Lesson 9 - Strawberries

LTC James K. Starling

Last compiled on 31 January, 2022

A study (Smith & Skog, “Postharvest Carbon Dioxide Treatment Enhances Firmness of Several Cultivars of Strawberry,” American Society for Horticulture Science, 27(5), 1992) was conducted in order to investigate the effects of storage method on the firmness of strawberries. As strawberries are harvested, they are placed into storage containers called “clamshells.” Because clamshells have holes in them, air can circulate from outside the clamshell into the clamshell and vice versa while the container still protects the fruit. To test the potential effect of changing the composition of the air in which strawberries are stored, researchers obtained 15 clamshells filled with strawberries of several different varieties. Five clamshells were randomly assigned to be stored in regular air (approximately 21% oxygen and 0.04% carbon dioxide) and five clamshells were randomly assigned to be stored in a modified air method (18% oxygen and 15% carbon dioxide). After two days of refrigeration at 0.5°C in one of the two different atmospheres, the strawberries were then left at room temperature in regular air for two more days. The firmness of the strawberries was then measured by the force (measured in newtons, N) needed for a probe to pierce the exterior of the strawberry. The remaining five clamshells were randomly assigned to a “control” condition which measured the strawberries immediately (not after four days like the other two conditions).

```
### Loading packages
library(tidyverse)
library(ggResidpanel)
## Load data

## three ways to read in the data:
# berries <- read.table('http://www.isi-stats.com/isi2/data/StrawberryStorageRCBD.txt',
#                        header = TRUE, stringsAsFactors = TRUE)
# berries <- read.table('https://raw.githubusercontent.com/jkstarling/MA376/main/StrawberryStorageRCBD.txt',
#                        header = TRUE, stringsAsFactors = TRUE)

setwd('C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons/Block II/LSN9/')
berries <- read.table('StrawberryStorageRCBD.txt', header = TRUE, stringsAsFactors = TRUE)
```

Background

1. Identify the observational units, the explanatory variable, and the response variable. Identify hypothesized Sources of explained variation and unexplained variation.

Observational Units: strawberry clamshells

Response Variable: Strawberry Firmness

Explanatory Variable: Air Composition.

Hypothesized Explained Variation: Air Composition

Hypothesized Unexplained Variation: Initial firmness, size, unknown

2. Calculate the effects for each storage method. (Hint: use `summarise()` and `group_by()` as necessary)

```
#Calculate the overall average
Mean <- mean(berries$Firmness)
#calculate the group effects
berries %>% group_by(Storage) %>% summarise(mean(Firmness)-Mean)
```

Storage	mean(Firmness) - Mean
<fct>	<dbl>
Air	-0.66
Control	-0.56
ModifiedAir	1.22

3 rows

3. Does the storage method appear to explain some of the variability in the firmness levels? How are you deciding?

Yes, the modified air strawberries tend to be the firmest of the air compositions (there are noticeable differences among the three group means

4. Calculate the ANOVA table for the (one variable) separate means model. (Use `aov()` or a linear model with `anova` to display the output.)

What is the proportion of variation explained by storage method?

```
anova.simple = aov(Firmness~Storage, data = berries)
summary(anova.simple)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Storage      2  11.19   5.594   0.995  0.398
## Residuals   12  67.44   5.620
```

```
#or
model.simple = lm(Firmness~Storage, data = berries)
anova(model.simple)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Storage	2	11.188	5.594000	0.9954327	0.3981255
Residuals	12	67.436	5.619667	NA	NA

2 rows

```
#R-squared
11.188/(11.188+67.436)
```

```
## [1] 0.1422975
```

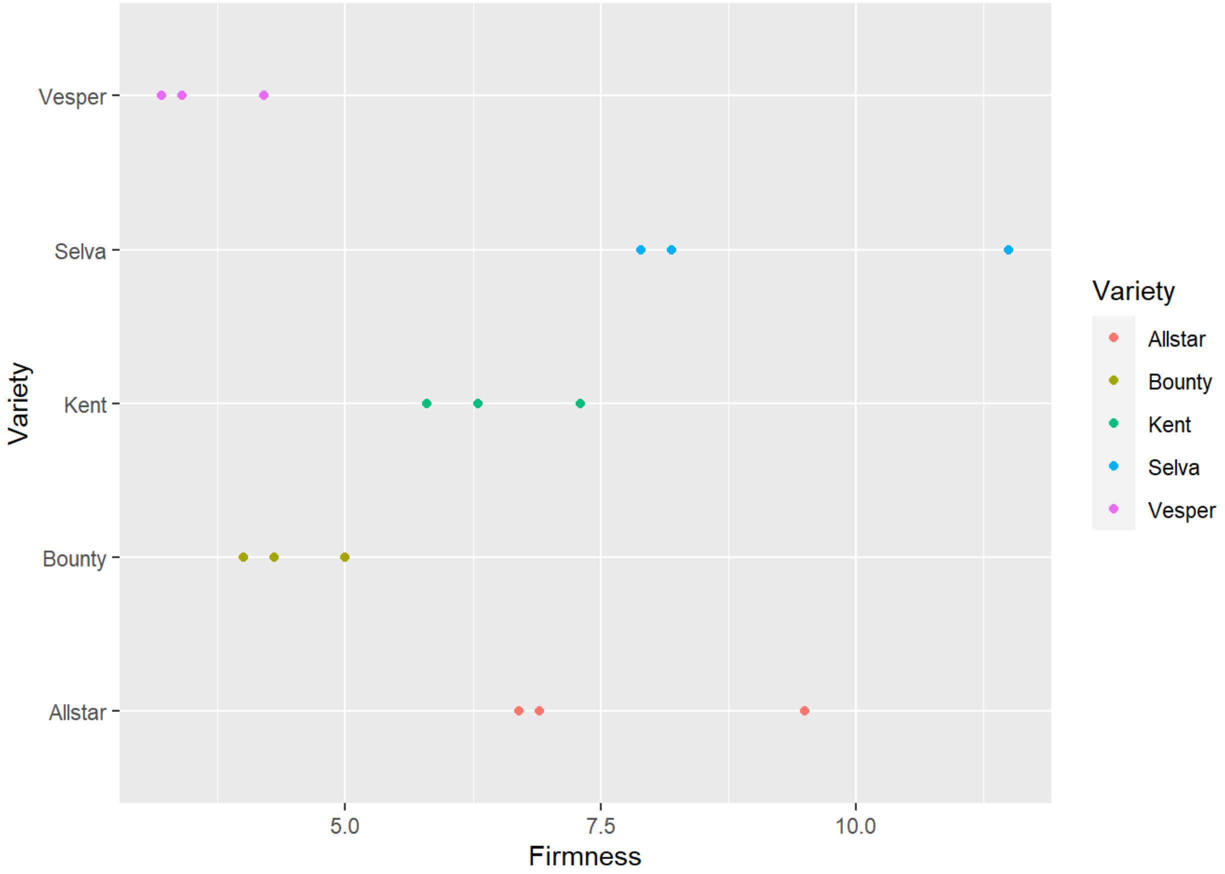
The storage method explains 14.2% of the variation in firmness.

Model Update

Rather than limiting themselves to one type of strawberry (i.e., using inclusion criteria of one type of strawberry) which would most likely have given them a smaller SSError, the researchers decided to include five different varieties of strawberries in the study allowing them to generalize to a more diverse population of strawberries. Suspecting that different varieties of strawberries have different firmness ratings, that is, believing variety is associated with firmness (the response variable), the researchers opted to **block** on strawberry variety. In this way the variation explained by variety will be accounted for in the analysis by removing it from the unexplained variation.

5. Create a plot of Firmness vs Variety? (Use `geom_point()` with `ggplot()` . Set `x=Firmness` , `y=Variety` , and `color=Variety`) Do the data confirm or refute the researchers’ prediction that different strawberry varieties tend to have different firmness ratings? Is there a strawberry variety that tends to have very low firmness? Very high?

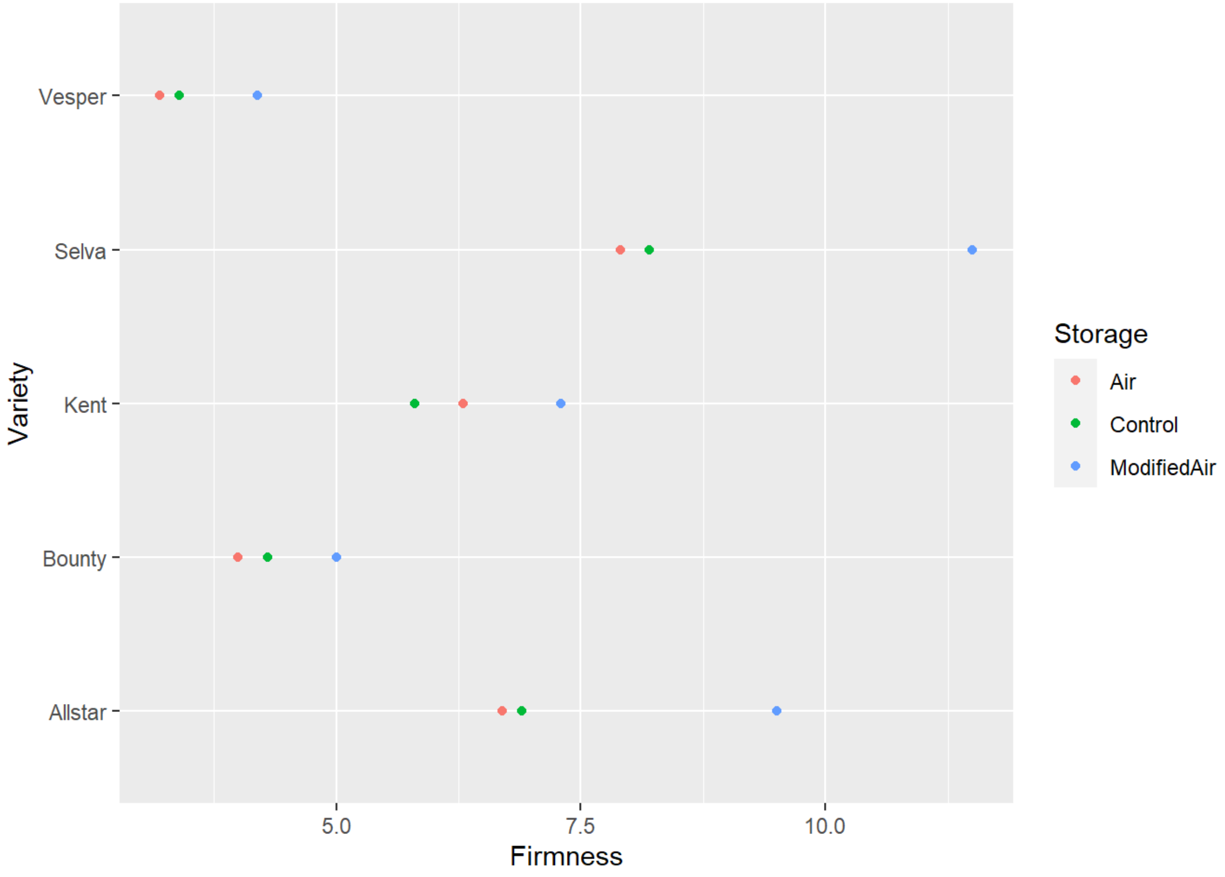
```
berries %>% ggplot(aes(x = Firmness, y = Variety, color = Variety))+
  geom_point()
```



The data confirms that strawberry varieties tend to have different firmness ratings. Vesper tends to have a low firmness whereas Selva tends to have a high firmness.

6. Color your plot from the previous question by `Storage`. For the Vesper variety, which storage method had the largest firmness? Is this the case for all of the varieties? Looking at this graph, do you think the storage method effect will be statistically significant after adjusting for variety? Explain your reasoning.

```
berries %>% ggplot(aes(x = Firmness, y = Variety, color = Storage))+
  geom_point()
```



Modified air has the largest firmness for Vesper and this extends to all varieties. Storage method will be statistically significant after adjusting for variety because modified air is noticeably higher for each variety.

A block study design creates blocks of experimental units that are similar to each other, randomly assigns the treatments within each block, and then analyzes the data in a way which accounts for block-to-block variations. When there are only two groups being compared a block study design is called a matched pairs design. The term block comes from the first block designs which were agricultural experiments in large fields where separate parts of the field were called "blocks."

7. How would we implement randomization when we block on strawberry type?

Since we have three clamshells of Allstar strawberries (and the other types) we can randomly place into three clamshells and then randomly assign to one of the three storage methods to each clamshell. This restricted random assignment guarantees that there is one clamshell of each variety for each storage method. General rule: block what you can, randomize what you cannot.

8. Consider the study we did last lesson with baking chips (chocolate and butterscotch). What were the 'blocks' that we considered? Explain how exploring the melting time of a white-chocolate chip would make the study a block study design.

The blocks in the previous chip study were the participants (students). If we were to introduce a white chocolate chip, we would need have each student melt a white chocolate chip in addition to the chocolate and butterscotch chips. For each student we would have to randomize the order in which they melt the three chips.

Modifying the Analysis to Consider Blocking. Now that we know more about the actual way this study was designed, we know that an analysis that ignores this “restricted randomization” is incorrect. Instead, the analysis should reflect the study design that was used (and account for the blocking).

9. Conduct an analysis of a Two-variable ANOVA accounting for treatment and blocks. (Use `aov()` or a linear model with `anova()` (after using effect encoding)). Calculate the R-squared value. Compare the results with the single-variable ANOVA calculated above.

```
## Using AOV
blocked.aov <- aov(Firmness~Storage+Variety, data = berries)
summary(blocked.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Storage      2  11.19   5.594   11.40  0.00455 ** 
## Variety      4  63.51  15.878   32.36 5.47e-05 ***
## Residuals    8   3.93   0.491                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Calculating an anova table from a linear regression model using Effect Encoding to account for variety
berries2 <- berries
contrasts(berries2$Storage) = contr.sum
contrasts(berries2$Variety) = contr.sum

blocked.model <- lm(Firmness ~ Storage + Variety, data = berries2)
summary(blocked.model)
```

```
##
## Call:
## lm(formula = Firmness ~ Storage + Variety, data = berries2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6533 -0.4133 -0.1067  0.3933  1.0800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.2800     0.1809  34.723 5.18e-10 ***
## Storage1     -0.6600     0.2558  -2.580 0.032598 *
## Storage2     -0.5600     0.2558  -2.189 0.059977 .
## Variety1      1.4200     0.3617   3.926 0.004383 **
## Variety2     -1.8467     0.3617  -5.105 0.000924 ***
## Variety3      0.1867     0.3617   0.516 0.619778
## Variety4      2.9200     0.3617   8.072 4.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7005 on 8 degrees of freedom
## Multiple R-squared:  0.9501, Adjusted R-squared:  0.9126
## F-statistic: 25.37 on 6 and 8 DF,  p-value: 8.59e-05
```

```
anova(blocked.model)
```

	Df <int>	Sum Sq <dbl>	Mean Sq <dbl>	F value <dbl>	Pr(>F) <dbl>
Storage	2	11.188000	5.5940000	11.40082	4.550563e-03
Variety	4	63.510667	15.8776667	32.35938	5.472671e-05
Residuals	8	3.925333	0.4906667	NA	NA

3 rows

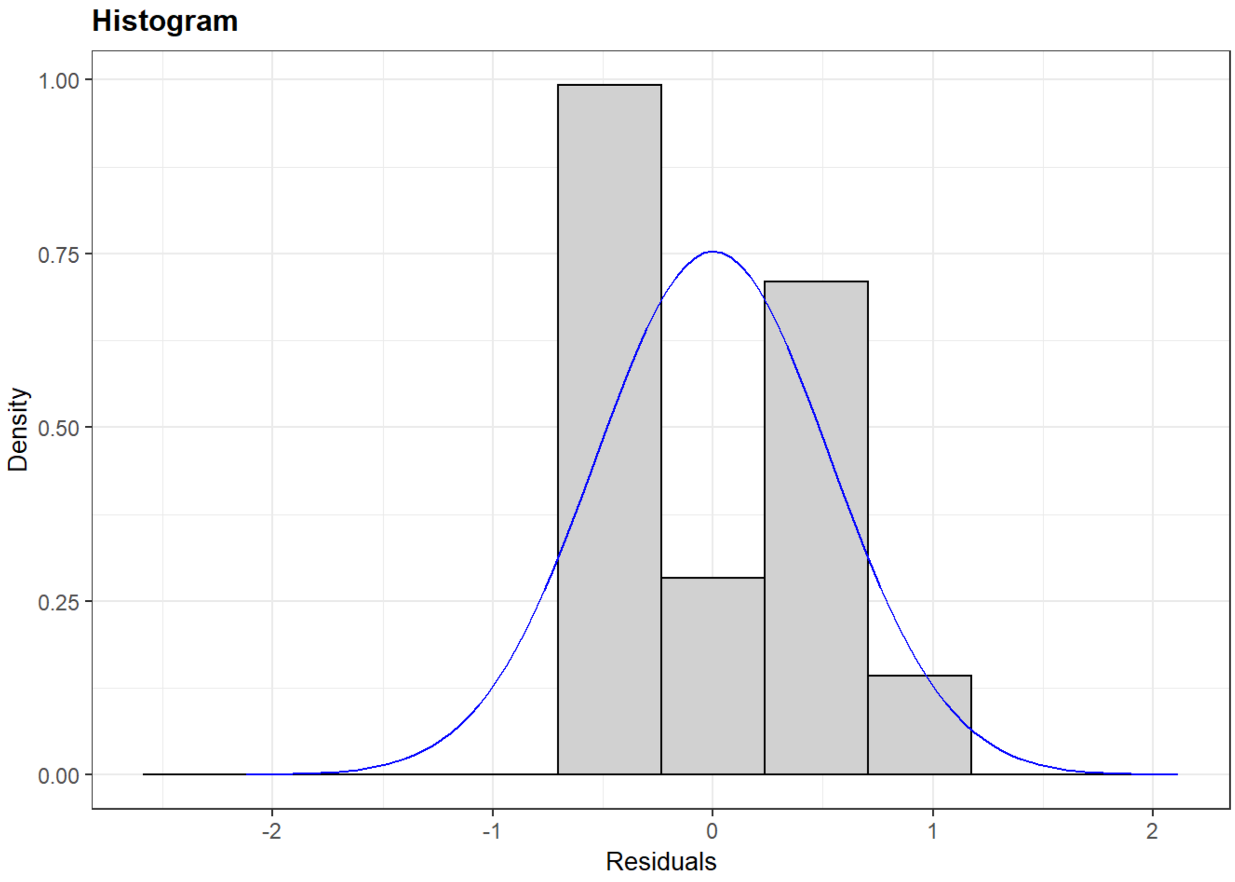
```
(11.118+63.511)/(11.118+63.511+3.925)
```

```
## [1] 0.9500344
```

We see that the r-squared for the model that accounts for blocking is 95% vs 14.2%. Thus our model accounts for much more of the variability than the single ANOVA model (without blocking).

10. Do we meet the validity conditions for a two variable ANOVA? (see p. 163)

```
resid_panel(blocked.model, plot = "hist", bins = 10)
```



```
berries %>% group_by(Storage) %>% summarise(sd(Firmness))
```

Storage	sd(Firmness)
<fct>	<dbl>
Air	1.956272
Control	1.933132

Storage <fct>	sd(Firmness) <dbl>
ModifiedAir	3.048770

3 rows

1) The clamshell data has no strong outliers or skewness although the data does not look exactly normal (small sample size?) meeting the first validity condition. 2) The standard deviations range are similar ranging from 1.933 to 3.049. 3) The clamshells were randomly assigned to blocks meeting the final validity conditions.

References

Nathan Tintle et al.(2019). Intermediate Statistical Investigations.