STAT 401A - Statistical Methods for Research Workers

Two-way ANOVA

Jarad Niemi (Dr. J)

Iowa State University

last updated: December 18, 2014

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

There is an expectation that planting density will have a different effect depending on the variety. Therefore a balanced, complete, randomized design was used.

- complete: each treatment (variety × density) is represented in the experiment
- balanced: each treatment in the experiment has the same number of replications
- randomized: treatment was randomly assigned to the plot

This is also referred to as a full factorial or fully crossed design.

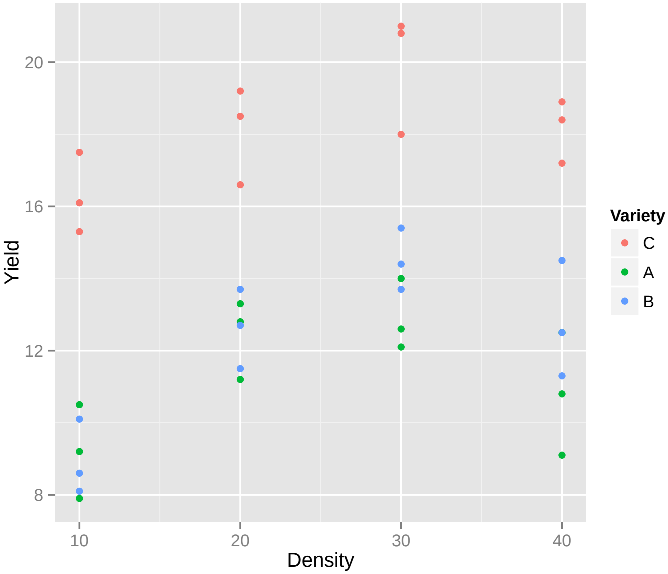
Hypotheses

- Does variety affect mean yield?
  - Is the mean yield for variety A different from B on average?
  - Is the mean yield for variety A different from B at a particular value for density?
- Does density affect mean yield?
  - Is the mean yield for density 10 different from density 20 on average?
  - Is the mean yield for density 10 different from density 20 at a particular value for variety?
- Does density affect yield differently for each variety?

For all of these questions, we want to know

- is there any effect and
- if yes, what is the nature of the effect.

Confidence intervals can answer these questions.


20
16
12
8
Yield
10
20
30
40
Density
Variety
C
A
B

Summary statistics

Number of replicates

```
  Variety 10 20 30 40
1       C  3  3  3  3
2       A  3  3  3  3
3       B  3  3  3  3
```

Mean Yield

```
  Variety        10       20       30       40
1       C 16.300000 18.10000 19.93333 18.16667
2       A  9.200000 12.43333 12.90000 10.80000
3       B  8.933333 12.63333 14.50000 12.76667
```

Standard deviation of yield

```
  Variety       10       20        30        40
1       C 1.113553 1.345362 1.6772994 0.8736895
2       A 1.300000 1.096966 0.9848858 1.7000000
3       B 1.040833 1.101514 0.8544004 1.6165808
```

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels
- Model:

$$Y_{ijk} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

  where Y_{ijk} is the
  - kth observation at the
  - ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
  - jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

Consider the models:

- Additive: $\mu_{ij} = \mu + \nu_i + \delta_j$
- Cell-means: $\mu_{ij} = \mu + \nu_i + \delta_j + \gamma_{ij}$

	10	20	30	40
A	μ_{11}	μ_{12}	μ_{13}	μ_{14}
B	μ_{21}	μ_{22}	μ_{23}	μ_{24}
C	μ_{31}	μ_{32}	μ_{33}	μ_{34}

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.
3. Build indicator variables, e.g. $\mathrm{I}(V_i = A)$ and $\mathrm{I}(D_i = 10)$.
4. The additive model:

$$\begin{aligned}\mu_i = \;& \beta_0 + \beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & + \beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30).\end{aligned}$$

   β_1 is the expected difference in yield between varieties A and C at any fixed density

5. The cell-means model:

$$\begin{aligned}\mu_i = \;& \beta_0 + \beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & + \beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30) \\ \\ & + \beta_6 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 10) + \beta_7 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 20) + \beta_8 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 30) \\ & + \beta_9 \mathrm{I}(V_i = B)\mathrm{I}(D_i = 10) + \beta_{10} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 20) + \beta_{11} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 30)\end{aligned}$$

   β_1 is the expected difference in yield between varieties A and C at a density of 40

ANOVA Table

ANOVA Table - Additive model

Source	SS	df	MS	F
Factor A	SSA	(I-1)	SSA/(I-1)	MSA/MSE
Factor B	SSB	(J-1)	SSB/(J-1)	MSB/MSE
Error	SSE	n-I-J+1	SSE/(n-I-J+1)	
Total	SST	n-1		

ANOVA Table - Cell-means model

Source	SS	df	MS	
Factor A	SSA	I-1	SSA/(I-1)	MSA/MSE
Factor B	SSB	J-1	SSB/(J-1)	MSB/MSE
Interaction AB	SSAB	(I-1)(J-1)	SSAB /(I-1)(J-1)	MSAB/MSE
Error	SSE	n-IJ	SSE/(n-IJ)	
Total	SST	n-1		

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant. Remember that an insignificant test does not prove that there is no interaction.

	Additive	Cell-means
Interpretation	Direct	Complicated
Estimate of σ^2	Biased	Unbiased

We will continue using the cell-means model to answer the scientific questions of interest.

Two-way ANOVA using PROC GLM

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;

PROC GLM DATA=tomato PLOTS=all;
  CLASS variety density;
  MODEL yield = variety|density / SOLUTION;
  LSMEANS variety / cl adjust=tukey;
  LSMEANS density / cl adjust=tukey;
  LSMEANS variety*density / cl adjust=tukey;
  RUN;
```

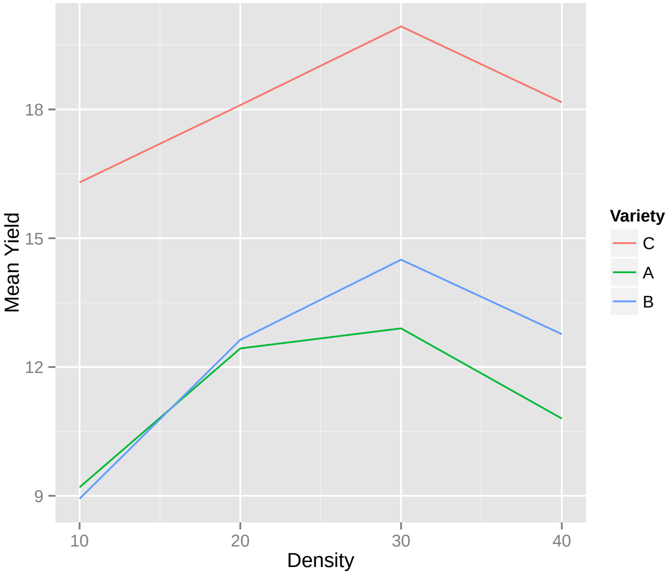
Two-way ANOVA using PROC GLM

```
                                        The GLM Procedure

Dependent Variable: yield

                                               Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                       11      422.3155556      38.3923232      24.22    <.0001
       Error                       24       38.0400000       1.5850000
       Corrected Total             35      460.3555556

                       R-Square     Coeff Var      Root MSE    yield Mean
                       0.917368      9.064568      1.258968      13.88889

       Source                      DF        Type I SS     Mean Square    F Value    Pr > F
       variety                      2      327.5972222     163.7986111     103.34    <.0001
       density                      3       86.6866667      28.8955556      18.23    <.0001
       variety*density              6        8.0316667       1.3386111       0.84    0.5484

       Source                      DF      Type III SS     Mean Square    F Value    Pr > F
       variety                      2      327.5972222     163.7986111     103.34    <.0001
       density                      3       86.6866667      28.8955556      18.23    <.0001
       variety*density              6        8.0316667       1.3386111       0.84    0.5484
```

The Type I and Type III SS are equal because the design is balanced.

Two-way ANOVA using PROC GLM

```
MODEL yield = variety|density / SOLUTION;
                                        The GLM Procedure

                                                          Standard
         Parameter                    Estimate               Error    t Value    Pr > |t|

         Intercept                 18.16666667 B        0.72686542      24.99      <.0001
         variety          A        -7.36666667 B        1.02794293      -7.17      <.0001
         variety          B        -5.40000000 B        1.02794293      -5.25      <.0001
         variety          C         0.00000000 B         .                .         .
         density          10       -1.86666667 B        1.02794293      -1.82      0.0819
         density          20       -0.06666667 B        1.02794293      -0.06      0.9488
         density          30        1.76666667 B        1.02794293       1.72      0.0986
         density          40        0.00000000 B         .                .         .
         variety*density  A 10      0.26666667 B        1.45373083       0.18      0.8560
         variety*density  A 20      1.70000000 B        1.45373083       1.17      0.2537
         variety*density  A 30      0.33333333 B        1.45373083       0.23      0.8206
         variety*density  A 40      0.00000000 B         .                .         .
         variety*density  B 10     -1.96666667 B        1.45373083      -1.35      0.1887
         variety*density  B 20     -0.06666667 B        1.45373083      -0.05      0.9638
         variety*density  B 30     -0.03333333 B        1.45373083      -0.02      0.9819
         variety*density  B 40      0.00000000 B         .                .         .
         variety*density  C 10      0.00000000 B         .                .         .
```


Mean Yield
18
15
12
9
10
20
30
40
Density
Variety
C
A
B

Is the mean yield for variety A different from B on average?

```
LSMEANS variety / cl adjust=tukey;
                               Least Squares Means
                      Adjustment for Multiple Comparisons: Tukey

...

                        Least Squares Means for effect variety
                         Pr > |t| for HO: LSMean(i)=LSMean(j)

                              Dependent Variable: yield

                      i/j             1             2             3
                         1                     0.2249        <.0001
                         2       0.2249                      <.0001
                         3       <.0001        <.0001

                    variety     yield LSMEAN      95% Confidence Limits
                    A              11.333333       10.583245    12.083422
                    B              12.208333       11.458245    12.958422
                    C              18.125000       17.374912    18.875088

                         Least Squares Means for Effect variety
                              Difference          Simultaneous 95%
                                 Between       Confidence Limits for
                     i    j        Means        LSMean(i)-LSMean(j)
                     1    2    -0.875000       -2.158534     0.408534
                     1    3    -6.791667       -8.075201    -5.508132
                     2    3    -5.916667       -7.200201    -4.633132
```

Is the mean yield at density 10 different from density 20 on average?

```
LSMEANS density / cl adjust=tukey;
                                    Least Squares Means
                          Adjustment for Multiple Comparisons: Tukey
                                            ...

                      density    yield LSMEAN      95% Confidence Limits
                      10            11.477778       10.611650    12.343905
                      20            14.388889       13.522762    15.255016
                      30            15.777778       14.911650    16.643905
                      40            13.911111       13.044984    14.777238

                             Least Squares Means for Effect density
                                  Difference          Simultaneous 95%
                                     Between      Confidence Limits for
                       i    j          Means       LSMean(i)-LSMean(j)

                       1    2       -2.911111       -4.548299    -1.273923
                       1    3       -4.300000       -5.937188    -2.662812
                       1    4       -2.433333       -4.070521    -0.796145
                       2    3       -1.388889       -3.026077     0.248299
                       2    4        0.477778       -1.159410     2.114966
                       3    4        1.866667        0.229479     3.503855
```

Is mean yield different for particular combinations?

```
LSMEANS variety*density / cl adjust=tukey;

                variety    density    yield LSMEAN      95% Confidence Limits

                A          10             9.200000        7.699824    10.700176
                A          20            12.433333       10.933157    13.933510
                A          30            12.900000       11.399824    14.400176
                A          40            10.800000        9.299824    12.300176
                B          10             8.933333        7.433157    10.433510
                B          20            12.633333       11.133157    14.133510
                B          30            14.500000       12.999824    16.000176
                B          40            12.766667       11.266490    14.266843
                C          10            16.300000       14.799824    17.800176
                C          20            18.100000       16.599824    19.600176
                C          30            19.933333       18.433157    21.433510
                C          40            18.166667       16.666490    19.666843
```

Is mean yield different for particular combinations?

```
LSMEANS variety*density / cl adjust=tukey;
```

```
                         Least Squares Means for Effect variety*density

                              Difference        Simultaneous 95%
                                 Between     Confidence Limits for
                      i     j      Means      LSMean(i)-LSMean(j)
                      1     2    -3.233333     -6.939704     0.473037
                      1     3    -3.700000     -7.406371     0.006371
                      1     4    -1.600000     -5.306371     2.106371
                      1     5     0.266667     -3.439704     3.973037
                      1     6    -3.433333     -7.139704     0.273037
                      1     7    -5.300000     -9.006371    -1.593629
                      1     8    -3.566667     -7.273037     0.139704
                      1     9    -7.100000    -10.806371    -3.393629
                      1    10    -8.900000    -12.606371    -5.193629
                      1    11   -10.733333    -14.439704    -7.026963
                      1    12    -8.966667    -12.673037    -5.260296
                      2     3    -0.466667     -4.173037     3.239704
                      2     4     1.633333     -2.073037     5.339704
                      2     5     3.500000     -0.206371     7.206371
                      2     6    -0.200000     -3.906371     3.506371
                      2     7    -2.066667     -5.773037     1.639704
                      2     8    -0.333333     -4.039704     3.373037
                      2     9    -3.866667     -7.573037    -0.160296
                      2    10    -5.666667     -9.373037    -1.960296
                      2    11    -7.500000    -11.206371    -3.793629
                      2    12    -5.733333     -9.439704    -2.026963
                      3     4     2.100000     -1.606371     5.806371
                      3     5     3.966667      0.260296     7.673037
                      3     6     0.266667     -3.439704     3.973037
```

```
tomato$Density = factor(tomato$Density)
m = lm(Yield~Variety*Density, tomato)
anova(m)


Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq  F value    Pr(>F)
Variety          2 327.60 163.799 103.3430 1.608e-12 ***
Density          3  86.69  28.896  18.2306 2.212e-06 ***
Variety:Density  6   8.03   1.339   0.8445    0.5484
Residuals       24  38.04   1.585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(lsmeans)
lsmeans(m, pairwise~Variety)


$lsmeans
 Variety   lsmean        SE df lower.CL upper.CL
 C       18.12500 0.3634327 24 17.37491 18.87509
 A       11.33333 0.3634327 24 10.58325 12.08342
 B       12.20833 0.3634327 24 11.45825 12.95842

Results are averaged over the levels of: Density
Confidence level used: 0.95

$contrasts
 contrast  estimate        SE df t.ratio p.value
 C - A     6.791667 0.5139715 24  13.214  <.0001
 C - B     5.916667 0.5139715 24  11.512  <.0001
 A - B    -0.875000 0.5139715 24  -1.702  0.2249

Results are averaged over the levels of: Density
P value adjustment: tukey method for a family of 3 means
```

```
lsmeans(m, pairwise~Density)


$lsmeans
 Density   lsmean        SE df lower.CL upper.CL
 10      11.47778 0.4196559 24 10.61165 12.34391
 20      14.38889 0.4196559 24 13.52276 15.25502
 30      15.77778 0.4196559 24 14.91165 16.64391
 40      13.91111 0.4196559 24 13.04498 14.77724

Results are averaged over the levels of: Variety
Confidence level used: 0.95

$contrasts
 contrast   estimate        SE df t.ratio p.value
 10 - 20  -2.9111111 0.5934831 24  -4.905  0.0003
 10 - 30  -4.3000000 0.5934831 24  -7.245  <.0001
 10 - 40  -2.4333333 0.5934831 24  -4.100  0.0022
 20 - 30  -1.3888889 0.5934831 24  -2.340  0.1169
 20 - 40   0.4777778 0.5934831 24   0.805  0.8514
 30 - 40   1.8666667 0.5934831 24   3.145  0.0213

Results are averaged over the levels of: Variety
P value adjustment: tukey method for a family of 4 means
```

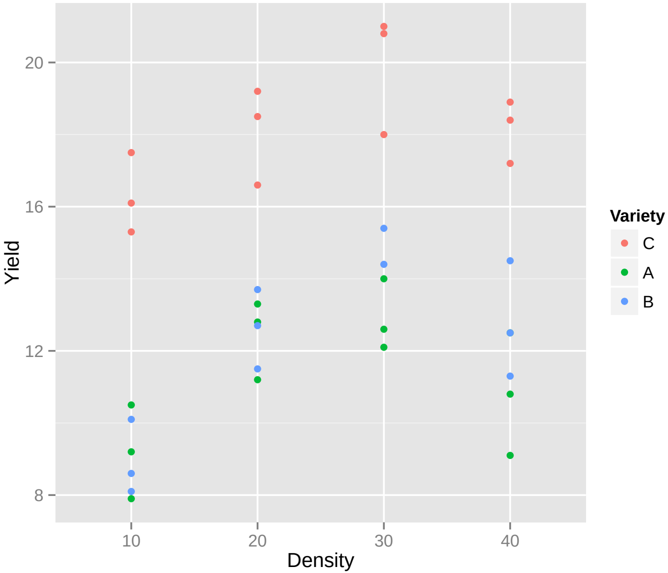
```
lsmeans(m, pairwise~Variety*Density)


$lsmeans
 Variety Density    lsmean        SE df  lower.CL upper.CL
 C       10      16.300000 0.7268654 24 14.799824 17.80018
 A       10       9.200000 0.7268654 24  7.699824 10.70018
 B       10       8.933333 0.7268654 24  7.433157 10.43351
 C       20      18.100000 0.7268654 24 16.599824 19.60018
 A       20      12.433333 0.7268654 24 10.933157 13.93351
 B       20      12.633333 0.7268654 24 11.133157 14.13351
 C       30      19.933333 0.7268654 24 18.433157 21.43351
 A       30      12.900000 0.7268654 24 11.399824 14.40018
 B       30      14.500000 0.7268654 24 12.999824 16.00018
 C       40      18.166667 0.7268654 24 16.666490 19.66684
 A       40      10.800000 0.7268654 24  9.299824 12.30018
 B       40      12.766667 0.7268654 24 11.266490 14.26684

Confidence level used: 0.95

$contrasts
 contrast       estimate       SE df t.ratio p.value
 C,10 - A,10   7.10000000 1.027943 24   6.907  <.0001
 C,10 - B,10   7.36666667 1.027943 24   7.166  <.0001
 C,10 - C,20  -1.80000000 1.027943 24  -1.751  0.8276
 C,10 - A,20   3.86666667 1.027943 24   3.762  0.0356
 C,10 - B,20   3.66666667 1.027943 24   3.567  0.0543
 C,10 - C,30  -3.63333333 1.027943 24  -3.535  0.0582
 C,10 - A,30   3.40000000 1.027943 24   3.308  0.0932
 C,10 - B,30   1.80000000 1.027943 24   1.751  0.8276
 C,10 - C,40  -1.86666667 1.027943 24  -1.816  0.7947
 C,10 - A,40   5.50000000 1.027943 24   5.350  0.0008
 C,10 - B,40   3.53333333 1.027943 24   3.437  0.0714
 A,10 - B,10   0.26666667 1.027943 24   0.259  1.0000
 A,10 - C,20  -8.90000000 1.027943 24  -8.658  <.0001
 A,10 - A,20  -3.23333333 1.027943 24  -3.145  0.1284
```

Summary

- Use LSMEANS to answer questions of scientific interest.
- Check model assumptions
- Consider alternative models, e.g. treating density as continuous

Unbalanced design

Suppose for some reason that a variety B, density 30 sample was contaminated. Although you started with a balanced design, the data is now unbalanced. Fortunately, we can still use the tools we have used previously.


Yield
20
16
12
8
10
20
30
40
Density
Variety
C
A
B

Summary statistics

Number of replicates

```
  Variety 10 20 30 40
1       C  3  3  3  3
2       A  3  3  3  3
3       B  3  3  2  3
```

Mean Yield

```
  Variety        10       20       30       40
1       C 16.300000 18.10000 19.93333 18.16667
2       A  9.200000 12.43333 12.90000 10.80000
3       B  8.933333 12.63333 14.90000 12.76667
```

Standard deviation of yield

```
  Variety       10       20        30        40
1       C 1.113553 1.345362 1.6772994 0.8736895
2       A 1.300000 1.096966 0.9848858 1.7000000
3       B 1.040833 1.101514 0.7071068 1.6165808
```

Two-way ANOVA using PROC GLM

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;
  i = _n_;

PROC GLM DATA=tomato PLOTS=all;
  WHERE i ~= 19; /* not equal to 19 */
  CLASS variety density;
  MODEL yield = variety|density / SOLUTION;
  LSMEANS variety / cl adjust=tukey;
  LSMEANS density / cl adjust=tukey;
  LSMEANS variety*density / cl adjust=tukey;
  RUN;
```

Two-way ANOVA using PROC GLM

```
                                     The GLM Procedure

Dependent Variable: yield

                                                Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                       11      423.2388571      38.4762597      23.87    <.0001
       Error                       23       37.0800000       1.6121739
       Corrected Total             34      460.3188571

                       R-Square     Coeff Var      Root MSE    yield Mean
                       0.919447      9.138391      1.269714      13.89429

       Source                      DF        Type I SS     Mean Square    F Value    Pr > F
       variety                      2      329.9878723     164.9939361     102.34    <.0001
       density                      3       84.4486608      28.1495536      17.46    <.0001
       variety*density              6        8.8023241       1.4670540       0.91    0.5052

       Source                      DF      Type III SS     Mean Square    F Value    Pr > F
       variety                      2      320.0374679     160.0187340      99.26    <.0001
       density                      3       86.0657613      28.6885871      17.79    <.0001
       variety*density              6        8.8023241       1.4670540       0.91    0.5052
```

Two-way ANOVA using PROC GLM

```
                                                Standard
Parameter                    Estimate              Error    t Value    Pr > |t|

Intercept                 18.16666667 B       0.73306978      24.78      <.0001
variety         A         -7.36666667 B       1.03671723      -7.11      <.0001
variety         B         -5.40000000 B       1.03671723      -5.21      <.0001
variety         C          0.00000000 B        .                .         .
density         10        -1.86666667 B       1.03671723      -1.80      0.0849
density         20        -0.06666667 B       1.03671723      -0.06      0.9493
density         30         1.76666667 B       1.03671723       1.70      0.1018
density         40         0.00000000 B        .                .         .
variety*density A 10       0.26666667 B       1.46613956       0.18      0.8573
variety*density A 20       1.70000000 B       1.46613956       1.16      0.2581
variety*density A 30       0.33333333 B       1.46613956       0.23      0.8222
variety*density A 40       0.00000000 B        .                .         .
variety*density B 10      -1.96666667 B       1.46613956      -1.34      0.1929
variety*density B 20      -0.06666667 B       1.46613956      -0.05      0.9641
variety*density B 30       0.36666667 B       1.55507584       0.24      0.8157
variety*density B 40       0.00000000 B        .                .         .
variety*density C 10       0.00000000 B        .                .         .
variety*density C 20       0.00000000 B        .                .         .
variety*density C 30       0.00000000 B        .                .         .
variety*density C 40       0.00000000 B        .                .         .
```

Two-way ANOVA using PROC GLM

```
                        The GLM Procedure
                       Least Squares Means
        Adjustment for Multiple Comparisons: Tukey-Kramer

             Least Squares Means for effect variety
              Pr > |t| for H0: LSMean(i)=LSMean(j)

                   Dependent Variable: yield

  i/j              1             2             3
     1                      0.1839        <.0001
     2        0.1839                      <.0001
     3        <.0001        <.0001

variety    yield LSMEAN      95% Confidence Limits
A             11.333333       10.575098    12.091569
B             12.308333       11.504103    13.112563
C             18.125000       17.366765    18.883235

              Least Squares Means for Effect variety

                   Difference         Simultaneous 95%
                      Between      Confidence Limits for
 i    j                 Means       LSMean(i)-LSMean(j)
 1    2             -0.975000       -2.313097     0.363097
 1    3             -6.791667       -8.089811    -5.493522
 2    3             -5.816667       -7.154763    -4.478570
```

Two-way ANOVA using PROC GLM

```
                      The GLM Procedure
                     Least Squares Means
      Adjustment for Multiple Comparisons: Tukey-Kramer

           Least Squares Means for effect density
            Pr > |t| for H0: LSMean(i)=LSMean(j)

                 Dependent Variable: yield

i/j              1             2             3             4
   1                      0.0004        <.0001        0.0025
   2        0.0004                      0.0967        0.8545
   3        <.0001        0.0967                      0.0189
   4        0.0025        0.8545        0.0189

    density    yield LSMEAN      95% Confidence Limits

    10            11.477778       10.602243    12.353312
    20            14.388889       13.513354    15.264423
    30            15.911111       14.965426    16.856797
    40            13.911111       13.035577    14.786646

            Least Squares Means for Effect density

                 Difference         Simultaneous 95%
                    Between      Confidence Limits for
     i    j           Means       LSMean(i)-LSMean(j)
     1    2       -2.911111       -4.567433    -1.254789
     1    3       -4.433333       -6.157288    -2.709379
     1    4       -2.433333       -4.089656    -0.777011
     2    3       -1.522222       -3.246177     0.201733
```

Two-way ANOVA using PROC GLM

```
                    The GLM Procedure
                  Least Squares Means
     Adjustment for Multiple Comparisons: Tukey-Kramer

                                                  LSMEAN
 variety     density     yield LSMEAN             Number
 A           10             9.2000000                  1
 A           20            12.4333333                  2
 A           30            12.9000000                  3
 A           40            10.8000000                  4
 B           10             8.9333333                  5
 B           20            12.6333333                  6
 B           30            14.9000000                  7
 B           40            12.7666667                  8
 C           10            16.3000000                  9
 C           20            18.1000000                 10
 C           30            19.9333333                 11
 C           40            18.1666667                 12
```

Two-way ANOVA using PROC GLM

```
                          The GLM Procedure
                         Least Squares Means
            Adjustment for Multiple Comparisons: Tukey-Kramer

              Least Squares Means for Effect variety*density

                     Difference         Simultaneous 95%
                        Between       Confidence Limits for
   i      j               Means        LSMean(i)-LSMean(j)
   1     11          -10.733333      -14.487164     -6.979502
   1     12           -8.966667      -12.720498     -5.212836
   2      3           -0.466667       -4.220498      3.287164
   2      4            1.633333       -2.120498      5.387164
   2      5            3.500000       -0.253831      7.253831
   2      6           -0.200000       -3.953831      3.553831
   2      7           -2.466667       -6.663577      1.730244
   2      8           -0.333333       -4.087164      3.420498
   2      9           -3.866667       -7.620498     -0.112836
   2     10           -5.666667       -9.420498     -1.912836
   2     11           -7.500000      -11.253831     -3.746169
   2     12           -5.733333       -9.487164     -1.979502
   3      4            2.100000       -1.653831      5.853831
   3      5            3.966667        0.212836      7.720498
   3      6            0.266667       -3.487164      4.020498
   3      7           -2.000000       -6.196911      2.196911
   3      8            0.133333       -3.620498      3.887164
   3      9           -3.400000       -7.153831      0.353831
   3     10           -5.200000       -8.953831     -1.446169
   3     11           -7.033333      -10.787164     -3.279502
   3     12           -5.266667       -9.020498     -1.512836
   4      5            1.866667       -1.887164      5.620498
```

```
m = lm(Yield~Variety*Density, tomato)
anova(m)


Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq  F value    Pr(>F)
Variety          2 327.60 163.799 103.3430 1.608e-12 ***
Density          3  86.69  28.896  18.2306 2.212e-06 ***
Variety:Density  6   8.03   1.339   0.8445    0.5484
Residuals       24  38.04   1.585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lsmeans(m, pairwise~Variety)


$lsmeans
 Variety   lsmean        SE df lower.CL upper.CL
 C       18.12500 0.3634327 24 17.37491 18.87509
 A       11.33333 0.3634327 24 10.58325 12.08342
 B       12.20833 0.3634327 24 11.45825 12.95842

Results are averaged over the levels of: Density
Confidence level used: 0.95

$contrasts
 contrast  estimate        SE df t.ratio p.value
 C - A     6.791667 0.5139715 24  13.214  <.0001
 C - B     5.916667 0.5139715 24  11.512  <.0001
 A - B    -0.875000 0.5139715 24  -1.702  0.2249

Results are averaged over the levels of: Density
P value adjustment: tukey method for a family of 3 means
```

```
lsmeans(m, pairwise~Density)


$lsmeans
 Density   lsmean        SE df lower.CL upper.CL
 10      11.47778 0.4196559 24 10.61165 12.34391
 20      14.38889 0.4196559 24 13.52276 15.25502
 30      15.77778 0.4196559 24 14.91165 16.64391
 40      13.91111 0.4196559 24 13.04498 14.77724

Results are averaged over the levels of: Variety
Confidence level used: 0.95

$contrasts
 contrast   estimate        SE df t.ratio p.value
 10 - 20  -2.9111111 0.5934831 24  -4.905  0.0003
 10 - 30  -4.3000000 0.5934831 24  -7.245  <.0001
 10 - 40  -2.4333333 0.5934831 24  -4.100  0.0022
 20 - 30  -1.3888889 0.5934831 24  -2.340  0.1169
 20 - 40   0.4777778 0.5934831 24   0.805  0.8514
 30 - 40   1.8666667 0.5934831 24   3.145  0.0213

Results are averaged over the levels of: Variety
P value adjustment: tukey method for a family of 4 means
```

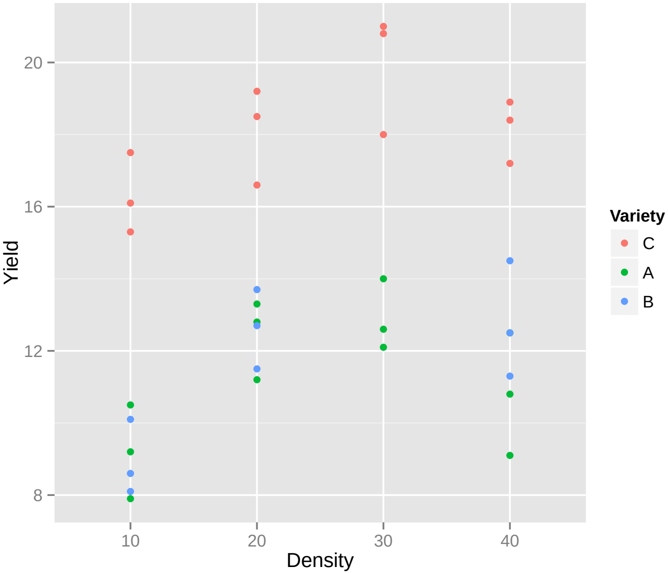
```
lsmeans(m, pairwise~Variety*Density)


$lsmeans
 Variety Density    lsmean        SE df  lower.CL upper.CL
 C       10      16.300000 0.7268654 24 14.799824 17.80018
 A       10       9.200000 0.7268654 24  7.699824 10.70018
 B       10       8.933333 0.7268654 24  7.433157 10.43351
 C       20      18.100000 0.7268654 24 16.599824 19.60018
 A       20      12.433333 0.7268654 24 10.933157 13.93351
 B       20      12.633333 0.7268654 24 11.133157 14.13351
 C       30      19.933333 0.7268654 24 18.433157 21.43351
 A       30      12.900000 0.7268654 24 11.399824 14.40018
 B       30      14.500000 0.7268654 24 12.999824 16.00018
 C       40      18.166667 0.7268654 24 16.666490 19.66684
 A       40      10.800000 0.7268654 24  9.299824 12.30018
 B       40      12.766667 0.7268654 24 11.266490 14.26684

Confidence level used: 0.95

$contrasts
 contrast        estimate       SE df t.ratio p.value
 C,10 - A,10   7.10000000 1.027943 24   6.907  <.0001
 C,10 - B,10   7.36666667 1.027943 24   7.166  <.0001
 C,10 - C,20  -1.80000000 1.027943 24  -1.751  0.8276
 C,10 - A,20   3.86666667 1.027943 24   3.762  0.0356
 C,10 - B,20   3.66666667 1.027943 24   3.567  0.0543
 C,10 - C,30  -3.63333333 1.027943 24  -3.535  0.0582
 C,10 - A,30   3.40000000 1.027943 24   3.308  0.0932
 C,10 - B,30   1.80000000 1.027943 24   1.751  0.8276
 C,10 - C,40  -1.86666667 1.027943 24  -1.816  0.7947
 C,10 - A,40   5.50000000 1.027943 24   5.350  0.0008
 C,10 - B,40   3.53333333 1.027943 24   3.437  0.0714
 A,10 - B,10   0.26666667 1.027943 24   0.259  1.0000
 A,10 - C,20  -8.90000000 1.027943 24  -8.658  <.0001
 A,10 - A,20  -3.23333333 1.027943 24  -3.145  0.1284
```

Summary

The analysis can be completed just like the balanced design using LSMEANS to answer scientific questions of interest.

Incomplete design

Suppose none of the samples from Variety B, density 30 were obtained. Now the analysis becomes more complicated.


Yield
20
16
12
8
10
20
30
40
Density
Variety
C
A
B

Summary statistics

Number of replicates

```
  Variety 10 20 30 40
1       C  3  3  3  3
2       A  3  3  3  3
3       B  3  3  0  3
```

Mean Yield

```
  Variety        10       20       30       40
1       C 16.300000 18.10000 19.93333 18.16667
2       A  9.200000 12.43333 12.90000 10.80000
3       B  8.933333 12.63333      NaN 12.76667
```

Standard deviation of yield

```
  Variety       10       20        30        40
1       C 1.113553 1.345362 1.6772994 0.8736895
2       A 1.300000 1.096966 0.9848858 1.7000000
3       B 1.040833 1.101514        NA 1.6165808
```

Two-way ANOVA using PROC GLM

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;

PROC GLM DATA=tomato PLOTS=all;
  WHERE ~(variety='B' & density=30);
  CLASS variety density;
  MODEL yield = variety|density / SOLUTION;
  LSMEANS variety / cl adjust=tukey;
  LSMEANS density / cl adjust=tukey;
  LSMEANS variety*density / cl adjust=tukey;
  RUN;
```

Two-way ANOVA using PROC GLM

```
                                        The GLM Procedure
Dependent Variable: yield
                                               Sum of
       Source                      DF         Squares     Mean Square    F Value    Pr > F
       Model                       10     421.0933333      42.1093333      25.33    <.0001
       Error                       22      36.5800000       1.6627273
       Corrected Total             32     457.6733333

                       R-Square     Coeff Var      Root MSE    yield Mean
                       0.920074      9.321454      1.289468      13.83333

       Source                      DF       Type I SS     Mean Square    F Value    Pr > F
       variety                      2     347.3819444     173.6909722     104.46    <.0001
       density                      3      66.6531019      22.2177006      13.36    <.0001
       variety*density              5       7.0582870       1.4116574       0.85    0.5300

       Source                      DF     Type III SS     Mean Square    F Value    Pr > F
       variety                      2     321.2233796     160.6116898      96.60    <.0001
       density                      3      66.6531019      22.2177006      13.36    <.0001
       variety*density              5       7.0582870       1.4116574       0.85    0.5300
```

Two-way ANOVA using PROC GLM

```
                                                   Standard
Parameter                        Estimate             Error    t Value    Pr > |t|
Intercept                     18.16666667 B      0.74447460      24.40      <.0001
variety           A           -7.36666667 B      1.05284607      -7.00      <.0001
variety           B           -5.40000000 B      1.05284607      -5.13      <.0001
variety           C            0.00000000 B       .               .          .
density           10          -1.86666667 B      1.05284607      -1.77      0.0901
density           20          -0.06666667 B      1.05284607      -0.06      0.9501
density           30           1.76666667 B      1.05284607       1.68      0.1075
density           40           0.00000000 B       .               .          .
variety*density   A 10         0.26666667 B      1.48894919       0.18      0.8595
variety*density   A 20         1.70000000 B      1.48894919       1.14      0.2658
variety*density   A 30         0.33333333 B      1.48894919       0.22      0.8249
variety*density   A 40         0.00000000 B       .               .          .
variety*density   B 10        -1.96666667 B      1.48894919      -1.32      0.2001
variety*density   B 20        -0.06666667 B      1.48894919      -0.04      0.9647
variety*density   B 40         0.00000000 B       .               .          .
variety*density   C 10         0.00000000 B       .               .          .
variety*density   C 20         0.00000000 B       .               .          .
variety*density   C 30         0.00000000 B       .               .          .
variety*density   C 40         0.00000000 B       .               .          .
```

Notice the missing variety*density B 30 line.

Two-way ANOVA using PROC GLM

```
                  The GLM Procedure
                 Least Squares Means
  Adjustment for Multiple Comparisons: Tukey-Kramer

                                      LSMEAN
         variety     yield LSMEAN     Number
         A             11.3333333          1
         B                Non-est          2
         C             18.1250000          3

       Least Squares Means for effect variety
        Pr > |t| for H0: LSMean(i)=LSMean(j)


variety    yield LSMEAN      95% Confidence Limits
A             11.333333       10.561360    12.105306
B                     .               .            .
C             18.125000       17.353027    18.896973

        Least Squares Means for Effect variety

             Difference         Simultaneous 95%
                Between      Confidence Limits for
 i    j           Means       LSMean(i)-LSMean(j)
 1    2               .               .            .
 1    3       -6.791667       -7.883358    -5.699975
 2    3               .               .            .
```

Two-way ANOVA using PROC GLM

```
                  The GLM Procedure
                 Least Squares Means
  Adjustment for Multiple Comparisons: Tukey-Kramer

                                      LSMEAN
         density     yield LSMEAN     Number
         10            11.4777778          1
         20            14.3888889          2
         30               Non-est          3
         40            13.9111111          4

density    yield LSMEAN      95% Confidence Limits
10            11.477778      10.586380    12.369175
20            14.388889      13.497491    15.280286
30                    .              .            .
40            13.911111      13.019714    14.802509

        Least Squares Means for Effect density

             Difference         Simultaneous 95%
                Between      Confidence Limits for
 i    j           Means       LSMean(i)-LSMean(j)
 1    2       -2.911111      -4.438096    -1.384126
 1    3               .              .            .
 1    4       -2.433333      -3.960319    -0.906348
 2    3               .              .            .
 2    4        0.477778      -1.049207     2.004763
 3    4               .              .            .
```

Two-way ANOVA using PROC GLM

```
                 The GLM Procedure
                Least Squares Means
   Adjustment for Multiple Comparisons: Tukey

                                              LSMEAN
variety     density      yield LSMEAN         Number
A           10              9.2000000              1
A           20             12.4333333              2
A           30             12.9000000              3
A           40             10.8000000              4
B           10              8.9333333              5
B           20             12.6333333              6
B           40             12.7666667              7
C           10             16.3000000              8
C           20             18.1000000              9
C           30             19.9333333             10
```

Two-way ANOVA using PROC GLM

```
               Difference        Simultaneous 95%
                  Between     Confidence Limits for
i      j            Means      LSMean(i)-LSMean(j)
1      2        -3.233333      -6.997053     0.530387
1      3        -3.700000      -7.463720     0.063720
1      4        -1.600000      -5.363720     2.163720
1      5         0.266667      -3.497053     4.030387
1      6        -3.433333      -7.197053     0.330387
1      7        -3.566667      -7.330387     0.197053
1      8        -7.100000     -10.863720    -3.336280
1      9        -8.900000     -12.663720    -5.136280
1     10       -10.733333     -14.497053    -6.969613
1     11        -8.966667     -12.730387    -5.202947
2      3        -0.466667      -4.230387     3.297053
2      4         1.633333      -2.130387     5.397053
2      5         3.500000      -0.263720     7.263720
2      6        -0.200000      -3.963720     3.563720
2      7        -0.333333      -4.097053     3.430387
2      8        -3.866667      -7.630387    -0.102947
2      9        -5.666667      -9.430387    -1.902947
2     10        -7.500000     -11.263720    -3.736280
2     11        -5.733333      -9.497053    -1.969613
3      4         2.100000      -1.663720     5.863720
3      5         3.966667       0.202947     7.730387
3      6         0.266667      -3.497053     4.030387
3      7         0.133333      -3.630387     3.897053
3      8        -3.400000      -7.163720     0.363720
3      9        -5.200000      -8.963720    -1.436280
3     10        -7.033333     -10.797053    -3.269613
3     11        -5.266667      -9.030387    -1.502947
4      5         1.866667      -1.897053     5.630387
```

Treat as a One-way ANOVA

When the data are incomplete, use a one-way ANOVA combined with contrasts to answer questions of interest. For example, to compare the average difference between B and C, we want to only compare at densities 10, 20, and 40.

	10	20	30	40
A	μ_{11}	μ_{12}	μ_{13}	μ_{14}
B	μ_{21}	μ_{22}	μ_{23}	μ_{24}
C	μ_{31}	μ_{32}	μ_{33}	μ_{34}

Thus, the contrast is

$$\begin{aligned} \gamma &= \frac{1}{3}(\mu_{31}+\mu_{32}+\mu_{34}) - \frac{1}{3}(\mu_{21}+\mu_{22}+\mu_{24}) \\ &= \frac{1}{3}(\mu_{31}+\mu_{32}+\mu_{34}-\mu_{21}-\mu_{22}-\mu_{24}) \end{aligned}$$

Two-way ANOVA using PROC GLM

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;

PROC GLM DATA=tomato PLOTS=all;
  WHERE ~(variety='B' & density=30);
  CLASS variety density;
  MODEL yield = variety*density / SOLUTION CLPARM;
  LSMEANS variety*density / cl adjust=tukey;
  /*                             A10 A20 A30 A40 B10 B20 B40 C10 C20 C30 C40 */
  ESTIMATE 'C-B' variety*density  0   0   0   0  -1  -1  -1   1   1   0   1 / DIVISOR=3;
  ESTIMATE 'C-A' variety*density -1  -1  -1  -1   0   0   0   1   1   1   1 / DIVISOR=4;
  ESTIMATE 'B-A' variety*density -1  -1   0  -1   1   1   1   0   0   0   0 / DIVISOR=3;
  /* we could do the densities similarly */
  RUN;
```

Two-way ANOVA using PROC GLM

```
                                   The GLM Procedure
Dependent Variable: yield
                                                Sum of
      Source                      DF           Squares     Mean Square    F Value    Pr > F
      Model                       10       421.0933333      42.1093333      25.33    <.0001
      Error                       22        36.5800000       1.6627273
      Corrected Total             32       457.6733333

                    R-Square     Coeff Var      Root MSE    yield Mean
                    0.920074      9.321454      1.289468      13.83333

      Source                      DF         Type I SS     Mean Square    F Value    Pr > F
      variety*density             10       421.0933333      42.1093333      25.33    <.0001

      Source                      DF       Type III SS     Mean Square    F Value    Pr > F
      variety*density             10       421.0933333      42.1093333      25.33    <.0001
```

Two-way ANOVA using PROC GLM

```
                                           Standard
Parameter                  Estimate           Error  t Value  Pr > |t|    95% Confidence Limits
Intercept               18.16666667 B    0.74447460    24.40    <.0001   16.62272085  19.71061248
variety*density A 10    -8.96666667 B    1.05284607    -8.52    <.0001  -11.15013578  -6.78319756
variety*density A 20    -5.73333333 B    1.05284607    -5.45    <.0001   -7.91680244  -3.54986422
variety*density A 30    -5.26666667 B    1.05284607    -5.00    <.0001   -7.45013578  -3.08319756
variety*density A 40    -7.36666667 B    1.05284607    -7.00    <.0001   -9.55013578  -5.18319756
variety*density B 10    -9.23333333 B    1.05284607    -8.77    <.0001  -11.41680244  -7.04986422
variety*density B 20    -5.53333333 B    1.05284607    -5.26    <.0001   -7.71680244  -3.34986422
variety*density B 40    -5.40000000 B    1.05284607    -5.13    <.0001   -7.58346911  -3.21653089
variety*density C 10    -1.86666667 B    1.05284607    -1.77    0.0901   -4.05013578   0.31680244
variety*density C 20    -0.06666667 B    1.05284607    -0.06    0.9501   -2.25013578   2.11680244
variety*density C 30     1.76666667 B    1.05284607     1.68    0.1075   -0.41680244   3.95013578
variety*density C 40     0.00000000 B     .              .       .         .            .
```

The Regression model

The regression model here considers variety-density combination as a single explanatory variable with 11 levels: A10, A20, A30, A40, B10, B20, B40, C10, C20, C30, and C40. By default, SAS chose C40 as our reference level. For observation i, let

- Y_i be the yield
- V_i be the variety
- D_i be the density

The model is then $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ and

$$
\begin{array}{lllllll}
\mu_i & = \beta_0 & +\beta_1 \mathrm{I}(V_i = A, D_i = 10) & +\beta_2 \mathrm{I}(V_i = A, D_i = 20) & +\beta_3 \mathrm{I}(V_i = A, D_i = 30) & +\beta_4 \mathrm{I}(V_i = A, D_i = 40) \\
 & & +\beta_5 \mathrm{I}(V_i = B, D_i = 10) & +\beta_6 \mathrm{I}(V_i = B, D_i = 20) & & +\beta_7 \mathrm{I}(V_i = B, D_i = 40) \\
 & & +\beta_8 \mathrm{I}(V_i = C, D_i = 10) & +\beta_9 \mathrm{I}(V_i = C, D_i = 20) & +\beta_{10} \mathrm{I}(V_i = C, D_i = 30) &
\end{array}
$$

Two-way ANOVA using PROC GLM

```
                                The GLM Procedure

Dependent Variable: yield

                                        Standard
 Parameter                 Estimate        Error  t Value  Pr > |t|    95% Confidence Limits
 C-B                     6.07777778   0.60786096    10.00    <.0001    4.81715130   7.33840426
 C-A                     6.79166667   0.52642304    12.90    <.0001    5.69993211   7.88340122
 B-A                     0.63333333   0.60786096     1.04    0.3088   -0.62729315   1.89395981
```

Two-way ANOVA using PROC GLM

```
                   The GLM Procedure
                  Least Squares Means
       Adjustment for Multiple Comparisons: Tukey

                                                  LSMEAN
variety     density      yield LSMEAN             Number
A           10              9.2000000                  1
A           20             12.4333333                  2
A           30             12.9000000                  3
A           40             10.8000000                  4
B           10              8.9333333                  5
B           20             12.6333333                  6
B           40             12.7666667                  7
C           10             16.3000000                  8
C           20             18.1000000                  9
C           30             19.9333333                 10
```

Two-way ANOVA using PROC GLM

```
                   Difference        Simultaneous 95%
                      Between      Confidence Limits for
 i      j               Means       LSMean(i)-LSMean(j)
 1      2           -3.233333       -6.997053      0.530387
 1      3           -3.700000       -7.463720      0.063720
 1      4           -1.600000       -5.363720      2.163720
 1      5            0.266667       -3.497053      4.030387
 1      6           -3.433333       -7.197053      0.330387
 1      7           -3.566667       -7.330387      0.197053
 1      8           -7.100000      -10.863720     -3.336280
 1      9           -8.900000      -12.663720     -5.136280
 1     10          -10.733333      -14.497053     -6.969613
 1     11           -8.966667      -12.730387     -5.202947
 2      3           -0.466667       -4.230387      3.297053
 2      4            1.633333       -2.130387      5.397053
 2      5            3.500000       -0.263720      7.263720
 2      6           -0.200000       -3.963720      3.563720
 2      7           -0.333333       -4.097053      3.430387
 2      8           -3.866667       -7.630387     -0.102947
 2      9           -5.666667       -9.430387     -1.902947
 2     10           -7.500000      -11.263720     -3.736280
 2     11           -5.733333       -9.497053     -1.969613
 3      4            2.100000       -1.663720      5.863720
 3      5            3.966667        0.202947      7.730387
 3      6            0.266667       -3.497053      4.030387
 3      7            0.133333       -3.630387      3.897053
 3      8           -3.400000       -7.163720      0.363720
 3      9           -5.200000       -8.963720     -1.436280
 3     10           -7.033333      -10.797053     -3.269613
 3     11           -5.266667       -9.030387     -1.502947
 4      5            1.866667       -1.897053      5.630387
```

```
m = lm(Yield~Variety:Density, tomato, subset=!(Variety=='B' & Density==30))
anova(m)


Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq F value    Pr(>F)    
Variety:Density 10 421.09  42.109  25.326 8.563e-10 ***
Residuals       22  36.58   1.663                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tomato$VarietyDensity = factor(paste(tomato$Variety, tomato$Density, sep=""))
# Note the -1 in order to construct the contrast
m = lm(Yield~VarietyDensity-1, tomato, subset=!(Variety=='B' & Density==30))
#                  A10 A20 A30 A40 B10 B20 B40 C10 C20 C30 C40
K = rbind('C-B' = c(  0,  0,  0,  0, -1, -1, -1,  1,  1,  0,  1)/3,
          'C-A' = c( -1, -1, -1, -1,  0,  0,  0,  1,  1,  1,  1)/4,
          'B-A' = c( -1, -1,  0, -1,  1,  1,  1,  0,  0,  0,  0)/3)

library(multcomp)
t = glht(m, linfct=K)
#summary(t)
confint(t, calpha=univariate_calpha())
```

```


 Simultaneous Confidence Intervals

Fit: lm(formula = Yield ~ VarietyDensity - 1, data = tomato, subset = !(Variety ==
    "B" & Density == 30))

Quantile = 2.0739
95% confidence level


Linear Hypotheses:
         Estimate lwr     upr
C-B == 0  6.0778   4.8172  7.3384
C-A == 0  6.7917   5.6999  7.8834
B-A == 0  0.6333  -0.6273  1.8940
```

```
m = lm(Yield~Variety:Density, tomato, subset=!(Variety=='B' & Density==30))
lsmeans(m, pairwise~Variety:Density)


$lsmeans
 Variety Density    lsmean        SE df  lower.CL upper.CL
 C       10      16.300000 0.7444746 22 14.756054 17.84395
 A       10       9.200000 0.7444746 22  7.656054 10.74395
 B       10       8.933333 0.7444746 22  7.389388 10.47728
 C       20      18.100000 0.7444746 22 16.556054 19.64395
 A       20      12.433333 0.7444746 22 10.889388 13.97728
 B       20      12.633333 0.7444746 22 11.089388 14.17728
 C       30      19.933333 0.7444746 22 18.389388 21.47728
 A       30      12.900000 0.7444746 22 11.356054 14.44395
 B       30             NA        NA NA        NA       NA
 C       40      18.166667 0.7444746 22 16.622721 19.71061
 A       40      10.800000 0.7444746 22  9.256054 12.34395
 B       40      12.766667 0.7444746 22 11.222721 14.31061

Confidence level used: 0.95

$contrasts
 contrast        estimate       SE df t.ratio p.value
 C,10 - A,10   7.10000000 1.052846 22   6.744  <.0001
 C,10 - B,10   7.36666667 1.052846 22   6.997  <.0001
 C,10 - C,20  -1.80000000 1.052846 22  -1.710  0.8458
 C,10 - A,20   3.86666667 1.052846 22   3.673  0.0465
 C,10 - B,20   3.66666667 1.052846 22   3.483  0.0688
 C,10 - C,30  -3.63333333 1.052846 22  -3.451  0.0734
 C,10 - A,30   3.40000000 1.052846 22   3.229  0.1136
 C,10 - B,30           NA       NA NA      NA      NA
 C,10 - C,40  -1.86666667 1.052846 22  -1.773  0.8156
 C,10 - A,40   5.50000000 1.052846 22   5.224  0.0014
 C,10 - B,40   3.53333333 1.052846 22   3.356  0.0887
 A,10 - B,10   0.26666667 1.052846 22   0.253  1.0000
 A,10 - C,20  -8.90000000 1.052846 22  -8.453  <.0001
```

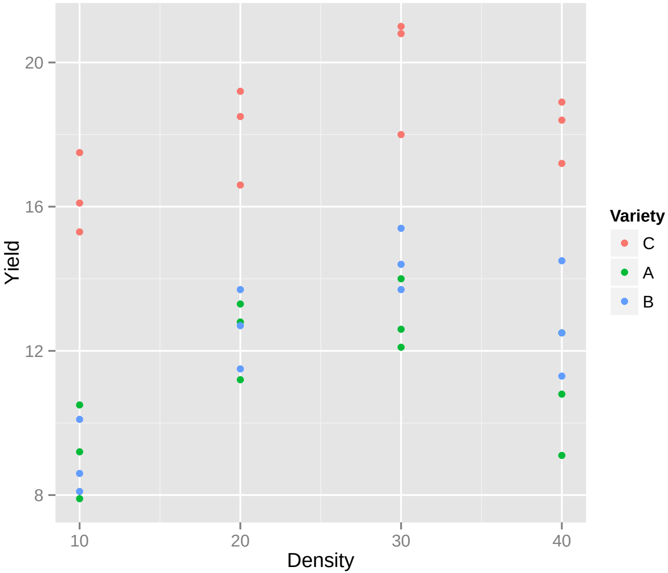
Summary

When dealing with an incomplete design, it is often easier to treat the analysis as a one-way ANOVA and use contrasts to answer scientific questions of interest.

Optimal yield

Now suppose you have the same data set, but your scientific question is different. Specifically, you are interested in choosing a variety and density that provide the optimal yield.

You can use the ANOVA analysis to choose from amongst the 3 varieties and one of the 4 densities, but there is no reason to believe that the optimal density will be one of those 4.


Yield
20
16
12
8
10
20
30
40
Density
Variety
C
A
B

Modeling

Considering a single variety, if we assume a linear relationship between Yield (Y_i) and Density (D_i) then the maximum Yield will occur at either $-\infty$ or $+\infty$ which is unreasonable. The easiest way to have a maximum (or minimum) is to assume a quadratic relationship, e.g.

$$E[Y_i] = \mu_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$$

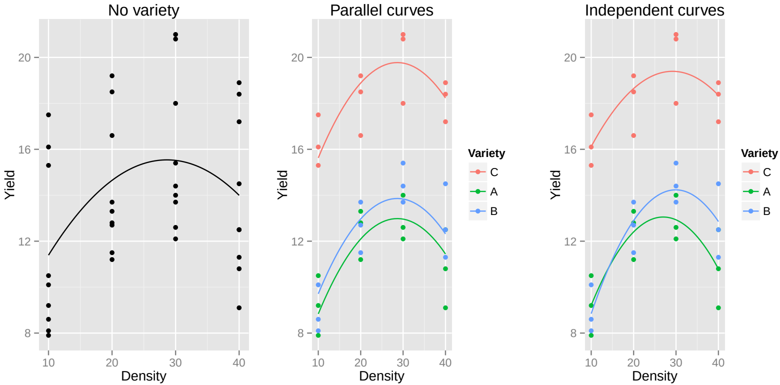
Now we can incorporate Variety (V_i) in many ways. Two options are parallel curves or completely independent curves.

Parallel curves:

$$\begin{aligned}\mu_i = \quad & \beta_0 + \beta_1 D_i + \beta_2 D_i^2 \\ & + \beta_3 \mathrm{I}(V_i = A) + \beta_4 \mathrm{I}(V_i = B)\end{aligned}$$

Independent lines:

$$\begin{aligned}\mu_i = \quad & \beta_0 + \beta_1 D_i + \beta_2 D_i^2 \\ & + \beta_3 \mathrm{I}(V_i = A) + \beta_4 \mathrm{I}(V_i = B) \\ & + \beta_5 \mathrm{I}(V_i = A) D_i + \beta_6 \mathrm{I}(V_i = B) D_i \\ & + \beta_7 \mathrm{I}(V_i = A) D_i^2 + \beta_8 \mathrm{I}(V_i = B) D_i^2\end{aligned}$$


No variety
Parallel curves
Independent curves
Yield
Density
20
16
12
8
10
20
30
40
Variety
C
A
B

Finding the maximum

For a particular variety, there will be an equation like

$$E[Y_i] = \mu_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$$

where these β_1 and β_2 need not correspond to any particular β_1 and β_2 we have discussed thus far.

If $\beta_2 < 0$, then the quadratic curve has a maximum and it occurs at $-\beta_1/2\beta_2$.

Analysis in SAS

```
DATA tomato;
  INFILE 'Ch13-tomato.csv' DSD FIRSTOBS=2;
  INPUT variety $ density yield;

/* No variety */
PROC GLM DATA=tomato PLOTS=all;
  CLASS variety; /* density is no longer here */
  MODEL yield = density|density / SOLUTION;
  RUN;

/* Parallel curves */
PROC GLM DATA=tomato PLOTS=all;
  CLASS variety; /* density is no longer here */
  MODEL yield = density|density variety/ SOLUTION;
  RUN;

/* Independent curves */
PROC GLM DATA=tomato PLOTS=all;
  CLASS variety; /* density is no longer here */
  MODEL yield = density|density|variety/ SOLUTION;
  RUN;
```

No variety

```
                                        The GLM Procedure

Dependent Variable: yield

                                                    Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                        2       85.3346667      42.6673333       3.75    0.0340
       Error                       33      375.0208889      11.3642694
       Corrected Total             35      460.3555556


...

       Source                      DF      Type III SS     Mean Square    F Value    Pr > F
       density                      1      65.30344358     65.30344358       5.75    0.0223
       density*density              1      51.36111111     51.36111111       4.52    0.0411

                                                         Standard
             Parameter                 Estimate             Error    t Value    Pr > |t|
             Intercept              5.744444444        3.12824210       1.84      0.0753
             density                0.684111111        0.28538383       2.40      0.0223
             density*density       -0.011944444        0.00561849      -2.13      0.0411
```

Parallel curves

```
                                   The GLM Procedure

Dependent Variable: yield

                                            Sum of
      Source                      DF       Squares     Mean Square    F Value    Pr > F
      Model                        4   412.9318889     103.2329722      67.48    <.0001
      Error                       31    47.4236667       1.5297957
      Corrected Total             35   460.3555556

...

      Source                      DF   Type III SS     Mean Square    F Value    Pr > F
      density                      1    65.3034436      65.3034436      42.69    <.0001
      density*density              1    51.3611111      51.3611111      33.57    <.0001
      variety                      2   327.5972222     163.7986111     107.07    <.0001

                                                      Standard
          Parameter                 Estimate             Error    t Value    Pr > |t|
          Intercept              9.980555556 B      1.18419286       8.43      <.0001
          density                0.684111111        0.10470690       6.53      <.0001
          density*density       -0.011944444        0.00206142      -5.79      <.0001
          variety        A      -6.791666667 B      0.50494153     -13.45      <.0001
          variety        B      -5.916666667 B      0.50494153     -11.72      <.0001
          variety        C       0.000000000 B       .                .          .
```

Independent curves

```
                                               Sum of
       Source                      DF         Squares     Mean Square    F Value    Pr > F
       Model                        8     419.8612222      52.4826528      34.99    <.0001
       Error                       27      40.4943333       1.4997901
       Corrected Total             35     460.3555556
...

       Source                      DF     Type III SS     Mean Square    F Value    Pr > F
       density                      1     65.30344358     65.30344358      43.54    <.0001
       density*density              1     51.36111111     51.36111111      34.25    <.0001
       variety                      2     21.66539427     10.83269713       7.22    0.0031
       density*variety              2      2.07850215      1.03925108       0.69    0.5088
       densit*densit*variet         2      1.65388889      0.82694444       0.55    0.5825

                                                         Standard
        Parameter                      Estimate             Error    t Value    Pr > |t|
        Intercept                   11.80833333 B      1.96836425       6.00      <.0001
        density                      0.52016667 B      0.17957029       2.90      0.0074
        density*density             -0.00891667 B      0.00353529      -2.52      0.0179
        variety              A      -8.45833333 B      2.78368742      -3.04      0.0052
        variety              B      -9.73333333 B      2.78368742      -3.50      0.0016
        variety              C       0.00000000 B        .               .         .
        density*variety      A       0.19916667 B      0.25395073       0.78      0.4397
        density*variety      B       0.29266667 B      0.25395073       1.15      0.2592
        density*variety      C       0.00000000 B        .               .         .
        densit*densit*variet A      -0.00441667 B      0.00499965      -0.88      0.3848
        densit*densit*variet B      -0.00466667 B      0.00499965      -0.93      0.3589
        densit*densit*variet C       0.00000000 B        .               .         .
```

No variety

```

Call:
lm(formula = Yield ˜ Density + I(Densityˆ2), data = tomato)

Residuals:
   Min     1Q Median     3Q    Max
-4.898 -2.721 -1.320  3.364  6.109

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.744444   3.128242   1.836   0.0753 .
Density       0.684111   0.285384   2.397   0.0223 *
I(Densityˆ2) -0.011944   0.005618  -2.126   0.0411 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.371 on 33 degrees of freedom
Multiple R-squared:  0.1854, Adjusted R-squared:  0.136
F-statistic: 3.755 on 2 and 33 DF,  p-value: 0.03395
```

Parallel curves

```
Call:
lm(formula = Yield ~ Density + I(Density^2) + Variety, data = tomato)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3422 -0.9039  0.1744  0.8082  2.1828 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.980556   1.184193   8.428 1.61e-09 ***
Density       0.684111   0.104707   6.534 2.71e-07 ***
I(Density^2) -0.011944   0.002061  -5.794 2.21e-06 ***
VarietyA     -6.791667   0.504942 -13.450 1.76e-14 ***
VarietyB     -5.916667   0.504942 -11.718 6.39e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 31 degrees of freedom
Multiple R-squared:  0.897, Adjusted R-squared:  0.8837 
F-statistic: 67.48 on 4 and 31 DF,  p-value: 7.469e-15
```

Independent curves

```
Call:
lm(formula = Yield ~ Density * Variety + I(Density^2) * Variety,
    data = tomato)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04500 -0.82125 -0.01417  0.94000  1.71000

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           11.808333   1.968364   5.999 2.12e-06 ***
Density                0.520167   0.179570   2.897  0.00739 **
VarietyA              -8.458333   2.783687  -3.039  0.00523 **
VarietyB              -9.733333   2.783687  -3.497  0.00165 **
I(Density^2)          -0.008917   0.003535  -2.522  0.01787 *
Density:VarietyA       0.199167   0.253951   0.784  0.43971
Density:VarietyB       0.292667   0.253951   1.152  0.25924
VarietyA:I(Density^2) -0.004417   0.005000  -0.883  0.38482
VarietyB:I(Density^2) -0.004667   0.005000  -0.933  0.35889
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.225 on 27 degrees of freedom
Multiple R-squared:  0.912, Adjusted R-squared:  0.886
F-statistic: 34.99 on 8 and 27 DF,  p-value: 2.678e-12
```

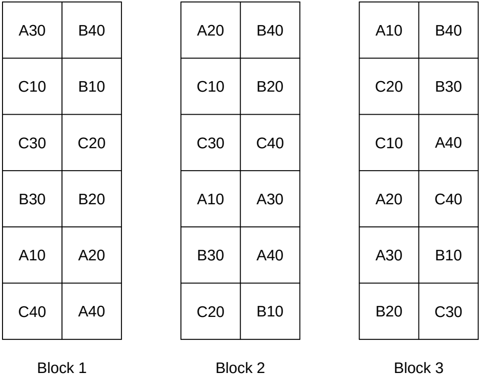
Completely randomized design (CRD)

This semester, we have assumed a completely randomized design. As an example, consider 36 plots and we are randomly assigning our variety-density combinations to the plots such that we have 3 reps of each combination. The result may look something like this

A20	A30	A40	C20	A40	B40
C20	C40	C40	B30	A10	A40
B40	C30	B40	C10	A20	C10
C10	B20	B20	A30	B10	A20
A10	C40	A10	B10	A30	B10
C20	B30	B20	B30	C30	C30

Complete randomized block design (RBD)

A randomized block design is appropriate when there is a nuisance factor that you want to control for. In our example, imagine you had 12 plots at 3 different locations and you expect these locations would have impact on yield. A randomized block design might look like this.



A30	B40
C10	B10
C30	C20
B30	B20
A10	A20
C40	A40

Block 1

A20	B40
C10	B20
C30	C40
A10	A30
B30	A40
C20	B10

Block 2

A10	B40
C20	B30
C10	A40
A20	C40
A30	B10
B20	C30

Block 3

RBD Analysis

Generally, you will want to model a randomized block design using an additive model for the treatment and blocking factor. If you have the replication, you should test for an interaction. Let's compute the degrees of freedom for the ANOVA tables for this current design considering the variety-density combination as the treatment.

V+D+B		T+B		Cell-means	
Factor	df	Factor	df	Factor	df
Variety	2				
Density	3	Treatment	11	Treatment	11
Block	2	Block	2	Block	2
				Treatment x Block	22
Error	28	Error	22	Error	0
Total	35	Total	35	Total	35

The cell-means model does not have enough degrees of freedom to estimate the interacion because there is no replication of the treatment within a block.

Why block?

Consider a simple experiment with 2 blocks each with 3 experimental units and 3 treatments (A, B, C).

Blocked

Block 1	Block 2
B	C
A	B
C	A

Unblocked

Block 1	Block 2
B	C
A	C
B	A

Let's consider 3 possible analyses:

- Blocked experiment using an additive model for treatment and block (RBD)
- Unblocked experiment using only treatment (CRD)
- Unblocked experiment using an additive model for treatment and block

Why block?

Now suppose, the true model is

$$\mu_{ij} = \mu + T_i + B_j$$

where $T_1 = T_2 = T_3$ and $B_1 = 0$ and $B_2 = \delta$.

In the Blocked experiment using an additive model for treatment and block, the expected treatment differences to all be zero.

In the Unblocked design using only treatment, the expected difference between treatments is

$$\mu_C - \mu_B = \delta \qquad \text{and} \qquad \mu_C - \mu_A = \delta/2.$$

In the Unblocked design using an additive model for treatment and block, we would have an unbalanced design and it would be impossible to compare B and C.

Summary

Block what you can control; randomize what you cannot.