TWO-WAY INTERACTION WHEN THERE IS A LINEAR RELATIONSHIP AMONG THE PAIRS OF MEANS IN A FACTORIAL EXPERIMENT

by Donald W. Seegrist, Leader, Biometrics Group,
USDA Forest Service, Northeastern Forest Experiment
Station, Upper Darby, Pennsylvania 19082

INTRODUCTION

SUPPOSE WE HAVE a genetic experiment with J families planted at two locations and find that the best families at one location are the best families at the second location. In other words, we find that there is a positive relationship among the J pairs of family means. An example of a positive relationship among means is shown in Figure 1 using maple sugar content for 25 families at two locations: Proctor, Vermont; and Hopkins, Massachusetts.

Will there be location x family interaction when the best families at one location are also the best at the second location? The answer depends on the form of the relationship among the means. To show this, we have to know what is meant by "no location x family interaction".

Location x family interaction when there is a relationship among the pairs of family means. --"No interaction" is a special case of a linear relationship among the family means. To illustrate, I write the linear relationship among the means as

$$H_0 : u_{1j} = \alpha + \beta u_{2j}$$

$$j = 1, \ldots, J$$

where u_{ij} = true mean of j^{th} family at i^{th}

The hypothesis can be written in a different form to bring out the relationship between the linear relationship and "no interaction". Taking the J-1 differences between family 1 and the other families we have

$$H_0 : u_{11} - u_{12} = \beta(u_{21} - u_{22})$$

$$\cdots$$

$$u_{11} - u_{1J} = \beta(u_{21} - u_{2J}).$$

How is H related to the hypothesis of no interaction? In statistical jargon, we say there is no location x family interaction when the J-1 pairs of line segments in the "profile of the means" are parallel. Figure 2 shows the profile of the means of the 25 families at Proctor and Hopkins. The J-1 pairs of line segments are parallel when

$$H_1 : u_{11} - u_{12} = u_{21} - u_{22}$$

$$\cdots$$

$$u_{11} - u_{1J} = u_{21} - u_{2J}.$$

Comparing H_1 with H_0, we see that there is "no interaction" when the relationship among the means is linear and the slope (B) is equal to 1.0.

The conditions for parallel line segments can be understood most readily by examining the profile of the means of two families at two locations (1 and 2) (Fig. 3). The line segment between u_{11}, and u_{12} is parallel to the segment between u_{21} and u_{22} when

$$u_{11} - u_{12} = u_{21} - u_{22}.$$

The condition can also be written as

$$u_{11} - u_{21} = u_{12} - u_{22}.$$

In analysis of variance, the alternative hypothesis (H_A) to "no two-way interaction" is that at least one of the conditions u_{11} - u⊟uu $u_{11}-u_{1j}=u_{21}-u_{2j}$ is not true. The alternate hypothesis says nothing about a possible relationship among the means. There could be no relationship; in which case, the slope S would be zero. There could be a linear relationship, but the slope is not equal to 1.0. Another possibility is that the relataionship is not linear.

Family selection when there *is* a relationship among the family means.--In most genetic studies, one would consider family selection if there is a positive relationship among the means, and if the character of interest has high heritability.

Most geneticists would probably carry out family selection if the ANOVA showed that there was no location x family interaction. *We* would select the superior families at one location (or both locations) and be reasonably confident that the offspring of selected families would be superior at the two (and possibly other) locations.

What do we do when there is location x family interaction? Family selection is feasible when- the interaction is such that the families at one location are also superior at the second location. I have demonstrated that H tells us nothing about the form of the interaction. A relationship among means can be demonstrated by plotting the means at one location against the means at the second location, or by comparing the ranks of families at the two locations.

Estimating the slope of a linear regression among the means.--Suppose the means are linearly related. In other words our model

$$u_{1j} = \alpha + \beta\ u_{2j}$$

holds. How should we estimate the slope ?

Ordinary regression on the means estimators.--The simplest way to estimate the slope (B) is to regress the family means at one location on the means at the other location. For example, one estimator of the slope is

$$\bar{\beta}_1 = \frac{\Sigma(\bar{y}_{1j} - \bar{\bar{y}}_1)(\bar{y}_{2j} - \bar{\bar{y}}_2).}{\Sigma(\bar{y}_{2j} - \bar{\bar{y}}_2)^2}$$

We have another estimator

$$\bar{\beta}_2 = \frac{\Sigma(\bar{y}_{2j} - \bar{\bar{y}}_2)(\bar{y}_{1j} - \bar{\bar{y}}_1).}{\Sigma(\bar{y}_{1j} - \bar{\bar{y}}_1)^2}$$

I have defined the linear relationship among the pairs of means as the regression of u_1 on u_2. Therefore we use the sample regression coefficient $\bar{\beta}_1$ to estimate β.

Are the Regression on the Means estimators good estimators of the slope? The ordinary regression estimators are the best linear unbiased estimators when the variables are known without error, which doesn't hold in our case. We know that sample means are variables with variances inversely proportional to the number of offspring that make up each location-family combination.

It is clear that we should not use the Regression on the Means standard error to set confidence intervals or to test hypotheses because the sample standard error does not take into account the variances of the means.

VARIANCE AND COVARIANCE COMPONENT ESTIMATORS

How should we estimate the slope among the family means? The slope depends on the variance of true family means within location, and the covariance between true family means at the two locations. I define the slope of the regression of the families at Location 1 on the families at Location 2 as

$$\beta(u_1/u_2) = \frac{\sigma(u_1, u_2)}{\sigma^2(u_2)}$$

where $\sigma(u_1, u_2)$ = the covariance component among means,

$\sigma^2(u_2)$ = variance component among means at Location 2.

Variance component estimates.--One possible model for the observations is to consider the families as nested within locations. The model can be written as

$$y_{ijk} = u_i + u^*_{ij} + e_{ijk}$$

for $i = 1,2$
$j = 1, \ldots, J$
$k = 1, \ldots, n_{ij}$

where y_{ijk} = value of k^{th} offspring of j^{th} family in i^{th} location

u_i = effect of i^{th} location

u^*_{ij} = effect of j^{th} family in i^{th} location

and e_{ijk} = "error" of k^{th} offspring of j^{th} family in i^{th} location.

The assumptions of the model are

$$\Sigma(e_{1jk}) = \Sigma(e_{2jk}) = \Sigma(u^*_{1j}) = \Sigma(u^*_{2j}) = 0$$

$$\Sigma(u^{*2}_{1j}) = \sigma^2(u_1),\ \Sigma(u^{*2}_{2j}) = \sigma^2(u_2)$$

$$\Sigma(e^2_{1jk}) = \Sigma(e^2_{2jk}) = \sigma^2_e.$$

The analysis of variance can be used to estimate the variance components $\sigma^2(u_1)$, $\sigma^2(u_2)$, andσ_e^2, which are denoted as $s^2(u_1)$, $s^2(u_2)$, and s_e^2, respectively.

Covariance component estimator.--How do we estimate the covariance component among means $a(u_1, u_2)$? Kempthorne (1957) discusses components of covariance, and introduces the subject by assuming that two values y and z are measured on each individual. The component of variance of (y+z) for the hth source of variatio is

$$\sigma_h^2(y + z) = \sigma_h^2(y) + \sigma_h^2(z) + 2\sigma_h(y,z).$$

Therefore the component of covariance is

$$\sigma_h(y,z) = 1/2\{\sigma_h^2(y + z) - \sigma_h^2(y) - \sigma_h^2(z)\}.$$

A simple method for estimating the component of covariance is to calculate $\hat{\sigma}_h^2(y + z)$, $\hat{\sigma}_h^2(y)$, and $\hat{\sigma}_h^2(z)$ by the analysis of variance, and to substitute the sample values into the above expression.

This procedure cannot be used if there is no correspondence between the y and z values at some level of classification. An example of unpaired variables would be an agricultural experiment to study the yield (y) and percent protein (z) for several varieties of grain. Suppose for each variety that r units are measured for yield, and r independent units are measured for percent protein. In this case, there is no covariance component within variety. The covariance among the variety means *is* estimated by the covariance among the sample means.

In our hypothetical genetic. experiment, the value of the offspring of each family measured at two locations is analogous to the two variables for each-variety measured on two sets of independent units. We have no covariance component within family. Our experiment is different in that we do not measure two different variables *(y)* and (z), but measure the same characteristic *(y)* at two locations. We have a bivariate situation because the experiment is repeated in space. The covariance among the sample means is

$$\hat{\sigma}(u_1, u_2) = \sum_{j=1}^{J} (\bar{y}_{1j} - \bar{\bar{y}}_1)(\bar{y}_{2j} - \bar{\bar{y}}_2)$$

The slope among means component estimator. --I suggest that we use an estimator of the slope (3) based on the covariance among sample means and the variance component estimator. The estimator is

$$\tilde{b}_1 = \frac{\hat{\sigma}(u_1, u_2).}{s^2(u_2)}$$

I believe that b is a better estimator of 3 than the Regression on the Means estimator 3 because b$_i$ has been adjusted for variation about the sample means. The statistical properties of b$_1$ need to be investigated.

The estimator B is larger than B_{1_1} because the denominator with component of variance s^2 (u) is less than the variance among the sample means s^{2}t7$_2$). In other words, the slope among the sample means underestimates B(u$_1$/u$_2$). It is said that 1 has been "adjusted for attenuation", a procedure that is well known in psychological testing theory (Walker and Lev 1953:305;

TESTING FOR A LINEAR RELATIONSHIP AMONG THE PAIRS OF FAMILY MEANS

How do we test the hypothesis that there is no linear relationship among the means? There is no linear relationship when the slope 3 is equal to zero, and the slope B is equal to zero when the covariance o(u$_1$, u$_7$) is equal to zero.

For the nested model, the covariance among the means is

$$\sigma(u_1, u_2) = \Sigma\ u^*_{1j}\ u^*_{2j}$$

which is a nonlinear function of the u^*_{ij}'s. If we replace the u^*_{1j}'s with their estimators ($\hat{u}^*_{1j}$), the hypothesis becomes

$$\Sigma\ \hat{u}^*_{1j}\ u^*_{2j} = 0$$

which is a linear combination of the u^*_{2j}'s.

The general linear hypothesis procedure can be used to test for linear combinations of parameters of a linear model. The procedure is well known. BIOMEDX63 can be used to compute the test statistics.

The sampling properties of the suggested test procedure need to be investigated. And I have used the standard statistical method for testing a linear hypothesis. The mean square error is the appropriate denominator for the F test. However the numerator of the F test would be a function of the sample means; this would introduce a bias into the testing procedure.

DISCUSSION

I have assumed that the variation among families was different at the two locations. One could test whether $\sigma^2(u_1) = \sigma^2(u_2)$ by analysis of variance. If the variance among families within location are the same, one could reanalyze the data under the assumption that $\sigma^2(u_1) = \sigma^2(u_2)$, and estimate the variance component among families by the ANOVA. Denote the estimator by $\hat{\sigma}^2(u.)$. In this case, we could estimate the slope by

$$\tilde{b}_3 = \frac{\hat{\sigma}(u_1,\ u2)}{\hat{\sigma}^2(u.)} .$$

What do we do if there are more than two locations? I would plot the family means for each pair of locations If the same families have high values at all locations, we will find that there is a positive relationship among the means for each pair of locations.

An analysis of the relationship among means should be carried out in any experiment design model that includes two-way interaction terms. For example, the two-way interactions in a three-factor experiment could be analyzed for a possible linear relation among pairs of means.

The analysis becomes more complex with increasing numbers of locations. Hopefully, one would find a number of superior families common to the locations. In that case, a family selection program could be carried out. If there are no superior families over all locations, the form of the locations x family interaction will greatly effect the breeding program.

The primary purpose of this paper is to show the relationships between "no two-way interaction" in a 2xJ two-factor experiment when there is a positive linear relation among the J pairs of means.

Also, I suggest a method of estimating the slope if such a linear relationship exists.

LITERATURE CITED

Kempthorne, O.
1957. AN INTRODUCTION TO GENETIC STATISTICS. J. Wiley and Sons, New York. 545 p.

Walker, H. M., and J. Lev.
1953. STATISTICAL INFERENCE. H. Holt and Co., New York. 510 p.

FIGURE LEGENDS

Figure I.--Family means at Proctor location versus that at Hopkins location.

Figure 2.--Profile of family means at Proctor and Hopkins locations.

Figure 3.--Profile of means for two families at two locations.

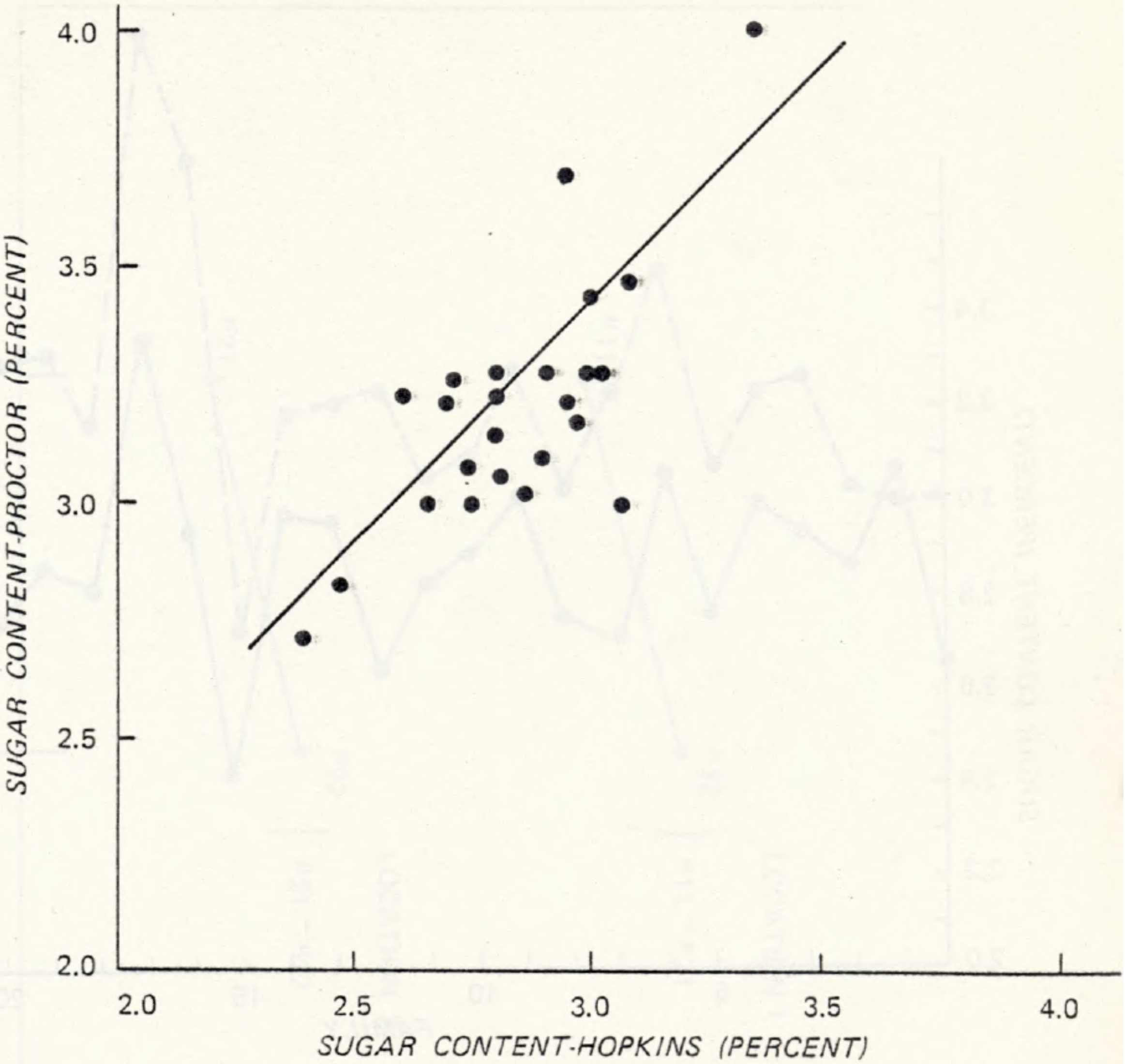

Figure 1. Family means at Proctor location versus that at Hopkins location.

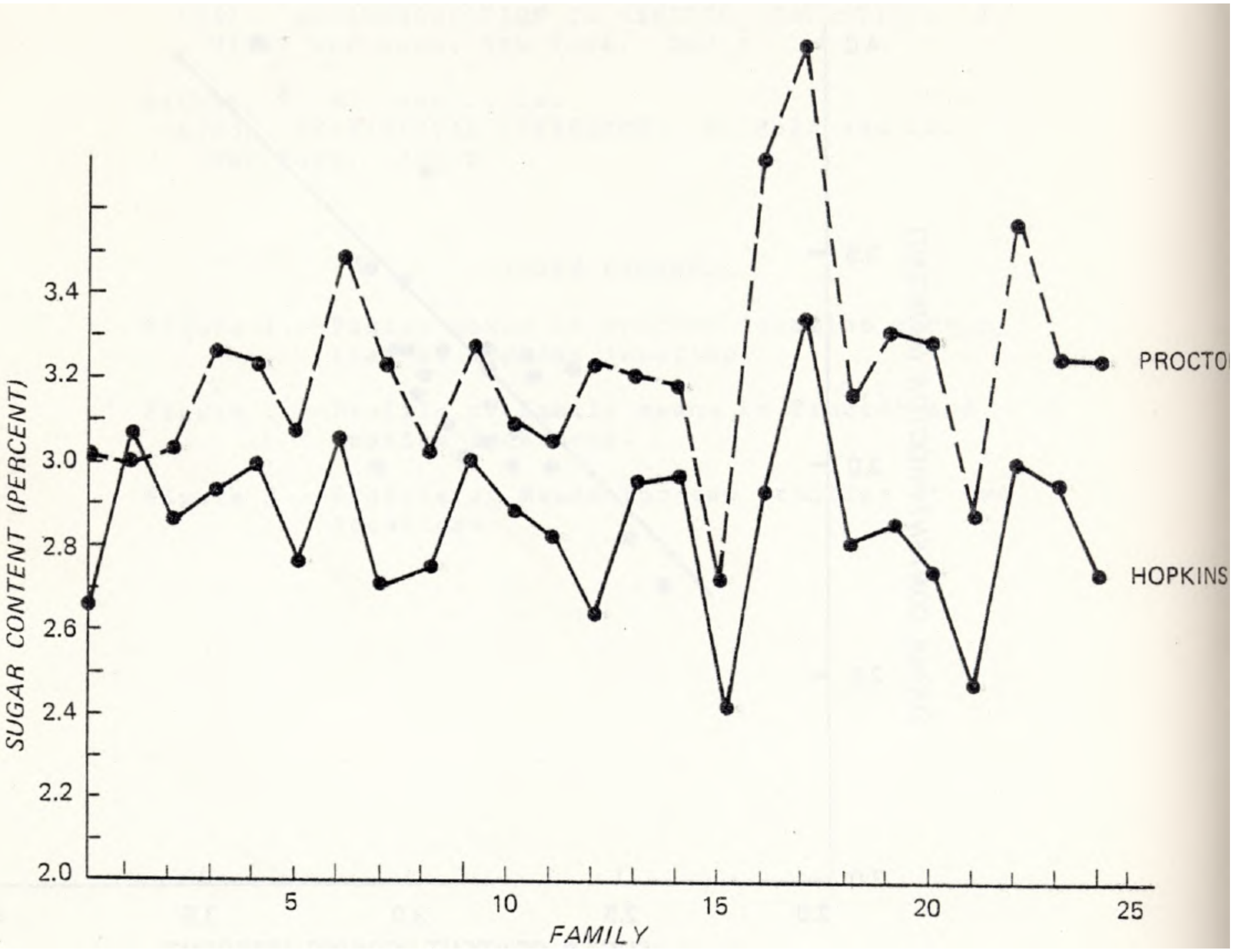

Figure 2. Profile of family means at Proctor and Hopkins locations.

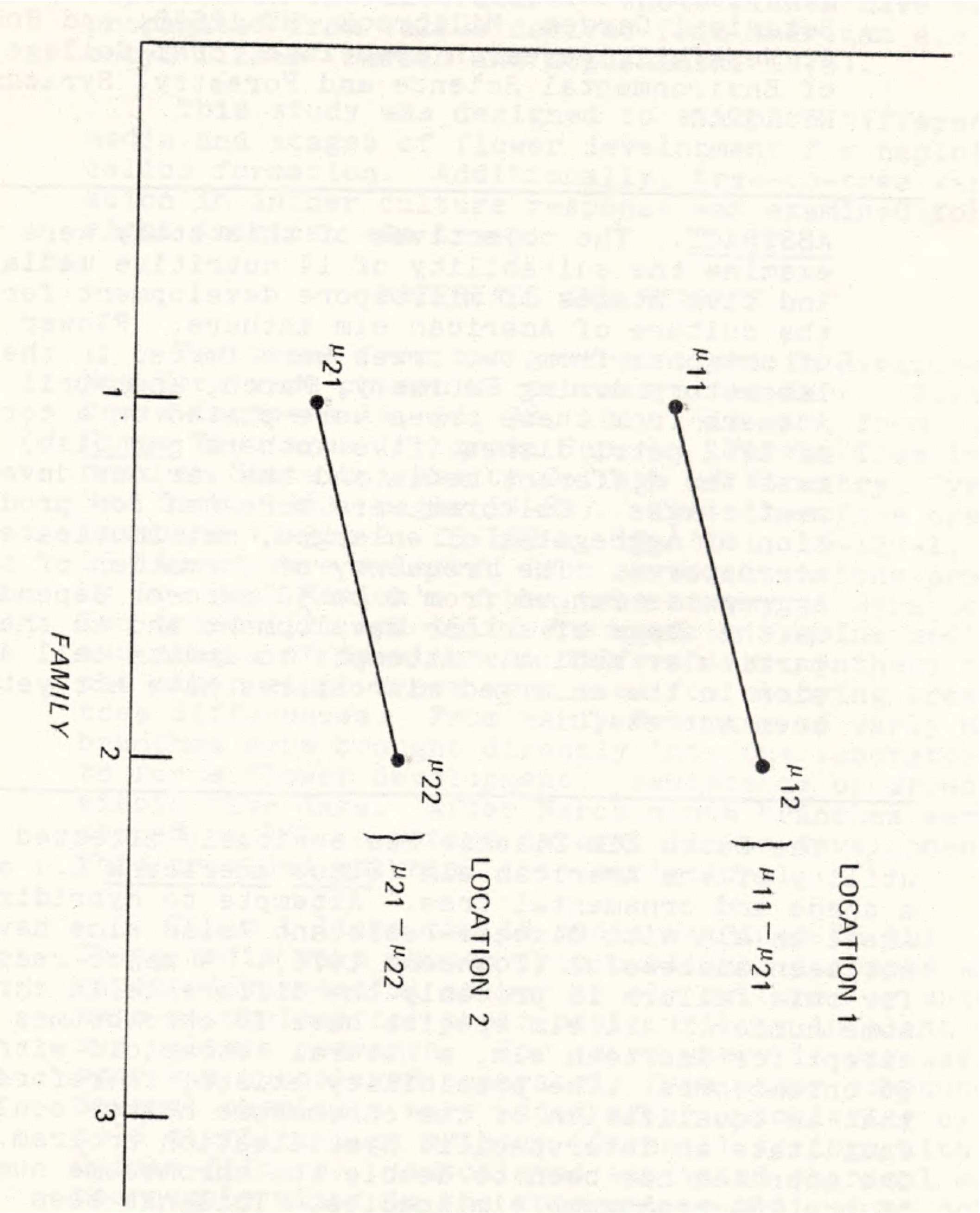

Figure 3. Profile of means for two families at two locations