

Set R09 - Two-way ANOVA

STAT 401 (Engineering) - Iowa State University

April 12, 2017

Two factors

Consider the question of the affect of variety and density on yield under various experimental designs:

- Balanced, complete design
- Unbalanced, complete
- Incomplete
- Optimization

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 \times 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.

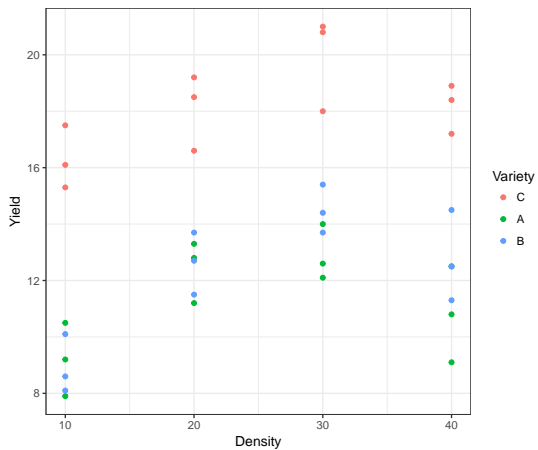
Hypotheses

- How does variety affect mean yield?
 - How is the mean yield for variety A different from B **on average**?
 - How is the mean yield for variety A different from B **at a particular value for density**?
- How does density affect mean yield?
 - How is the mean yield for density 10 different from density 20 **on average**?
 - How is the mean yield for density 10 different from density 20 **at a particular value for variety**?
- How 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/credible intervals can answer these questions.



Summary statistics

```
sm = tomato %>%
  group_by(Variety, Density) %>%
  summarize(n = n(),
            mean = mean(Yield),
            sd = sd(Yield))

sm
```

Source: local data frame [12 x 5]
Groups: Variety [?]

| | Variety | Density | n | mean | sd |
|----|---------|---------|-------|-----------|-----------|
| | <fctr> | <int> | <int> | <dbl> | <dbl> |
| 1 | C | 10 | 3 | 16.300000 | 1.1135529 |
| 2 | C | 20 | 3 | 18.100000 | 1.3453624 |
| 3 | C | 30 | 3 | 19.933333 | 1.6772994 |
| 4 | C | 40 | 3 | 18.166667 | 0.8736895 |
| 5 | A | 10 | 3 | 9.200000 | 1.3000000 |
| 6 | A | 20 | 3 | 12.433333 | 1.0969655 |
| 7 | A | 30 | 3 | 12.900000 | 0.9848858 |
| 8 | A | 40 | 3 | 10.800000 | 1.7000000 |
| 9 | B | 10 | 3 | 8.933333 | 1.0408330 |
| 10 | B | 20 | 3 | 12.633333 | 1.1015141 |
| 11 | B | 30 | 3 | 14.500000 | 0.8544004 |
| 12 | B | 40 | 3 | 12.766667 | 1.6165808 |

Two-way ANOVA

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

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

where Y_{ijk} is the

- k th observation at the
- i th level of variable 1 (variety) with $i = 1, \dots, I$ and the
- j th level of variable 2 (density) with $j = 1, \dots, J$.

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. $I(V_i = A)$ and $I(D_i = 10)$.
4. The additive model:

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

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

5. The cell-means model:

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

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 | $n-1$ | | |

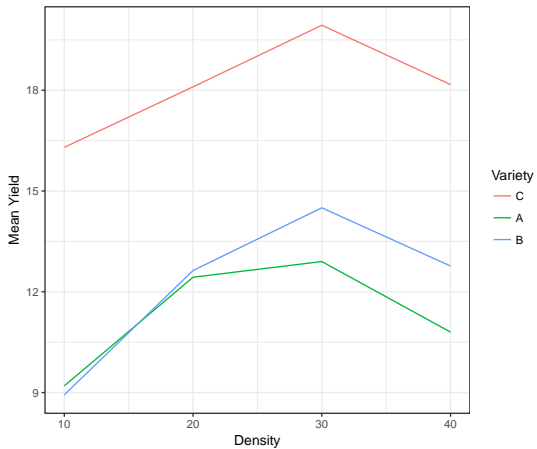
Additive vs cell-means

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.

```
ggplot(sm, aes(x=Density, y=mean, col=Variety)) + geom_line() + labs(y="Mean Yield") + theme_bw()
```



Two-way ANOVA in R

```
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

Variety comparison

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

```
$lsmmeans
  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 comparing a family of 3 estimates

Density comparison

```
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 comparing a family of 4 estimates

```
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 | 2.83333333 | 1.027943 | 24 | 2.754 | 0.1084 |
| A,10 - B,20 | 2.83333333 | 1.027943 | 24 | 2.754 | 0.1084 |
| A,10 - C,30 | -3.63333333 | 1.027943 | 24 | -3.535 | 0.0582 |
| A,10 - A,30 | 3.40000000 | 1.027943 | 24 | 3.308 | 0.0932 |
| A,10 - B,30 | 1.80000000 | 1.027943 | 24 | 1.751 | 0.8276 |
| A,10 - C,40 | -1.86666667 | 1.027943 | 24 | -1.816 | 0.7947 |
| A,10 - A,40 | 5.50000000 | 1.027943 | 24 | 5.350 | 0.0008 |
| B,10 - B,20 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| B,10 - B,30 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| B,10 - B,40 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| B,20 - B,30 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| B,20 - B,40 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| B,30 - B,40 | 0.26666667 | 1.027943 | 24 | 0.259 | 1.0000 |
| C,20 - C,30 | -1.80000000 | 1.027943 | 24 | -1.751 | 0.8276 |
| C,20 - C,40 | -1.86666667 | 1.027943 | 24 | -1.816 | 0.7947 |
| C,30 - C,40 | -0.06666667 | 1.027943 | 24 | -0.065 | 0.9814 |

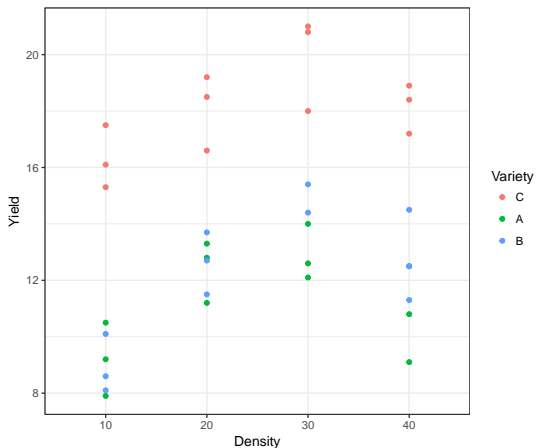
Summary

- 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.

```
tomato_unbalanced = tomato[-19,]
ggplot(tomato_unbalanced, aes(x=Density, y=Yield, color=Variety)) + geom_point() + theme_bw()
```



Summary statistics

```
sm_unbalanced = tomato_unbalanced %>%
  group_by(Variety, Density) %>%
  summarize(n = n(),
            mean = mean(Yield),
            sd = sd(Yield))
sm_unbalanced
```

Source: local data frame [12 x 5]
Groups: Variety [?]

| | Variety | Density | n | mean | sd |
|----|---------|---------|-------|-----------|-----------|
| | <fctr> | <fctr> | <int> | <dbl> | <dbl> |
| 1 | C | 10 | 3 | 16.300000 | 1.1135529 |
| 2 | C | 20 | 3 | 18.100000 | 1.3453624 |
| 3 | C | 30 | 3 | 19.933333 | 1.6772994 |
| 4 | C | 40 | 3 | 18.166667 | 0.8736895 |
| 5 | A | 10 | 3 | 9.200000 | 1.3000000 |
| 6 | A | 20 | 3 | 12.433333 | 1.0969655 |
| 7 | A | 30 | 3 | 12.900000 | 0.9848858 |
| 8 | A | 40 | 3 | 10.800000 | 1.7000000 |
| 9 | B | 10 | 3 | 8.933333 | 1.0408330 |
| 10 | B | 20 | 3 | 12.633333 | 1.1015141 |
| 11 | B | 30 | 2 | 14.900000 | 0.7071068 |
| 12 | B | 40 | 3 | 12.766667 | 1.6165808 |

Two-way ANOVA in R

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

Analysis of Variance Table

Response: Yield

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|---------------|
| Variety | 2 | 329.99 | 164.994 | 102.343 | 3.552e-12 *** |
| Density | 3 | 84.45 | 28.150 | 17.461 | 3.947e-06 *** |
| Variety:Density | 6 | 8.80 | 1.467 | 0.910 | 0.5052 |
| Residuals | 23 | 37.08 | 1.612 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Variety comparison

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

```
$lsmeans
```

| Variety | lsmean | SE | df | lower.CL | upper.CL |
|---------|----------|-----------|----|----------|----------|
| C | 18.12500 | 0.3665349 | 23 | 17.36676 | 18.88324 |
| A | 11.33333 | 0.3665349 | 23 | 10.57510 | 12.09157 |
| B | 12.30833 | 0.3887690 | 23 | 11.50410 | 13.11256 |

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 | 0.5183586 | 23 | 13.102 | <.0001 |
| C - B | 5.816667 | 0.5343118 | 23 | 10.886 | <.0001 |
| A - B | -0.975000 | 0.5343118 | 23 | -1.825 | 0.1839 |

Results are averaged over the levels of: Density

P value adjustment: tukey method for comparing a family of 3 estimates

Density comparison

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

```
$lsmeans
```

| Density | lsmean | SE | df | lower.CL | upper.CL |
|---------|----------|-----------|----|----------|----------|
| 10 | 11.47778 | 0.4232380 | 23 | 10.60224 | 12.35331 |
| 20 | 14.38889 | 0.4232380 | 23 | 13.51335 | 15.26442 |
| 30 | 15.91111 | 0.4571493 | 23 | 14.96543 | 16.85680 |
| 40 | 13.91111 | 0.4232380 | 23 | 13.03558 | 14.78665 |

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 | 0.5985490 | 23 | -4.864 | 0.0004 |
| 10 - 30 | -4.4333333 | 0.6229895 | 23 | -7.116 | <.0001 |
| 10 - 40 | -2.4333333 | 0.5985490 | 23 | -4.065 | 0.0025 |
| 20 - 30 | -1.5222222 | 0.6229895 | 23 | -2.443 | 0.0967 |
| 20 - 40 | 0.4777778 | 0.5985490 | 23 | 0.798 | 0.8545 |
| 30 - 40 | 2.0000000 | 0.6229895 | 23 | 3.210 | 0.0189 |

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

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

```
$lsmeans
```

| Variety | Density | lsmean | SE | df | lower.CL | upper.CL |
|---------|---------|-----------|-----------|----|-----------|----------|
| C | 10 | 16.300000 | 0.7330698 | 23 | 14.783530 | 17.81647 |
| A | 10 | 9.200000 | 0.7330698 | 23 | 7.683530 | 10.71647 |
| B | 10 | 8.933333 | 0.7330698 | 23 | 7.416863 | 10.44980 |
| C | 20 | 18.100000 | 0.7330698 | 23 | 16.583530 | 19.61647 |
| A | 20 | 12.433333 | 0.7330698 | 23 | 10.916863 | 13.94980 |
| B | 20 | 12.633333 | 0.7330698 | 23 | 11.116863 | 14.14980 |
| C | 30 | 19.933333 | 0.7330698 | 23 | 18.416863 | 21.44980 |
| A | 30 | 12.900000 | 0.7330698 | 23 | 11.383530 | 14.41647 |
| B | 30 | 14.900000 | 0.8978235 | 23 | 13.042711 | 16.75729 |
| C | 40 | 18.166667 | 0.7330698 | 23 | 16.650196 | 19.68314 |
| A | 40 | 10.800000 | 0.7330698 | 23 | 9.283530 | 12.31647 |
| B | 40 | 12.766667 | 0.7330698 | 23 | 11.250196 | 14.28314 |

Confidence level used: 0.95

```
$contrasts
```

| contrast | estimate | SE | df | t.ratio | p.value |
|-------------|-------------|-----------|----|---------|---------|
| C,10 - A,10 | 7.10000000 | 1.036717 | 23 | 6.849 | <.0001 |
| C,10 - B,10 | 7.36666667 | 1.036717 | 23 | 7.106 | <.0001 |
| C,10 - C,20 | -1.80000000 | 1.036717 | 23 | -1.736 | 0.8341 |
| C,10 - A,20 | 3.86666667 | 1.036717 | 23 | 3.730 | 0.0396 |
| C,10 - B,20 | 3.66666667 | 1.036717 | 23 | 3.537 | 0.0597 |
| C,10 - C,30 | -3.63333333 | 1.036717 | 23 | -3.505 | 0.0638 |
| C,10 - A,30 | 3.40000000 | 1.036717 | 23 | 3.280 | 0.1008 |
| C,10 - B,30 | 1.40000000 | 1.159085 | 23 | 1.208 | 0.9828 |
| C,10 - C,40 | -1.86666667 | 1.036717 | 23 | -1.801 | 0.8022 |
| C,10 - A,40 | 5.50000000 | 1.036717 | 23 | 5.305 | 0.0011 |
| C,10 - B,40 | 3.53333333 | 1.036717 | 23 | 3.408 | 0.0778 |
| A,10 - B,10 | 0.26666667 | 1.036717 | 23 | 0.257 | 1.0000 |
| A,10 - C,20 | -8.90000000 | 1.036717 | 23 | -8.585 | <.0001 |
| A,10 - A,20 | 2.83333333 | 1.036717 | 23 | 2.736 | 0.1235 |
| A,10 - B,20 | 2.83333333 | 1.036717 | 23 | 2.736 | 0.1235 |
| A,10 - C,30 | -3.63333333 | 1.036717 | 23 | -3.505 | 0.0638 |
| A,10 - A,30 | 3.40000000 | 1.036717 | 23 | 3.280 | 0.1008 |
| A,10 - B,30 | 1.40000000 | 1.159085 | 23 | 1.208 | 0.9828 |
| A,10 - C,40 | -1.86666667 | 1.036717 | 23 | -1.801 | 0.8022 |
| A,10 - A,40 | 5.50000000 | 1.036717 | 23 | 5.305 | 0.0011 |
| A,10 - B,40 | 3.53333333 | 1.036717 | 23 | 3.408 | 0.0778 |
| B,10 - C,20 | -1.80000000 | 1.036717 | 23 | -1.736 | 0.8341 |
| B,10 - A,20 | 3.86666667 | 1.036717 | 23 | 3.730 | 0.0396 |
| B,10 - B,20 | 3.66666667 | 1.036717 | 23 | 3.537 | 0.0597 |
| B,10 - C,30 | -3.63333333 | 1.036717 | 23 | -3.505 | 0.0638 |
| B,10 - A,30 | 3.40000000 | 1.036717 | 23 | 3.280 | 0.1008 |
| B,10 - B,30 | 1.40000000 | 1.159085 | 23 | 1.208 | 0.9828 |
| B,10 - C,40 | -1.86666667 | 1.036717 | 23 | -1.801 | 0.8022 |
| B,10 - A,40 | 5.50000000 | 1.036717 | 23 | 5.305 | 0.0011 |
| B,10 - B,40 | 3.53333333 | 1.036717 | 23 | 3.408 | 0.0778 |
| C,20 - A,20 | 8.93333333 | 0.7330698 | 23 | 12.187 | <.0001 |
| C,20 - B,20 | 9.16666667 | 0.7330698 | 23 | 12.500 | <.0001 |
| C,20 - C,30 | -6.46666667 | 0.7330698 | 23 | -8.821 | <.0001 |
| C,20 - A,30 | -6.46666667 | 0.7330698 | 23 | -8.821 | <.0001 |
| C,20 - B,30 | -4.43333333 | 0.7330698 | 23 | -6.048 | <.0001 |
| C,20 - C,40 | -6.46666667 | 0.7330698 | 23 | -8.821 | <.0001 |
| C,20 - A,40 | -6.46666667 | 0.7330698 | 23 | -8.821 | <.0001 |
| C,20 - B,40 | -4.43333333 | 0.7330698 | 23 | -6.048 | <.0001 |
| C,30 - A,30 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,30 - B,30 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,30 - C,40 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,30 - A,40 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,30 - B,40 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,40 - A,40 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |
| C,40 - B,40 | 17.03333333 | 0.7330698 | 23 | 23.236 | <.0001 |

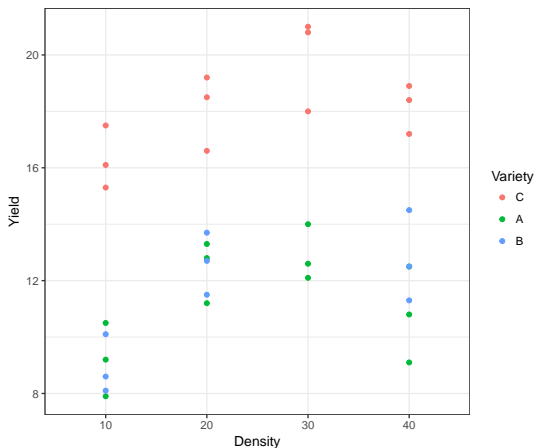
Summary

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.

```
tomato_incomplete = tomato %>%
  filter(!(Variety == "B" & Density == 30)) %>%
  mutate(VarietyDensity = paste0(Variety,Density))
ggplot(tomato_incomplete, aes(x=Density, y=Yield, color=Variety)) + geom_point() + theme_bw()
```



Summary statistics

```
sm_incomplete = tomato_incomplete %>%
  group_by(Variety, Density) %>%
  summarize(n      = n(),
            mean    = mean(Yield),
            sd      = sd(Yield))
sm_incomplete
```

Source: local data frame [11 x 5]

Groups: Variety [?]

| | Variety | Density | n | mean | sd |
|----|---------|---------|-------|-----------|-----------|
| | <fctr> | <fctr> | <int> | <dbl> | <dbl> |
| 1 | C | 10 | 3 | 16.300000 | 1.1135529 |
| 2 | C | 20 | 3 | 18.100000 | 1.3453624 |
| 3 | C | 30 | 3 | 19.933333 | 1.6772994 |
| 4 | C | 40 | 3 | 18.166667 | 0.8736895 |
| 5 | A | 10 | 3 | 9.200000 | 1.3000000 |
| 6 | A | 20 | 3 | 12.433333 | 1.0969655 |
| 7 | A | 30 | 3 | 12.900000 | 0.9848858 |
| 8 | A | 40 | 3 | 10.800000 | 1.7000000 |
| 9 | B | 10 | 3 | 8.933333 | 1.0408330 |
| 10 | B | 20 | 3 | 12.633333 | 1.1015141 |
| 11 | B | 40 | 3 | 12.766667 | 1.6165808 |

Treat as a One-way ANOVA

When the design is 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

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

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. Let C40 be the 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 $Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$ and

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

Two-way ANOVA in R

```
m <- lm(Yield ~ Variety*Density, data=tomato_incomplete)
anova(m)
```

Analysis of Variance Table

Response: Yield

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|-----------------|----|--------|---------|---------|---------------|
| Variety | 2 | 347.38 | 173.691 | 104.462 | 5.868e-12 *** |
| Density | 3 | 66.65 | 22.218 | 13.362 | 3.514e-05 *** |
| Variety:Density | 5 | 7.06 | 1.412 | 0.849 | 0.53 |
| Residuals | 22 | 36.58 | 1.663 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

How can you tell the design is not complete?

One-way ANOVA in R

```
m = lm(Yield~Variety:Density, tomato_incomplete)
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

Contrasts

```
# Note the -1 in order to construct the contrast
m = lm(Yield~VarietyDensity-1, tomato_incomplete)
#
#           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_incomplete)
```

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_incomplete)
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 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - A,20 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - B,20 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - C,30 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - A,30 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - B,30 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - C,40 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - A,40 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - B,40 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |
| A,10 - C,40 | 0.00000000 | 1.052846 | 22 | 0.000 | 1.0000 |

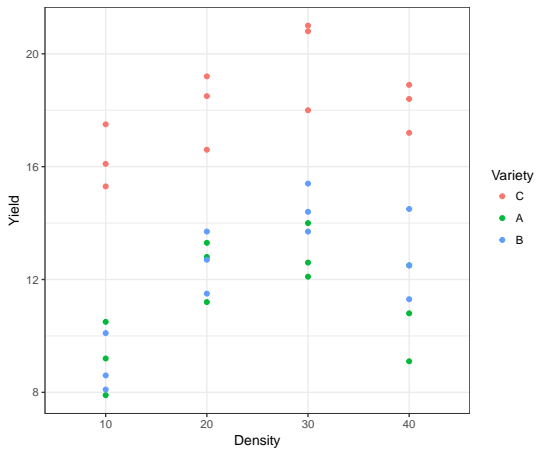
Summary

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-density combination that provides 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.



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.

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

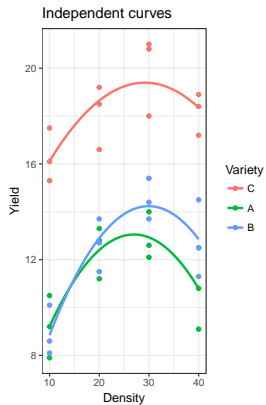
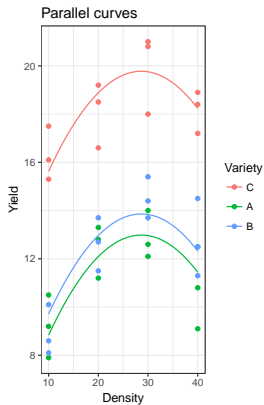
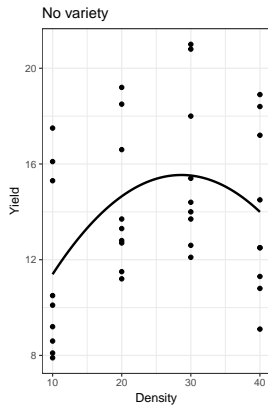
Now we can incorporate Variety (V_i) in many ways. Two options are parallel curves or completely independent curves.

Parallel curves:

$$\mu_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \beta_3 I(V_i = A) + \beta_4 I(V_i = B)$$

Independent lines:

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



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 $-\beta_1/2\beta_2$.

No variety

```
summary(lm(Yield~Density+I(Density^2), tomato))
```

Call:
lm(formula = Yield ~ Density + I(Density^2), 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(Density^2) | -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

```
summary(lm(Yield~Density+I(Density^2) + Variety, tomato))
```

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

```
summary(lm(Yield~Density*Variety+I(Density^2)*Variety, tomato))
```

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 interaction 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 | | Unblocked | |
|---------|---------|-----------|---------|
| B | C | B | C |
| A | B | A | C |
| C | A | B | A |
| Block 1 | Block 2 | Block 1 | Block 2 |

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

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

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.