Measuring Variation (From yield in field trials to spots on an array) Part II.

Review from Lecture 2

ANOVA

\[
\text{SUM } \left( \frac{(X)^2}{N} \right)
\]

where \(X\) = either individual data point and treatment or rep means.
\(N\) = the number of data points used to calculate \(X\).

The analysis of variance we preformed by hand can be written as an equation:

\[
Y_{ij} = \mu + \text{trt}_i + \text{rep}_j + e_{ij}
\]

Where \(Y_{ij}\) is the trait value of the \(i^{th}\) treatment class of the \(j^{th}\) replicate, \(\mu\) is the population mean, \(\text{trt}_i\) is the effect of the \(i^{th}\) treatment class, \(\text{rep}_j\) is the effect of the \(j^{th}\) replicate, and \(e_{ij}\) is the experimental error. The appropriate F-test for treatment is equal to \(\frac{\text{trt}_i}{e_{ij}}\).

Why was the model \(Y_i = \mu + \text{trt}_i + e_i\) less conservative?

Additional SAS resources

(basic info for novice sas users)
(introduction to sas)
http://www.stat.wisc.edu/computing/sas/intro.html
(introduction to sas)
ANOVA examples using different experimental designs
http://www.sscl.uwo.ca/sscl/statsexamples/sas/anova/

New Topics

Fixed and Random effects
Latin Square design for gene expression arrays

In this class, we will discuss more complicated ANOVA models and random and fixed affects. The file, (chsdata2.xls), contains the same data as chszero.xls, but a second experiment has been added (Lect. 3 Appendix 1).
What is the appropriate F test for treatment for the chsdata2.xls data set? First we need to
decide if the effects are fixed or random. The question becomes, are the effects being
tested considered a sample of all possible effects (replicates, blocks, locations, years) or
are they meant to represent a specific event, place, etc…?

If there are random effects in the model SAS can be used to help us answer that question
using the following program by using the random / test; option. This option will generate
expected mean squares for random effects. The syntax for this option follow:

Proc GLM;
   Class list effects here;
   Model var = model;
Random list random effects here / test;

Program

data chs;
   infile 'a:chsdata2.csv' delimiter = ',' firstobs = 2;
   input trt exp time rep chs;
proc sort;
   by trt exp rep;
proc glm;
   class trt exp rep;
   model chs = trt|exp|rep(exp);
   random trt|exp|rep(exp) / test;
run;

Note: all effects have been listed as random, simply to generate the equations. Partial output
follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>Type III Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>exp</td>
<td>Var(\text{Error})+\text{Var}(\text{trt}\text{*rep}(\text{exp}))+4 \text{Var}(\text{rep}(\text{exp}))+3\text{Var}(\text{trt}\text{*exp})+ 12\text{Var}(\text{exp})</td>
</tr>
<tr>
<td>trt</td>
<td>Var(\text{Error})+\text{Var}(\text{trt}\text{*rep}(\text{exp}))+ 3\text{Var}(\text{trt}\text{*exp})+6 \text{Var}(\text{trt})</td>
</tr>
<tr>
<td>rep(exp)</td>
<td>Var(\text{Error}) +Var(\text{trt}\text{*rep}(\text{exp}))+ 4 \text{Var}(\text{rep}(\text{exp}))</td>
</tr>
<tr>
<td>trt*exp</td>
<td>Var(\text{Error}) +Var(\text{trt}\text{*rep}(\text{exp}))+ 3 \text{Var}(\text{trt}\text{*exp})</td>
</tr>
<tr>
<td>trt*rep(exp)</td>
<td>Var(\text{Error}) +Var(\text{trt}\text{*rep}(\text{exp}))</td>
</tr>
</tbody>
</table>

From this output we can determine that the appropriate error for testing the significance
of “treatment” is trt*exp. We can now re-write the program.

data chs;
   infile 'a:chsdata2.csv' delimiter = ',' firstobs = 2;
   input trt exp time rep chs;
proc sort;
   by trt exp rep;
proc glm;
  class trt exp rep;
  model chs = trt exp rep(exp) trt*exp;
  random exp rep(exp);
  test h = trt e = trt*exp;
run;
quit;
OUTPUT

Dependent Variable: chs

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>11</td>
<td>1.35893640</td>
<td>0.12353967</td>
<td>2.22</td>
<td>0.0928</td>
</tr>
<tr>
<td>Error</td>
<td>12</td>
<td>0.66673831</td>
<td>0.05556153</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>23</td>
<td>2.02567471</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square  Coef Var  Root MSE  chs Mean
0.670856  20.54038  0.235715  1.147569

Fixed Treatment Fixed experiment M2/ M5

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>0.60948852</td>
<td>0.20316284</td>
<td>3.66</td>
<td>0.0442</td>
</tr>
<tr>
<td>exp</td>
<td>1</td>
<td>0.10440347</td>
<td>0.10440347</td>
<td>1.88</td>
<td>0.1955</td>
</tr>
<tr>
<td>rep(exp)</td>
<td>4</td>
<td>0.24350184</td>
<td>0.06087546</td>
<td>1.10</td>
<td>0.4024</td>
</tr>
<tr>
<td>trt*exp</td>
<td>3</td>
<td>0.40154257</td>
<td>0.13384752</td>
<td>2.41</td>
<td>0.1178</td>
</tr>
</tbody>
</table>

t Tests (LSD) for chs

NOTE: This test controls the type I comparisonwise error rate not the experimentwise error rate.

Alpha          0.05
Error Degrees of Freedom  12
Error Mean Square  0.055562
Critical Value of t  2.17881
Least Significant Difference  0.2965

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>t Grouping  Mean</th>
<th>N</th>
<th>trt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1.4235</td>
<td>6 21    + Agro - Sonication</td>
</tr>
<tr>
<td>B</td>
<td>1.0582</td>
<td>6 22    + Agro + Sonication</td>
</tr>
<tr>
<td>B</td>
<td>1.0577</td>
<td>6 11    - Agro - Sonication</td>
</tr>
<tr>
<td>B</td>
<td>1.0508</td>
<td>6 12    - Agro + Sonication</td>
</tr>
</tbody>
</table>

Fixed Treatment Random experiment M2/ M4

Tests of Hypotheses Using the Type III MS for trt*exp as an Error Term
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>trt</td>
<td>3</td>
<td>0.60948852</td>
<td>0.20316284</td>
<td>1.52</td>
<td>0.3700</td>
</tr>
</tbody>
</table>

Note the difference if we consider the effects as Fixed (treatment p = 0.0442) or Random (treatment p = 0.37). It is important to know what you consider your effects before doing the experiment…

Comments on PROC GLM: Uses Satterthwaite approximation to estimate degrees of freedom as a default.

Use of ANOVA to set objective cut-off values in gene-expression studies.

Kerr and Churchill (2001) provide the rationale for analyzing dye-swap gene expression experiments as latin square designs.

Review the experiment
- DNA Chips or “Arrays”
- Goal of experiments to detect differentially expressed genes

For our purposes the “intensity” of a spot could just as easily represent yield.

Review issues of data normalization and data transformation
- The use of ratios and loss of information
- Normalization by other means
- Transformations

<table>
<thead>
<tr>
<th>Array</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>Treatment 1</td>
<td>Treatment 2</td>
</tr>
<tr>
<td>Green</td>
<td>Treatment 2</td>
<td>Treatment 1</td>
</tr>
</tbody>
</table>

The statistical model being tested is:

\[ Y_{hijk} = \mu + \text{Array}_h + \text{Dye}_i + \text{Treatment}_j + \text{Gene}_k + (A \times G)_{hk} + (T \times G)_{jk} + \text{Error}_{hijk} \]

They present the following ANOVA table (Table 3, Kerr et al.)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array</td>
<td>1</td>
<td>92.34</td>
<td>92.34</td>
</tr>
<tr>
<td>Dye</td>
<td>1</td>
<td>0.74</td>
<td>0.74</td>
</tr>
<tr>
<td>Treatment</td>
<td>1</td>
<td>2.97</td>
<td>2.97</td>
</tr>
<tr>
<td>Gene</td>
<td>1285</td>
<td>1885.89</td>
<td>1.47</td>
</tr>
<tr>
<td>A x G</td>
<td>1285</td>
<td>160.01</td>
<td>0.12</td>
</tr>
<tr>
<td>T x G</td>
<td>1285</td>
<td>1357.28</td>
<td>1.06</td>
</tr>
<tr>
<td>Error</td>
<td>1285</td>
<td>82.75</td>
<td>0.0644</td>
</tr>
<tr>
<td>Corrected total</td>
<td>5143</td>
<td>3581.99</td>
<td></td>
</tr>
</tbody>
</table>

The SAS code for this model would be:

Proc glm;
  Class array dye treatment gene;
  Model [expression] = array dye treatment gene array*gene treatment*gene / SS;
Note: The issues normalization and transformation of the data must be considered before the final analysis.

To determine an objective cut-off for the experiment, we can use the formula for least significant difference (LSD): 
\[
\text{LSD} = t \times (\text{MSE} \times 2/n)^{1/2}
\]

Our value of \( n = 2 \) for the number of times each gene is hybridized to a specific treatment RNA, and the value \( t \) is the tabulated \( t \) value for the desired confidence level and the degrees of freedom of the initial sample. Remember that a significant \( T \times G \) effect indicates that some genes are differentially expressed. Now what we want to do is estimate a cut-off that we can use to determine when an individual gene can be considered differentially expressed between the two treatments.

The LSD can be used to set the “cut-off” and, because the data were transformed using LN, can be interpreted as “Fold changes” using \( e^{\text{LSD}/2} \).

Likewise, the confidence interval can be used to establish a cut-off, \( I = 2 \times t \times (\sigma^2/n)^{1/2} \), and fold changes interpreted using \( e^{\text{CI}/2} \).

\[
\begin{array}{cccccc}
\text{P-value} & t & \text{mse} & \text{LSD} & \text{fold} \\
0.001 & 3.291 & 0.0644 & 0.835162 & 1.518284 \\
0.01 & 2.576 & 0.0644 & 0.653716 & 1.386604 \\
0.02 & 2.326 & 0.0644 & 0.590273 & 1.343309 \\
0.05 & 1.96 & 0.0644 & 0.497392 & 1.282352 \\
0.1 & 1.645 & 0.0644 & 0.417454 & 1.232109 \\
0.2 & 1.282 & 0.0644 & 0.325335 & 1.176645 \\
\end{array}
\]

\[
\begin{array}{cccccc}
\text{P-value} & t & \text{mse} & \text{CI} & \text{fold} \\
0.001 & 3.291 & 0.0644 & 1.181098 & 1.804978 \\
0.01 & 2.576 & 0.0644 & 0.924493 & 1.587636 \\
0.02 & 2.326 & 0.0644 & 0.834772 & 1.517988 \\
0.05 & 1.96 & 0.0644 & 0.703419 & 1.421495 \\
0.1 & 1.645 & 0.0644 & 0.590369 & 1.343374 \\
0.2 & 1.282 & 0.0644 & 0.460093 & 1.258659 \\
\end{array}
\]

From Kerr et al.

\[
\begin{array}{cccc}
p=0.01 & \text{resampling} & 1.61 & 2.236695 \\
p=0.01 & \text{normal} & 1.29 & 1.905986 \\
\end{array}
\]

\(^1\) Assumes a two-tailed test with df = infinity.
The problem of multiple comparisons:

Bonferroni inequality CER = alpha / C where C = the number of comparisons and alpha is the desired probability level.

Discussion:

Why the difference in cut-off values between standard ANOVA and ANOVA with re-sampling?

Possibilities
Value of t
Df
Method of estimating confidence interval

References:


Exercise
I am providing you with a data set, 2353master.xls, that contains data for 96 gene probes spotted onto a nylon membrane (a dot blot). Replicate membranes were probed with cDNA extracted from plants that were infected (I) or not infected (U) by a bacterial pathogen. The intensity of hybridization was estimated using a phosphor imager. Data are the Sum of pixels above background (SumABG) by the phosphor imager, data normalized for each membrane using the expression of a “steady state” gene elf4a, and LN transformations of the SumABG and normalized data.

What model would you use to test for differential gene expression in the data set? Is there evidence for differential gene expression in this data set? Of the four variables, is one better for the ANOVA? (The following code may help you decide, I’ll leave it to you to work out the DATA step)

```plaintext
proc glm;
   class gene exp blot trt ;
   model sumabg = you need to figure out the model / ss3 ;
   output out=two rstudent=r1 ;
proc univariate normal plot;
   var sumabg r1; quit;
```
# Appendix 1

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Experiment</th>
<th>time</th>
<th>rep</th>
<th>chs/18S</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.96552</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.1</td>
<td>2</td>
<td>0.9875</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.1</td>
<td>3</td>
<td>0.901493</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>1.46939</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.1</td>
<td>2</td>
<td>0.899244</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.1</td>
<td>3</td>
<td>1.307377</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>2.07353</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0.1</td>
<td>2</td>
<td>1.4375</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0.1</td>
<td>3</td>
<td>1.37788</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>0.9304</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.1</td>
<td>2</td>
<td>0.896</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.1</td>
<td>3</td>
<td>1.316456</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>1.25862</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>0.945679</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.1</td>
<td>3</td>
<td>1.287313</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>0.99301</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>0.987562</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.1</td>
<td>3</td>
<td>0.648256</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>1.27014</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>1.039216</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.1</td>
<td>3</td>
<td>1.342975</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0.1</td>
<td>1</td>
<td>0.86545</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0.1</td>
<td>2</td>
<td>0.990172</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>0.1</td>
<td>3</td>
<td>1.350962</td>
</tr>
</tbody>
</table>