MULTIPLE MEASUREMENTS (SUBSAMPLING)

Multiple measurements

Sometimes multiple measurements are made on a single experimental unit.

Example. Sizes of zooplankton in six experimental ponds, three with fish and three without; 48 zooplankton measurements per pond (data set cep3).

Pond is the experimental unit.

This is a case where the smallest observational unit, fish, is not the experimental unit.

To correctly identify the experimental units, we have to see how the treatments are assigned:

- In this example, fish/no fish is the treatment, and the treatment varies at the pond level.

The 48 zooplankton measurements made on the fish within each pond are multiple measurements or subsampling.

Sometimes multiple measurements are made on a single experimental unit.

Let's look at an example.

Sizes of zooplankton in six experimental ponds, three with fish and three without; 48 zooplankton measurements per pond (data set cep3).

First note that in this example, pond is the experimental unit.

This is a case where the smallest observational unit, fish, is not the experimental unit.

To correctly identify the experimental units, we have to see how the treatments are assigned: In this example, fish/no fish is the treatment, and the treatment varies at the pond level.

The 48 zooplankton measurements made on the fish within each pond are multiple measurements or subsampling.

Model for multiple measurements

Model:

$$y_{ijk} = \mu + \tau_i + e_{ij} + d_{ijk} (i = 1, \dots, t; j = 1, \dots, r; k = 1, \dots, n)$$

where

- μ is the overall mean;
- τ_i is the fixed effect of treatment i;
- e_{ij} is the random error for experimental unit j in treatment i; and
- d_{ijk} is the random error for subsample k in unit j of treatment i.

Model assumptions: $e_{ij} i.i.d. N(0, \sigma_e^2)$, $d_{ijk} i.i.d. N(0, \sigma_d^2)$, e_{ij}'s and d_{ijk}'s are independent.

Note that we need two random terms e_{ij} and d_{ijk} in the regression equation.

(Recall that the "error" in a regression model means data variation not explained by the fixed part, $\mu + \tau_i$, of the model.)

To model the multiple measurements or subsampling, we need two random terms in the regression equation: one is the random error for the experimental unit, the other is the random error for subsample.

The random error terms are denoted by (e i j) and (d i j k) here. They are assumed iid normal with variances (sigma e squared) and (sigma d squared) respectively.The two types of error terms are assumed independent.

Recall that the "error" in a regression model means data variation not explained by the fixed part of the model.

Decomposition of the sum of squares

The total variability (SS Total) can be partitioned as follows:

$$\text{SS Total} = \text{SS Treatment} + \text{SS Error} + \text{SS Sampling}$$

See the ANOVA table for the zooplankton data on the next page.

The important thing to remember, with multiple measurements or sub-sampling, **MSE corresponding to between-treatment variation is the correct denominator to use in an F-test or t-test for treatment comparison**, not MS Sampling (corresponding to between-measurement variation within the same experimental unit)!

The total variability, SS Total, can be partitioned as follows:

(SS Treatment + SS Error + SS Sampling)

See the ANOVA table for the zooplankton data on the next page.

The important thing to remember, with multiple measurements or sub-sampling, MSE corresponding to between-treatment variation is the correct denominator to use in an F-test or t-test for treatment effect, not MS Sampling (corresponding to between-measurement variation within the same experimental unit)!

The ANOVA table for the zooplankton data

	df	SS	MS
Treaments	$t-1$	SS Treatment	MST
(Fish)	1	7.009	7.009
Error	$t(r-1)$	SSE	MSE
(Pond)	4	0.503	0.126
Sampling	$tr(n-1)$	SSS	MSS
	282	13.080	0.046
Total	$trn-1$	SS Total	
	287	20.592	

t=2, r=3, n=38

For the zooplankton data, the F-test for the fish effect should use the MSE corresponding to the pond as the denominator.

Fit this model in R

To fit this model in R, you must use a formula with nesting (using the "%in%" operator):

```
fish.f = factor(fish);
pond.f = factor(pond);
lm(size ~ fish.f + pond.f %in% fish.f)
```

To test whether treatment means differ, use $F_0 = MST/MSE$.

In the example, $F_0 = 7.009/0.126 = 55.78$, and p-value $= Pr(F_{1,4} > 55.78) = 0.0017$.

The 1 numerator d.f. corresponds to the fish/no-fish treatment and 4 denominator d.f. correspond to the six ponds (with two estimated means).

We have strong evidence of an association between zooplankton size and the presence/absence of fish.

See the R script for this example, `cep3.html`, for more details. There we presented different ways to get the correct test for this data sets.

To fit this model in R, you must use a formula with nesting.

To test whether treatment means differ, use (MST over MSE) as the F test statistic.

In this example, the F statistic value equals 7.009/0.126 equals 55.78. Comparing this value to a F distribution with 1 and 4 degrees of freedom gives a p-value of 0.0017.

The 1 numerator d.f.corresponds to the fish/no-fish treatment and 4 denominator d.f. correspond to the six ponds (with two estimated means).

We have strong evidence of an association between zooplankton size and the presence/absence of fish.

See the R script for this example, cep3.html, for more details. There we presented different ways to get the correct test for this data set.

The expected mean squares on p. 162 of Kuehl.

Table 5.5 Analysis of variance for the completely randomized design with subsamples[2]

Source of Variation	*Degrees of Freedom*	*Sum of Squares*	*Mean Square*	*Expected Mean Square*
Total	$trn - 1$	*SS* Total		
Treatments	$t - 1$	*SST*	*MST*	$\sigma_d^2 + n\sigma_e^2 + rn\theta_t^2$
Error	$t(r - 1)$	*SSE*	*MSE*	$\sigma_d^2 + n\sigma_e^2$
Sampling	$tr(n - 1)$	*SSS*	*MSS*	σ_d^2

$$SS\text{ Total} = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{...})^2$$

$$SST = SS\text{ Treatment} = rn\sum_{i=1}^{t}(\overline{y}_{i..} - \overline{y}_{...})^2$$

$$SSE = SS\text{ Error} = n\sum_{i=1}^{t}\sum_{j=1}^{r}(\overline{y}_{ij.} - \overline{y}_{i..})^2$$

$$SSS = SS\text{ Sampling} = \sum_{i=1}^{t}\sum_{j=1}^{r}\sum_{k=1}^{n}(y_{ijk} - \overline{y}_{ij.})^2$$

Here we see the ANOVA table corresponding to the a one-way ANOVA design with subsampling.

In the column of “expected mean squares”, we see the expected value of the MST is (sigma d squared + n sigma e squared + r n theta t squared).

Recall that (sigma e squared) corresponds to the random error for experimental unit and (sigma d squared) correspond to the random error subsample.

Importantly, (theta t squared) measures the treatment mean differences, and (theta t squared) is 0 if all treatment means are equal.

So we see that if all treatment means are the same, MST = MSE. That is why MSE should be used as the denominator is the F-test for equal treatment means.

Summary

With multiple measurements or sub-sampling, it is important to know what is the correct denominator to use in an F-test or t-test for treatment comparisons:

It is the between-treatment variation that matters here. Not the between-measurement variation within the same experimental unit.

The expected values of the mean squares in an ANOVA table provide hints on what is the correct term to use in the denominator of an F-test.

With multiple measurements or sub-sampling, it is important to know what is the correct denominator to use in an F-test or t-test for treatment comparisons:

It is the between-treatment variation that matters here. Not the between-measurement variation within the same experimental unit.

The expected values of the mean squares in an ANOVA table provide hints on what is the correct term to use in the denominator of an F-test.