Design Analysis of Experiments
"Finding the thickest noodle"

4 Components:

1. Design – The structure and components of the experiment and the purpose of testing
2. Modeling – Regression analysis of a 2^3 full Factorial design
3. Analysis – Outputs of estimated effects and coefficients with analysis of variance
4. Scientific Inference – Conclusions with various plots of visualizations

Design / Background:

People love eating pasta and I am also a huge fan of pasta since it is easy to cook and eat for a short period of time. A person like me is known as gourmand because we enjoy large amounts of food and feel stuffed. Then the design of my experiment become obvious; find out which characteristics of cooking noodle will make it the thickest in a given time so that it provides me with maximum satiety. There are several limits and problems with this initial simple proposal before testing: Choice of characteristics to test (heat temperature, amount of water and time of boiling) and under what conditions should the test be experimented (Constant heat temperature).

For the design of my experiment I chose the following factors and levels to test:

Factor	Variable	Low (-)	High (+)
Heat temperature	A	Medium (210 degrees Fahrenheit)	High (300 degrees Fahrenheit)
Amount of water	B	8 Oz	12 Oz
Boiling time	C	8 minutes	13 minutes

Data Collection:

I experimented with the noodle called “Fettuccine” since it is the thickest noodle I could get from the market. A total of 8 runs were observed with the various characteristics described in the above diagram. In order to maximize control of the experiment and decrease any potential variability, all 8 runs were experimented in my kitchen. The bottom right corner of the range was always used to control potential variable. My largest pan is chosen to boil noodle by laying it down into the pan so that I could avoid partial boiling in the beginning of the experiment till it gets soft enough to bend over into the water. In order to have heat consistency, I waited water to boil for each run and then put noodle into the pan because putting noodle form the beginning of heating would incur more complexity and noodle may not be edible. In addition, I used stopwatch to increase accuracy of the time. Lastly, I used both measuring cup and digital caliper for my water measurement and thickness of the noodle. Digital caliper was especially useful in this experiment that it is hard to measure minute thickness of noodle with human’s naked eye. Due to the development of technology, my experiment was able to increase accuracy.

Each run was replicated 5 times and the mean of each replicate was computed and used in the design of experiment analysis in R. Each run was randomly experimented, and I just wrote down the levels of each factor and its result of each observation in the piece of paper and then rearranged the numbers one through eight. The resulting order of the numbers 1 through 8 were then recorded and printed out below including each of the 5 trials.

Planning Matrix:

Planning Matrix			
Run	Heat	Water Amount	Time
1	Medium	8 Oz	8 minutes
2	Medium	8 Oz	13 minutes
3	Medium	12 Oz	8 minutes
4	Medium	12 Oz	13 minutes
5	High	8 Oz	8 minutes
6	High	8 Oz	13 minutes

7	High	12 Oz	8 minutes
8	High	12 Oz	13 minutes

Design Matrix:

Factors			Replicates (in mm)					Average (y bar)
Heat (A)	Water (B)	Time (C)	1	2	3	4	5	
-	-	-	5.73	5.61	5.69	5.47	5.55	5.610
-	-	+	6.19	6.38	6.07	6.10	6.15	6.178
-	+	-	5.30	5.45	5.58	5.53	5.52	5.476
-	+	+	6.38	6.07	6.19	6.23	6.25	6.224
+	-	-	5.86	5.82	5.83	5.88	5.88	5.846
+	-	+	6.59	6.50	6.50	6.53	6.66	6.556
+	+	-	5.76	5.82	5.83	5.94	5.86	5.842
+	+	+	6.38	6.60	6.26	6.27	6.52	6.406

Modeling:

The generic regression model to fit the data above is:

$$Y = \mu + \frac{A}{2}x_1 + \frac{B}{2}x_2 + \frac{C}{2}x_3 + \frac{AB}{2}x_1x_2 + \frac{AC}{2}x_1x_3 + \frac{BC}{2}x_2x_3 + \frac{ABC}{2}x_1x_2x_3 + \epsilon$$

In this model, A, B, and C correspond to the three factors and each factor represents factor variables of x_1, x_2, and x_3 respectively. The x_i (i = 1,2,3) variables = -1 if the factor is at its low level or x=1 if at its high level. The ϵ, known as the error term, is independently and identically distributed with mean 0 and constant variance. Thus, ϵ = y – yhat, where yhat stands for the fitted value.

Based on the Estimated Effects and Coefficients for Data, best fitting model for the data is given below. At first, I used p-value justification to simplify the model as much as possible. Two of the three main effects turn out to be significant and one 3 factor interaction is also significant in this model. This phenomenon demonstrates the Effect Hierarchy Principle that low order effects are more important than higher orders. Furthermore, the below model supports the idea of Effect Sparsity Principle that 3 of the 7 effects are significant:

$$Y = 5.6285 + \frac{.2905}{2} x_1 + \frac{.6475}{2} x_3 + \frac{-.0815}{2} x_1x_2x_3 + \epsilon$$

After doing the Lenth's Method of effect significance testing, the model for the 2^3 Factorial Design becomes:

$$Y = 5.6285 + \frac{.2905}{2} x_1 + \frac{.6475}{2} x_3$$

From this model we are able to find y the largest by having the following results:

x_1: + High Temperature / x_3: + 13 Minutes

With this result, our value for y is 6.097

Analysis:

Factorial Fit: Data versus A, B, C

Estimated Effects and Coefficients for Data

```
Call:
lm(formula = y ~ A + B + C + AB + AC + BC + ABC, data = design_lm)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1760 -0.0560 -0.0090  0.0465  0.2020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.62850    0.04544 123.868  < 2e-16 ***
A+           0.29050    0.03213   9.041 2.51e-10 ***
B+          -0.06050    0.03213  -1.883   0.0688 .
C+           0.64750    0.03213  20.152  < 2e-16 ***
AB+         -0.01650    0.03213  -0.514   0.6111
AC+         -0.01050    0.03213  -0.327   0.7460
BC+          0.00850    0.03213   0.265   0.7931
ABC+        -0.08150    0.03213  -2.537   0.0163 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1016 on 32 degrees of freedom
Multiple R-squared:  0.9397,    Adjusted R-squared:  0.9265
F-statistic: 71.18 on 7 and 32 DF,  p-value: < 2.2e-16
```

➔ There are 3 significant factorial effects: A, B and ABC. We need to do Lenth's Method to check whether each effect satisfies for this model.

```
> summary(aov.out)
            Df Sum Sq Mean Sq F value   Pr(>F)
A            1  0.844   0.844  81.744 2.51e-10 ***
B            1  0.037   0.037   3.545   0.0688 .
C            1  4.193   4.193 406.108  < 2e-16 ***
AB           1  0.003   0.003   0.264   0.6111
AC           1  0.001   0.001   0.107   0.7460
BC           1  0.001   0.001   0.070   0.7931
ABC          1  0.066   0.066   6.434   0.0163 *
Residuals   32  0.330   0.010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Scientific Inference:

The motivation for the project is to find out the optimum condition for cooking Fettuccine so as to achieve the maximum thickness of the noodle. The optimal thickness of noodle varies by people, so it is worthwhile to know the difference in thickness from the factors listed above. After doing the research, I realized that there exists a marginal thickness of noodle that it does not go thick anymore as it reaches its maximum thickness and the noodle cuts into pieces since it absorbed too much amount of water.

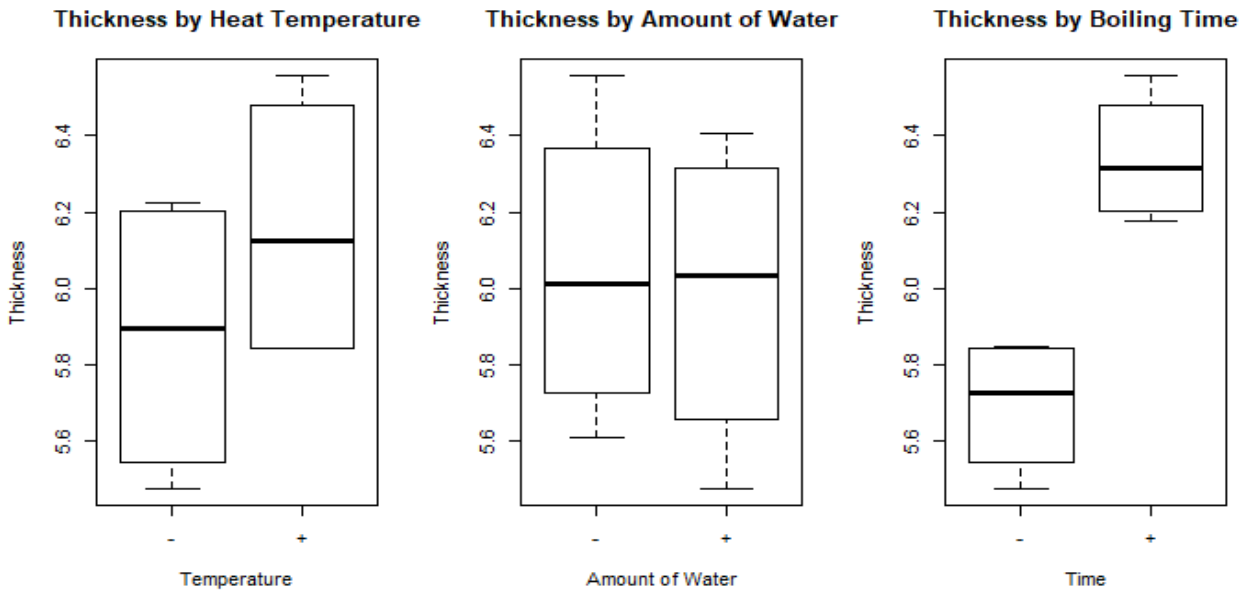

Figure 1

Figure 1 displays the Main Effects Plot explaining about the impact of three individual factor under the condition of two different level. By looking at the plots, we can easily understand which effects are significant or not. If two boxplots of single factor have substantial difference from each other, then it is considered to be significant because this means that changing the level of the factor will affect the thickness considerably. Boxplots of first and third factors are great examples of showing significant difference while the 2nd plot has slight difference between different two levels. Visualizations are great tools for understanding basic structure of the model for those who look at my experiment.

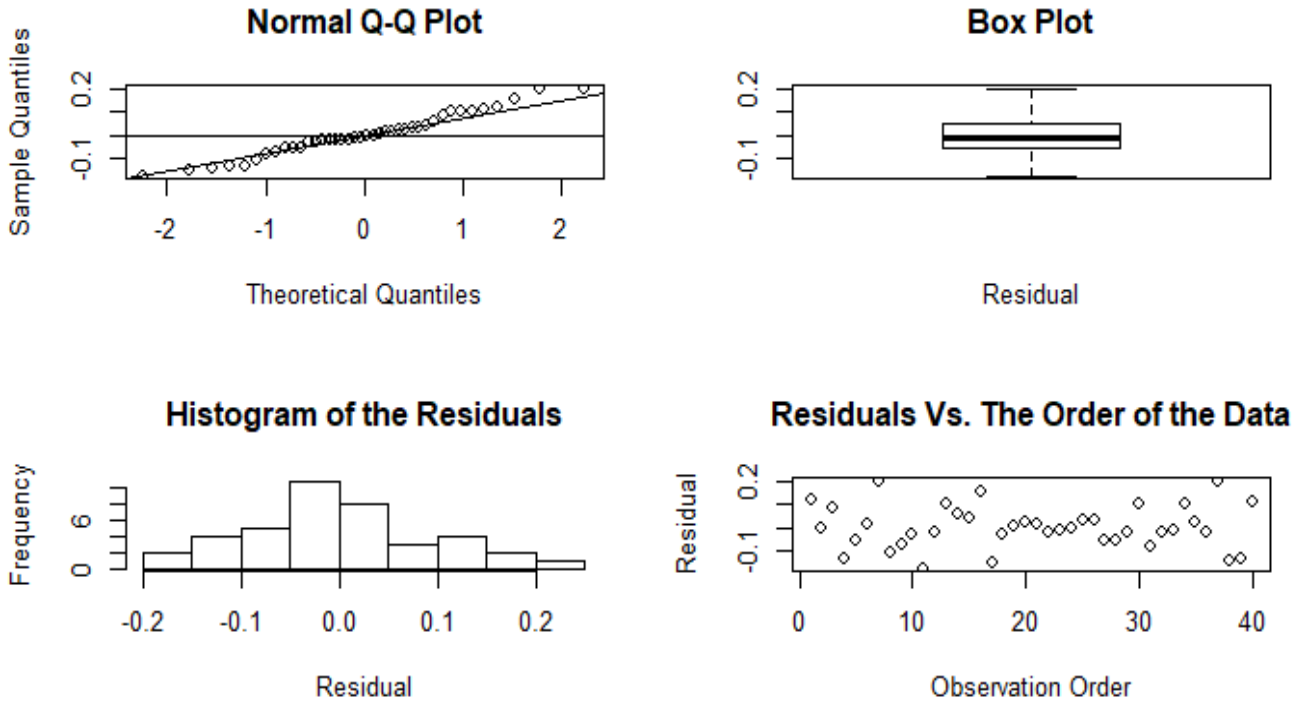

Figure 2

Four different types of plots are illustrated in the Figure 2. All of them are explaining about the distribution of the residuals. First plot verifies the normality and it is evenly distributed around 0 except for some outliers from both tails. Boxplot is another visualization showing the degree of distribution and we can conclude that most of them are close to 0. The histogram is slightly off from the normal curve, but it is still acceptable

from the results of analysis. The last plot is fairly scattered from zero and it supports the idea of stationarity that every observation is independent and identically distributed without any impact on one another.

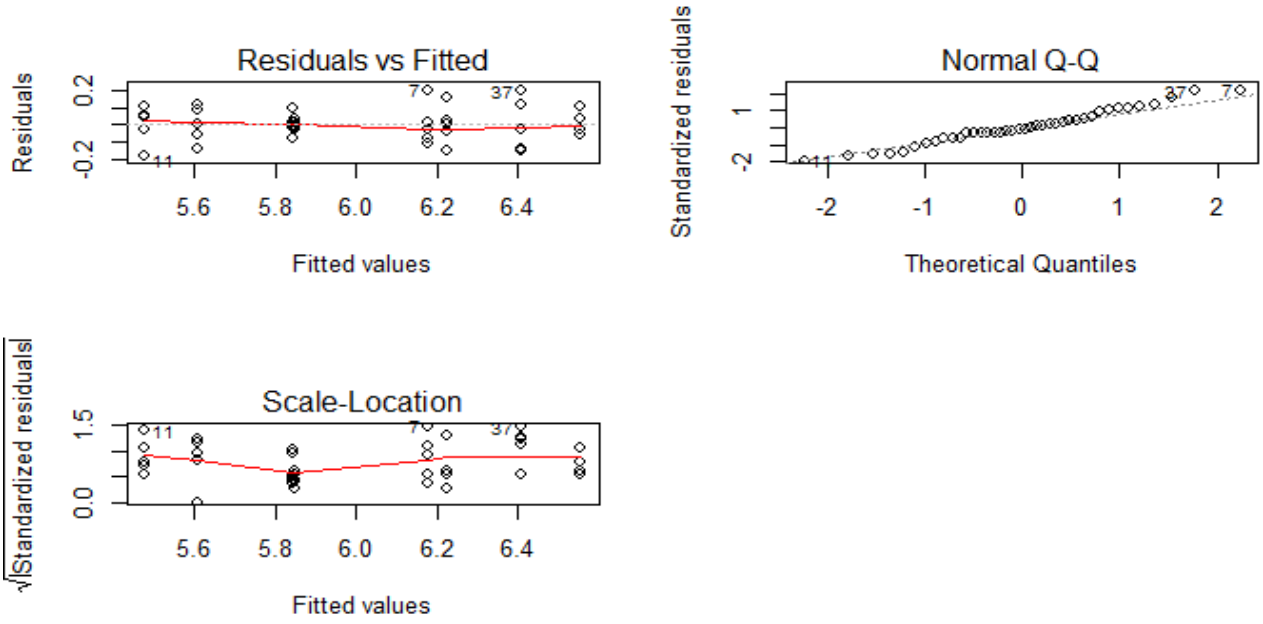

Figure 3

Figure 3 illustrates deeper analysis of residuals from previous diagrams. A plot of residuals versus fitted (predicted) values shows that the residuals are the part of the dependent variable that the model that the model could not explain, and they are our best available estimate of the error term from the regression model (Berkeley, 2011). There exist some outliers in this plot with the number labeled but the plot is reasonable enough to say that residuals are following linearity. Second plot is another quantile-quantile plot of the errors which satisfies to have a normal distribution. The last plot is similar to the first plot in Figure 3, but it has square root of the standardized residuals. Again, the red fitted line well explains the pattern of residuals from each single factorial effect.

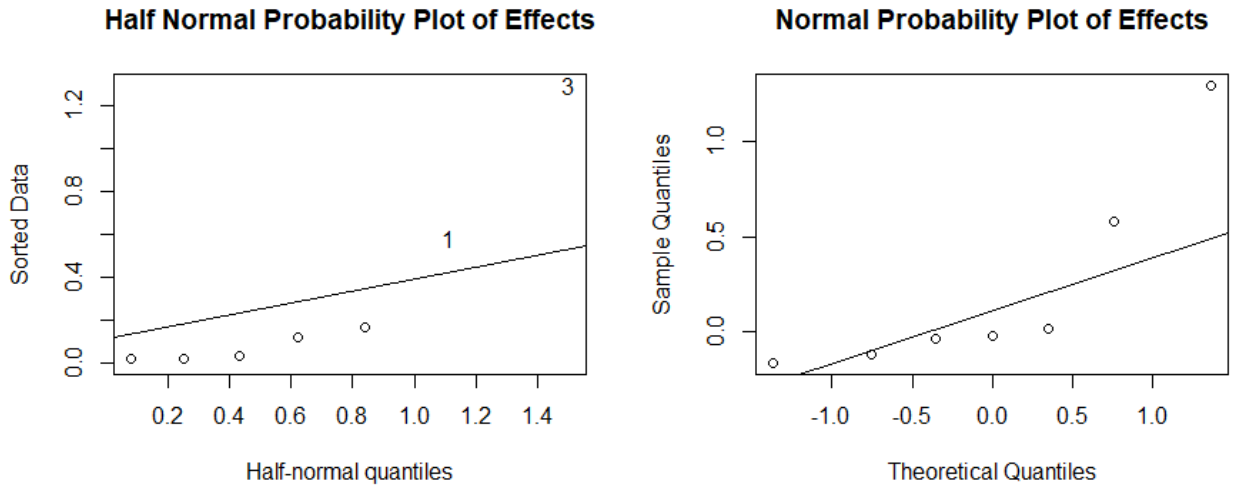

Figure 4

Both plots in Figure 4 are important graphical tools for identifying which factorial effects have significance. Both plots are very similar to each other that they have 2 outliers (Variable A and C) in the upper right corner with positive upward slope. However, a half normal distribution is the distribution of the absolute(X) that concentrates showing the magnitude of the effects while the normal probability contains rough information about direction.

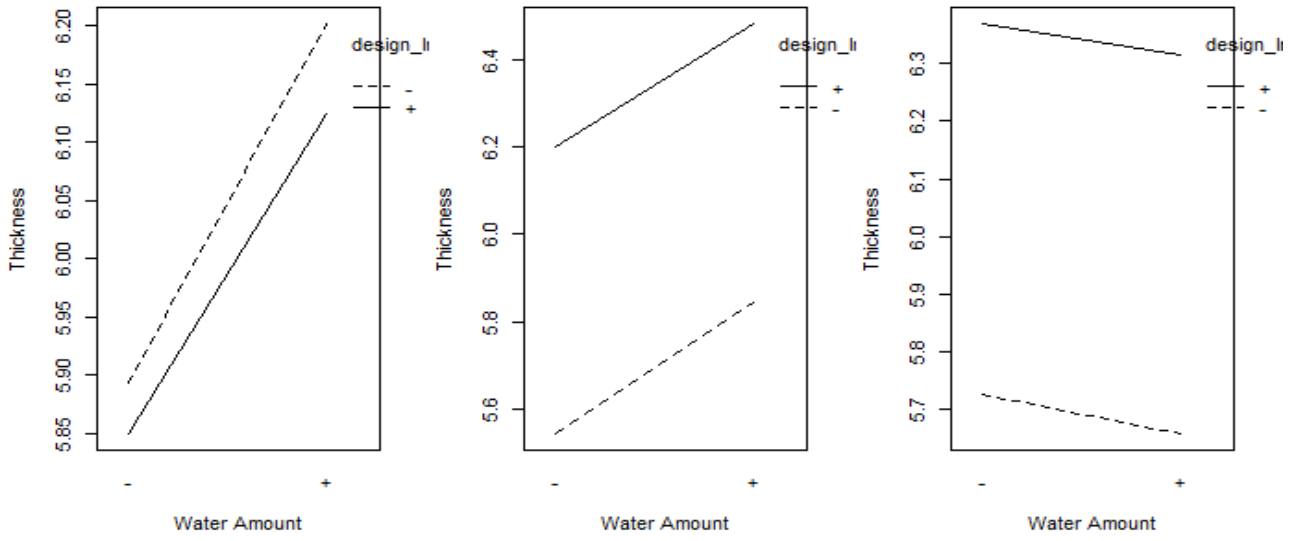

Figure 5

Figure 5 contains interesting information between the interaction of 2 factors. All the three graphs are synergistic meaning that there is a little relationship between the two factors. Since two slopes of each graph are almost parallel to each other, we could not find any significance. This graph well substantiates the analysis of effects above.

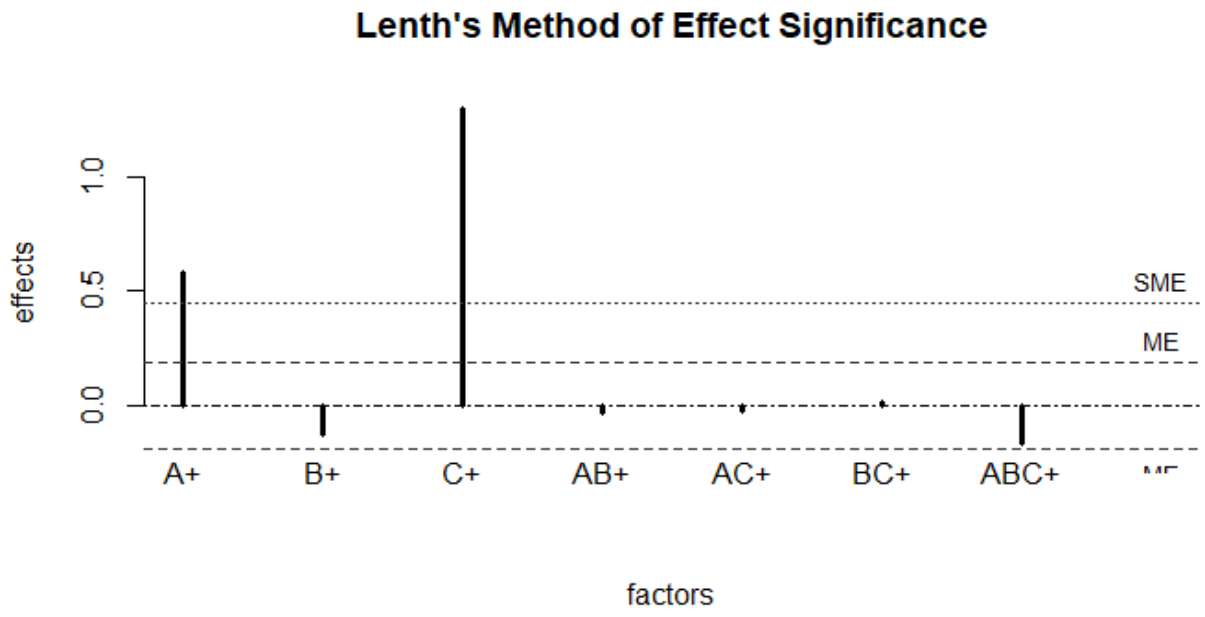

Figure 6

Figure 6 is the last plot identifying whether those significant effects are still acceptable after doing Lenth's Method testing. In the analysis, we found 3 significant effects (A, C, and ABC) that are used to find out the maximum thickness, however, ABC turns out to be inactive from the test and it is unnecessary to chosen for optimization. Since the height of ABC did not even exceed the Mean Squared threshold, we conclude it be insignificant eventually.

Unfortunately, there exist some limitations on my experiment due to long period of cooking and factors that are hard for me to control. When it comes to replicates, I chose

five samples from each run as five replicates because of time limit. In other words, every single replicate is technically the variance of rice in one replicate. To control this limitation, I assumed each noodle is independent from one another because we are not focusing the variance of each noodle but rather dealing with mean value of noodle from several factorial effects. In addition, there are hundreds of pasta manufacturing companies in the world that the outcome of the experiment from each noodle could slightly differ from one another since the contents of noodle are different.

References

Barrios, Ernesto. 2013, https://cran.r-project.org/web/packages/BsMD/BsMD.pdf

Berkeley, 2011. https://www.stat.berkeley.edu/classes/s133/Lr0.html

Faraway, Julian. 2016, https://rdrr.io/cran/faraway/man/halfnorm.html

NIST/SEMATECH, 2012, http://www.itl.nist.gov/div898/handbook/pri/section4/pri471.htm

https://www.itl.nist.gov/div898/handbook/pri/section4/pri471,r