Notes for Tuesday February 10, 1998

Today, and on Thursday, we are just going through some more examples to become more familiar with what one can do with ANOVA models.

Today we will look at two familiar examples, a factorial 3×4 design and a (another) randomized block design. The twists are:

- In the first example, well-chosen transformations can make interactions (as well as other problems) disappear;
- In the second example, we illustrate a simple analysis using SAS as well as SPLUS, and review multiple comparisons procedures.

1 A Replicated, Three-By-Four Design

The following data are the survival times of groups of four animals randomly allocated to three poisons and four treatments. Then experiment was part of an investigation to combat the effects of certain toxic agents. The is a 4-fold replication of a 3×4 factorial design. Do an analysis of these data.

	Treatment			
poison	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

As always the first step is to enter the data. I'll just enter the data and create the factor levels on the fly.

Here is the raw data, in a file called `poison.dat`

```
0.31 0.82 0.43 0.45
0.45 1.10 0.45 0.71
0.46 0.88 0.63 0.66
0.43 0.72 0.76 0.62
```



```
0.36 0.92 0.44 0.56
0.29 0.61 0.35 1.02
0.40 0.49 0.31 0.71
0.23 1.24 0.40 0.38
0.22 0.30 0.23 0.30
0.21 0.37 0.25 0.36
0.18 0.38 0.24 0.31
0.23 0.29 0.22 0.33
```

and here is my first pass at reading in the data and creating appropriate factors (note the use of `as.factor` again).

```
402 > poison <- scan("poison.dat")
402 > poison
 [1] 0.31 0.82 0.43 0.45 0.45 1.10 0.45 0.71 0.46 0.88 0.63 0.66 0.43 0.72 0.76
[16] 0.62 0.36 0.92 0.44 0.56 0.29 0.61 0.35 1.02 0.40 0.49 0.31 0.71 0.23 1.24
[31] 0.40 0.38 0.22 0.30 0.23 0.30 0.21 0.37 0.25 0.36 0.18 0.38 0.24 0.31 0.23
[46] 0.29 0.22 0.33
402 > length(poison)
[1] 48
```

Create a vector of the Tmt levels,

```
402 > rep(1:4, len=48)
 [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2 3 4 1 2
[39] 3 4 1 2 3 4 1 2 3 4
402 > Tmt <- as.factor( rep(1:4, len=48 ) )
```

and one of the Poison levels

```
402 > rep(1:3, rep(16,3))
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
[39] 3 3 3 3 3 3 3 3 3 3

402 > Poison <- as.factor(rep(1:3, rep(16,3)))
```

and finally a data frame to hold all the data neatly together.

```
402 > poison.frame <- data.frame(poison, Tmt, Poison)
402 > poison.frame
   poison Tmt Poison
 1   0.31   1      1
 2   0.82   2      1
 3   0.43   3      1
 4   0.45   4      1
402 > rm(poison, Tmt, Poison)
```

So what sort of model *might* make sense here? How about

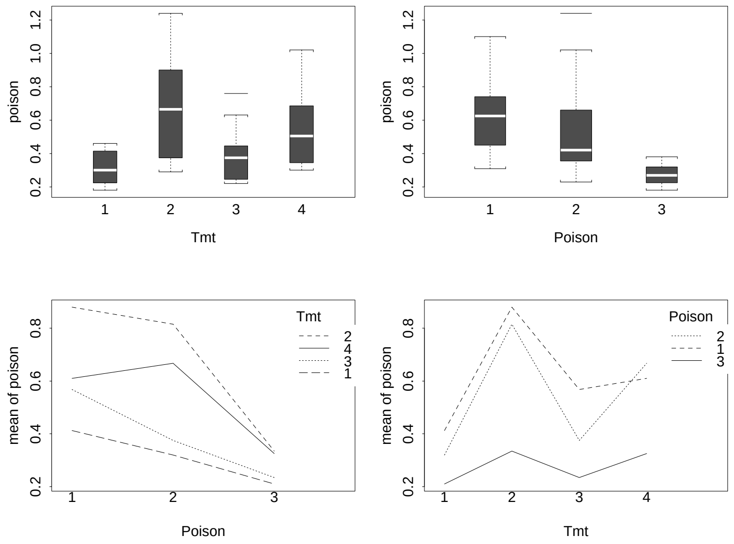
$$Y_{ij} = \mu + \tau_i + \beta_j + \omega_{ij} + \epsilon_{ijk}$$

where k implicitly indexes the replications?

That model allows us to estimate main effects and interactions, and still leaves room for random error? First thing to do is look at some pictures.

```
402 > plot.factor(poison.frame)
402 > attach(poison.frame, pos=1)
402 > interaction.plot(Poison, Tmt, poison)
402 > interaction.plot(Tmt, Poison, poison)
```

The plot of the RAW DATA is



It looks like there are treatment differences and some potential interactions. Why not let the ANOVA do its job and see where we go.

```
402 > anova(mymod <- aov(poison ~ Poison + Tmt))
Analysis of Variance Table

Response: poison
```



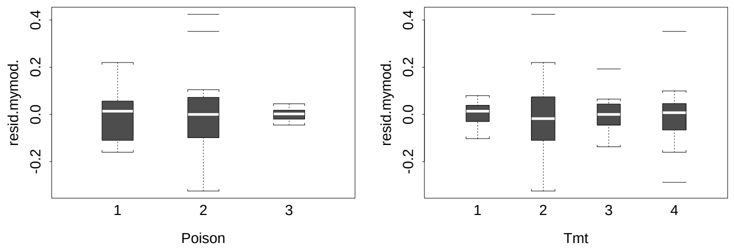
```
Terms added sequentially (first to last)
          Df Sum of Sq   Mean Sq  F Value        Pr(F)
Poison     2  1.033012 0.5165062 20.64329 5.703730e-07
Tmt        3  0.921206 0.3070688 12.27267 6.696971e-06
Residuals 42  1.050862 0.0250205
```

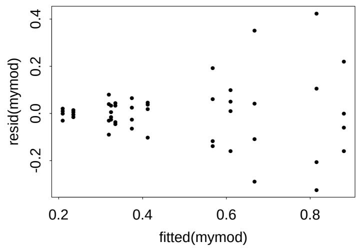
That analysis looks OK, except we forgot to add the interaction term. Here is a version with interactions.

```
402 > anova(mymod <- aov(poison ~ Poison + Tmt + Poison*Tmt))
Analysis of Variance Table

Response: poison

Terms added sequentially (first to last)
           Df Sum of Sq   Mean Sq  F Value     Pr(F)
Poison      2  1.033012 0.5165062 23.22174 0.0000003
Tmt         3  0.921206 0.3070688 13.80558 0.0000038
Poison:Tmt  6  0.250138 0.0416896  1.87433 0.1122506
Residuals  36  0.800725 0.0222424
```

It appears there are strong main effects, but little or no interaction.

At the very least, we should now try a residual plot of these data.

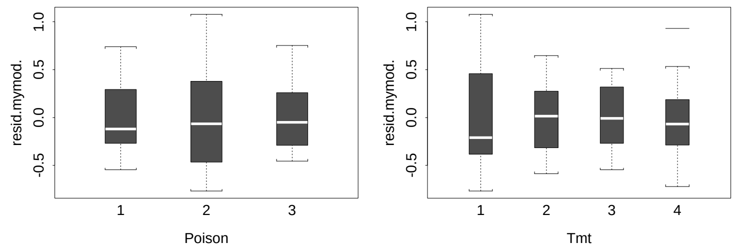


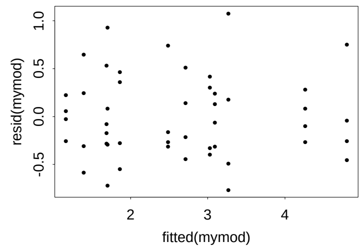


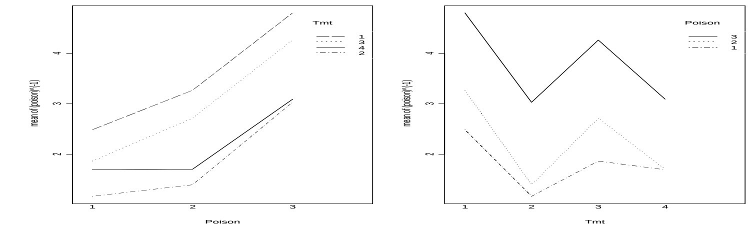
Those are not stellar looking plots. The Box plots do not look uniform, and there seems to be a relationship between residual and fitted value. What to do?

One could go through a formal process of selecting a transformation, perhaps from the Box-Cox family. I've decided to just use an inverse, and sqrt.

```
402 > anova(mymod <- aov( (poison)^(-1) ~ Poison + Tmt + Poison*Tmt))
Analysis of Variance Table

Response: (poison)^(-1)

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value     Pr(F)
Poison      2  34.87712 17.43856 72.63475 0.0000000
Tmt         3  20.41429  6.80476 28.34307 0.0000000
Poison:Tmt  6   1.57077  0.26180  1.09042 0.3867329
Residuals  36   8.64308  0.24009
402 > plot(1:10)
402 > plot.factor(data.frame(resid(mymod), Poison, Tmt))
402 > plot(fitted(mymod),resid(mymod))
402 > interaction.plot(Poison, Tmt, (poison)^(-1))
402 > interaction.plot(Tmt, Poison, (poison)^(-1))
```

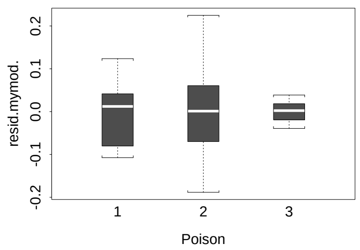



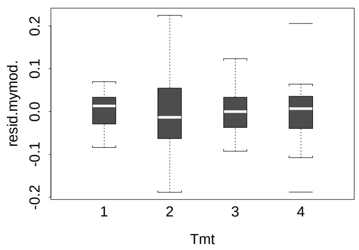


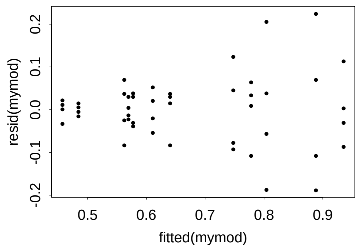


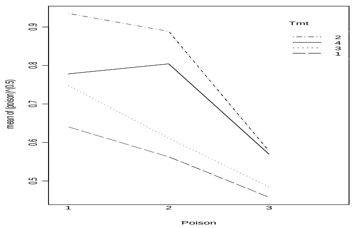

```
402 > anova(mymod <- aov( (poison)^(0.5) ~ Poison + Tmt + Poison*Tmt))
Analysis of Variance Table

Response: (poison)^(0.5)

Terms added sequentially (first to last)
           Df Sum of Sq   Mean Sq  F Value     Pr(F)
Poison      2 0.5609264 0.2804632 34.38932 0.0000000
Tmt         3 0.4306422 0.1435474 17.60123 0.0000003
Poison:Tmt  6 0.0793987 0.0132331  1.62259 0.1692592
```

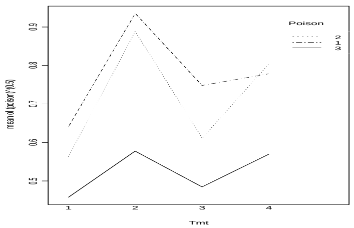
```
Residuals  36 0.2935992 0.0081555
402 > plot.factor(data.frame(resid(mymod), Poison, Tmt))
402 > plot(fitted(mymod),resid(mymod))
402 > interaction.plot(Poison, Tmt, (poison)^(0.5))
402 > interaction.plot(Tmt, Poison, (poison)^(0.5))
```











Note that the inverse transformation seems to be the ticket. Note particularly that any hint of interaction is completely missing on the inverse scale.

2 An example in Splus and SAS

2.1 SAS Basics

We do not have a textbook for SAS, but there are many on-line resources for learning about SAS. These include

- `http://www.stolaf.edu/people/single/stats/sasintro/sas.html`, an introduction to SAS that is focused on PROC GLM, one of the two standard procedures for doing ANOVA problems in SAS.
- `http://lib.stat.cmu.edu/ giovanni/745/`, the home page of 36-462 and 36-745. This course is about multivariate statistical methods for the social and behavioral sciences, which SAS is quite well suited to do.
- `/afs/andrew/stat/data/SAS-Tutorials`, a collection of examples and tutorials in SAS for Statistics courses at CMU (similar to `/afs/andrew/stat/data/S-Tutorials`, which you have probably used before).
- `Interactive SAS Help files online.` If you just say `sas` at the UNIX prompt on an Andrew workstation, you will get into SAS interactive mode, which consists of a number of windows controlled by two control panel windows. One of the control panels has a `Help` button, which brings up help for SAS that is similar to WINDOWS or Macintosh hypertext help.

There are two ways to run SAS on Andrew.

1. <u>Command-line method.</u> Type

   ```
   [unix16]bj20% sas myfile.sas
   ```

   at the UNIX prompt on an Andrew workstation, where `myfile.sas` is a collection of SAS PROC's (procedures) and other commands, as in the example below.

2. <u>Interactive method.</u> Type just

   ```
   [unix16]bj20% sas
   ```

   at the UNIX prompt on an ANDREW workstation. This gets you into SAS interactive mode, which consists of a number of windows controlled by two control panel windows. You can edit commands, submit jobs, view output, etc., in these various windows, which sounds very nice; but I find them very hard to work with. The most useful part of SAS interactive mode is the online interactive hypertext help.

In the next example, we will do the same problem two ways: once in SPLUS, and once in SAS using PROC ANOVA. As an exercise, you should try to do the exact same analysis (or as near as you can get) using PROC GLM (this is a good way to explore the various online SAS resources mentioned above!).

2.2 Review of Multiple Comparison Methods

Recall that if we want a (family) of tests at the α level, then we form confidence intervals of the form

$$\hat{D} \pm Z \times \hat{s}(\hat{D})$$

All that differs in the various tests is the multiplier, Z. The following table summarized the various multipliers.

Test	Critical Statistic	Remarks
Single Comparison	$t[1-\alpha/2;\nu]$	ν df for Error
Tukey all pairs	$T = \frac{1}{\sqrt{2}}q[1-\alpha;r;\nu]$	r groups
Scheffe's method	$S^2 = (r-1)F[1-\alpha;r-1;\nu]$	
Bonferroni method	$B = t[1-\alpha/2g;\nu]$	g comparisons

2.3 A Randomized Block Design

The following data are found on page 28, of *Planning of Experiments* by D. R .Cox, Wiley, 1958, and were previously described in Cochran and Cox's (a different Cox) book.

The treatments are five levels of potash (26, 54, 7, 108, 144 lb K_2O per acre applied to a cotton crop. On observation analyzed was a measure of single-fiber strength in arbitrary units, obtained as an average of a number of test on the cotton from each plot. There were three blocks each containing five plots. We have no details of the arrangement of plots within blocks. A summary of the data (which ignores the order in which the data were collected) is

	T_1	T_2	T_3	T_4	T_5
Block 1	7.62	8.14	7.76	7.17	7.46
Block 2	8.00	8.15	7.73	7.57	7.68
Block 3	7.93	7.87	7.74	7.80	7.21

We are interested in knowing if the treatments differ. We are prepared to live with block to block variations, but don't really care about them.

I entered the data into a computer file as follows,

```
Yield   Tmt    Block
7.62    T1     B1
8.14    T2     B1
7.76    T3     B1
7.17    T4     B1
7.46    T5     B1
8.00    T1     B2
8.15    T2     B2
7.73    T3     B2
7.57    T4     B2
7.68    T5     B2
7.93    T1     B3
7.87    T2     B3
7.74    T3     B3
```

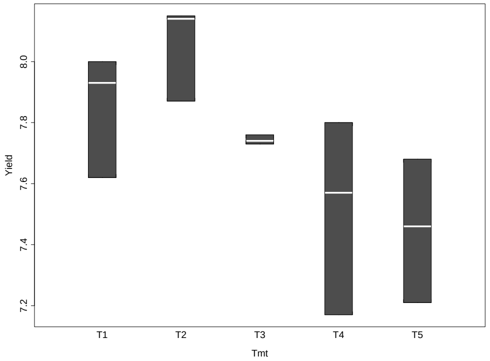


```
7.80    T4     B3
7.21    T5     B3
```

and then read the data into Splus

```
402 > coxdat <- read.table("coxdata.dat", head=T)
402 > coxdat
   Yield Tmt Block
 1  7.62  T1    B1
 2  8.14  T2    B1
......
```

and now, to a simple analysis (assume I've presented all the plots—I did look at them and Treatment 3 is a little less variable than one might expect.).




```
402 > mymod <- aov(Yield ~ Block + Tmt)

402 > anova(mymod)
Analysis of Variance Table

Response: Yield

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
Block      2   0.09712 0.048560 1.111594 0.3749855
Tmt        4   0.73244 0.183110 4.191599 0.0403682
Residuals  8   0.34948 0.043685
402 >
```

It is quite clear that there is a reasonable treatment effect, and an ignorable Block effect (more on that in a moment), but which treatment is the best? In the absence of any particular interesting contrasts we might want to test all pairwise contrasts to see which treatments are better than the others. The plot shows us that the highest treatment level doesn't seem to generate the highest yield.

2.3.1 The multiple comparison procedures

With 5 treatments (and hence 10 possible pairwise comparisons) we can't do simple t-tests and the simple Bonferroni adjusts would be too severe, so lets try both the Tukey test and the Scheffe test. To get started we need to calculate the treatment means. (We could do this with a fancy use of contrasts, but brute force is just as good.) This is yet another one of those "cool Splus tricks"—you should look up the `tapply` and `apply` functions.

```
402 > mymeans<-tapply(Yield, Tmt, mean)
402 > mymeans
   T1       T2       T3       T4   T5
 7.85 8.053333 7.743333 7.513333 7.45
```

The rest of the calculation is essentially by hand. Lets compare level T2 (the biggest) with level T5 (the smallest). We expect this to be significant.

```
402 > mymeans[2] - mymeans[5]
        T2
 0.6033333
```

A 90% confidence interval for this difference, using Tukey's studentized range intervals, is

$$\hat{D} \pm Z \times \hat{s}(\hat{D})$$

where

$$D = .60$$

,

$$\hat{s}(\hat{D}) = \sqrt{MSE \times (\frac{1}{3} + \frac{1}{3})} = \sqrt{0.043685 * 2/3} = 0.17$$

and

$$Z = \frac{1}{\sqrt{2}} q[0.90; 5; 8] = 2.95$$

Hence the interval is

$$0.6 \pm 2.95 \times 0.17 = 0.6 \pm 0.50$$

.

From the calculation it is clear that no pairwise difference that is less than 0.5 will be significant, which rules out everything else expect T2 against T4 or T5.

If we had done the test just as a pre-known contrast, using the t-test, the confidence 90% confidence region would have been

$$0.6 \pm 1.86 \times 0.17 = 0.6 \pm 0.31$$

. Thus, in addition, the comparison of T2 with T3 and T1 with T4 or T5 would all have been significant. (But we would have overstated things).

How about as an F-test? For Scheffe's method we need,

$$S^2 = (r-1)F[1-\alpha; r-1; \nu]$$

. If we are only going to do the pairwise differences, then $\hat{s}(\hat{D}$ is the same as before. Now

```
402 > qf(.90, 4, 8)
[1] 2.806426
```

so, $S = \sqrt{(r-1)F[1-\alpha; r-1; \nu]}$ is

```
402 > sqrt(4*qf(.90, 4, 8))
[1] 3.350478
```

Now the multiplier is 3.35, so the confidence interval is

$$0.6 \pm 3.35 \times 0.17 = 0.6 \pm 0.57$$

So with this method we only get a significant difference between T2 and T5.

2.3.2 An analysis in SAS

Here is a SAS program to do the same thing we did in Splus.

```
options linesize=78;
data cox;
        input yield tmt $ block $ ;
        cards;
7.62     T1     B1
8.14     T2     B1
7.76     T3     B1
7.17     T4     B1
7.46     T5     B1
8.00     T1     B2
8.15     T2     B2
7.73     T3     B2
7.57     T4     B2
7.68     T5     B2
7.93     T1     B3
7.87     T2     B3
7.74     T3     B3
7.80     T4     B3
7.21     T5     B3
;
proc anova;
        class tmt block;
        model yield = tmt block;
        means tmt/bon tukey scheffe alpha=0.10;
```

(I will explain what the various lines mean in class—you should annotate this). The log file from that SAS run was

```
......
NOTE: AUTOEXEC processing beginning; file is /usr/local/lib/sas/autoexec.sas.

NOTE: SAS initialization used:
      real time           0.19 seconds
      cpu time            0.18 seconds


NOTE: AUTOEXEC processing completed.

1          options linesize=78;
2          data cox;
3               input yield tmt $ block $ ;
4               cards;

NOTE: The data set WORK.COX has 15 observations and 3 variables.
NOTE: DATA statement used:
      real time           0.28 seconds
      cpu time            0.06 seconds


20         ;
21         proc anova;
22              class tmt block;
23              model yield = tmt block;
24              means tmt/bon tukey scheffe alpha=0.10;
NOTE: The PROCEDURE ANOVA printed pages 1-5.
NOTE: PROCEDURE ANOVA used:
      real time           0.54 seconds
      cpu time            0.09 seconds


NOTE: The SAS System used:
      real time           1.06 seconds
      cpu time            0.37 seconds
.....
```

and the susbstantive output follows. Notice that the significant differences we found are confirmed with SAS.


```
                                The SAS System                               1
                                             09:19 Wednesday, January 31, 1996

                        Analysis of Variance Procedure
                           Class Level Information

                      Class    Levels    Values

                      TMT           5    T1 T2 T3 T4 T5

                      BLOCK         3    B1 B2 B3


                   Number of observations in data set = 15


                                The SAS System                               2
                                             09:19 Wednesday, January 31, 1996

                        Analysis of Variance Procedure

Dependent Variable: YIELD
                                     Sum of          Mean
Source                  DF          Squares        Square   F Value     Pr > F

Model                    6       0.82956000    0.13826000      3.16     0.0677

Error                    8       0.34948000    0.04368500

Corrected Total         14       1.17904000

                  R-Square             C.V.       Root MSE          YIELD Mean

                  0.703589         2.706677       0.209010            7.722000


Source                  DF         Anova SS    Mean Square   F Value     Pr > F

TMT                      4       0.73244000     0.18311000      4.19     0.0404
BLOCK                    2       0.09712000     0.04856000      1.11     0.3750
```

```
                     The SAS System                                3
                                   09:19 Wednesday, January 31, 1996
               Analysis of Variance Procedure
  Tukey's Studentized Range (HSD) Test for variable: YIELD
NOTE: This test controls the type I experimentwise error rate, but
      generally has a higher type II error rate than REGWQ.

              Alpha= 0.1  df= 8  MSE= 0.043685
         Critical Value of Studentized Range= 4.169
           Minimum Significant Difference= 0.503

Means with the same letter are not significantly different.

        Tukey Grouping             Mean      N  TMT

                     A           8.0533      3  T2
                     A
             B       A           7.8500      3  T1
             B       A
             B       A           7.7433      3  T3
             B
             B                   7.5133      3  T4
             B
             B                   7.4500      3  T5


                     The SAS System                                4
                                   09:19 Wednesday, January 31, 1996
               Analysis of Variance Procedure
       Bonferroni (Dunn) T tests for variable: YIELD
NOTE: This test controls the type I experimentwise error rate, but
      generally has a higher type II error rate than REGWQ.

              Alpha= 0.1  df= 8  MSE= 0.043685
                 Critical Value of T= 3.36
           Minimum Significant Difference= 0.5726

Means with the same letter are not significantly different.

         Bon Grouping              Mean      N  TMT

                    A           8.0533      3  T2
                    A
            B       A           7.8500      3  T1
            B       A
            B       A           7.7433      3  T3
            B       A
            B       A           7.5133      3  T4
            B
            B                   7.4500      3  T5
```



```
                     The SAS System                                5
                                   09:19 Wednesday, January 31, 1996

               Analysis of Variance Procedure

             Scheffe's test for variable: YIELD

NOTE: This test controls the type I experimentwise error rate but
      generally has a higher type II error rate than REGWF for all
      pairwise comparisons

              Alpha= 0.1  df= 8  MSE= 0.043685
                Critical Value of F= 2.80643
           Minimum Significant Difference= 0.5718

Means with the same letter are not significantly different.

       Scheffe Grouping             Mean      N  TMT

                     A           8.0533      3  T2
                     A
             B       A           7.8500      3  T1
             B       A
             B       A           7.7433      3  T3
             B       A
             B       A           7.5133      3  T4
             B
             B                   7.4500      3  T5
```