model could be

relevant

$$Y_{i,j} = \mu + L_i + Z_{i,j} \quad , \quad L_i \in NID(0, \sigma_L^2) \quad , \quad Z_{i,j} \in NID(0, \sigma_Z^2)$$

The investigation is concerned with testing whether the variation between the localities is small compared to the variation within them. Thus we want to test whether $\sigma_L^2 = 0$.

This is done by means of a usual one-way analysis of variance for random effects:

Source of variation	SSQ	d.f.	s^2	$E\{s^2\}$	F–value
Localities	5.764	5–1	1.441	$\sigma_Z^2 + 3\sigma_L^2$	13.67
Residual variation	1.054	10	0.1054	σ_Z^2	
Total	6.818	14			

The critical value for the F–test quantity is $F > F(4, 10)_{0.95} = 3.48$. Thus a significant variation between localities is found and we conclude that $\sigma_L^2 > 0$.

We now find the following estimates for the two components of variance in the model:

$$\hat{\sigma}_Z^2 = 0.1054 = 0.325^2 \text{ and } \hat{\sigma}_L^2 = (s_L^2 - s_Z^2)/3 = 0.4452 = 0.667^2$$

Note that, allthough the computations in the two ANOVA tables are the same, the two models considered are very different. Which of these models is relevant is determined by the concrete problem at hand.

The estimates are different: In the first model individual effects for each locality are computed, but in the second case we compute the variance between localities describing a general random variation.

Question 2:

In this question we do not take into account any a priori informations about particular differences between certain localities. Thus the random effect model is relevant.

In this case it is reasonable to use a test based on the range to assess whether there is a grouping of the localities. The Duncans Multiple Range test or the Newmam-Keuls test can do this. Here we choose the Newman-Keuls test which , in fact, is a little more conservative (less likely to point out a false grouping).

We need the standard deviation corresponding to the average of one group (locality):

$$s_{\bar{y}}^2 = s_y^2/3 = 0.1054/3 \simeq 0.19^2$$

By means of the table for the 5% 'Studentized Range Statistics' critical values are found as 'least significant ranges' $= q \cdot s_{\bar{y}}$, using that the degrees af freedom for the residual variation is $f = 10$:

p =	2	3	4	5
critical value $= q$	3.15	3.88	4.33	4.66
$q \cdot s_{\bar{y}} = q \cdot 0.19 =$	0.60	0.74	0.82	0.89

The averages for the 5 localities are ordered and compared with this table as follows:

$\bar{y}_3$	$\bar{y}_1$	$\bar{y}_5$	$\bar{y}_4$	$\bar{y}_2$
0.73	0.87	0.93	2.03	2.17

The largest versus the smallest: $(2.17 - 0.73) = 1.44 > 0.89$, next $(2.03 - 0.73) = 1.30 > 0.82$, but $(0.93 - 0.73) = 0.20 < 0.74$. The same procedure from the other end: The smalles vesus the largest $(2.17 - 0.73) = 1.44 > 0.89$, next $(2.17 - 0.87) = 1.30 > 0.82$, and $(2.17 - 0.93) = 1.24 > 0.74$, but $(2.17 - 2.03) = 0.14 < 0.60$ (do a suitable drawing and convince yourself that the result is reasonable).

It is concluded that the 5 localities can be grouped in two groups containing localities (2 and 4) respectively (5, 1 and 3).

Question 3:

If we have the information given in this question it would be natural to use the deterministic effect model and to split the variation between localities according to suitably chosen orthogonal contrasts.

Especially the following hypothesis could be relevant to consider:

$$H_0 : 3(\alpha_2 + \alpha_4) = 2(\alpha_1 + \alpha_3 + \alpha_5)$$

by means of the contrast

$$C = 3(T_2 + T_4) - 2(T_1 + T_3 + T_5) = 3(6.5 + 6.1) - 2(2.6 + 2.2 + 2.8) = 22.60$$

with sum of squares

$$\mathrm{SSQ}_C = 22.60^2/(3^2(3 + 3) + 2^2(3 + 3 + 3)) = 5.675$$

This gives the F–test quantity:

$$F = 5.675/s^2_{rest} = 5.675/0.1054 = 53.9 >> F(1,10)_{0.95} = 4.96$$

A significant difference between the two groups of localities is thus found, confirming our 'a priori' idea, that they could be different for reasons as described.

The remaning variation between localities (that is the variation within the two groups) is (5.764–5.675)=0.089 with (4–1) degrees of freedom. This can be illustrated in the following modified ANOVA table:

Source of variation	SSQ	d.f.	s^2	F–value
Ground water/surface w.	5.675	1	5.675	53.84
Within groups	0.089	3	0.030	0.28
Residuatl variation	1.054	10	0.1054	
Total	6.818	14		

We note that the variation within the two types of water is not significant (far from, actually) and we may therefore pool the corresponding variance with the residual variance to obtain an improved estimate:

$$\hat{\sigma}^2_Z = (1.054 + 0.089)/(10 + 3) = 0.088 \simeq 0.30^2$$

having 13 degrees of freedom.

The average difference between the two types of water is estimated from the averages of the two groupes:

$(2.03 + 2.17)/2 - (0.73 + 0.87 + 0.93)/3 = 1.26$.

(one more page)

SAS program for exercise 24:

```
Title SAS code for exercise 24;
options linesize=76;

data ex24;
do local = 1 to 5;
do r = 1 to 3 ;
input y @ ; output;
end;
end;
cards;
1.1 0.7 0.8
2.0 2.0 2.5
0.5 0.7 1.0
1.6 2.1 2.4
1.4 0.8 0.6
;
proc print;

Title2  Random effect model;run;
proc glm ; class local;
model y=local;
means local / snk ; /* snk = Newman-Keuls test on local */
random local      ; /* declare local random */
run;

Title2  Fixed effect model;run;
proc glm ; class local;
model y=local;
contrast '3(2,4)-2(1,3,5)'   local -2 3 -2 3 -2 ;
contrast '(2)-(4)      '   local 0 1 0 -1 0 ;
contrast '(1)-(5)      '   local 1 0 0 0 -1 ;
contrast '(1)-2(3)+(5) '   local 1 0 -2 0 1 ;
run;
```