

Tel: (814) 863-0281

Fax: (814) 863-7114

The Pennsylvania State University

323 Thomas Building

University Park, Pa 16802

Statistical Consulting Center

Department of Statistics

Eberly College of Science

**SCCID**: N/A

**Client**: Awesome bakery company, Address: PA, USA

**Consultant**: Daljeet Singh

**Date**: 4/10/2018

**Project Title**: Effects of amount of sugar and freshness on the taste preference of customers.

**Project Description:**

We have recognized that we are dealing with a 2-factor split-plot completely randomized design (CRD).

**Variables of Interest**

* Amount of sugar – 3 levels: half, double, or regular
* Amount of freshness – 3 levels: 1 day, 2 days, and 3 days
* Taste measure that is rated on a 1 to 10 scale with 1 = inedible and 10 = best
* Batch - Nine batches of cookies that were made, following the bakery recipe, except the amount of sugar

**Study Diagram**

**Factors:**

1st Factor: amount of sugar – Fixed effect, 3 levels: half, double, or regular.

2nd Factor: amount of freshness – Fixed effect, 3 levels: 1 day, 2 days, and 3 days.

Response: Taste measure that is rated on a 1 to 10 scale with 1 = inedible and 10 = best.

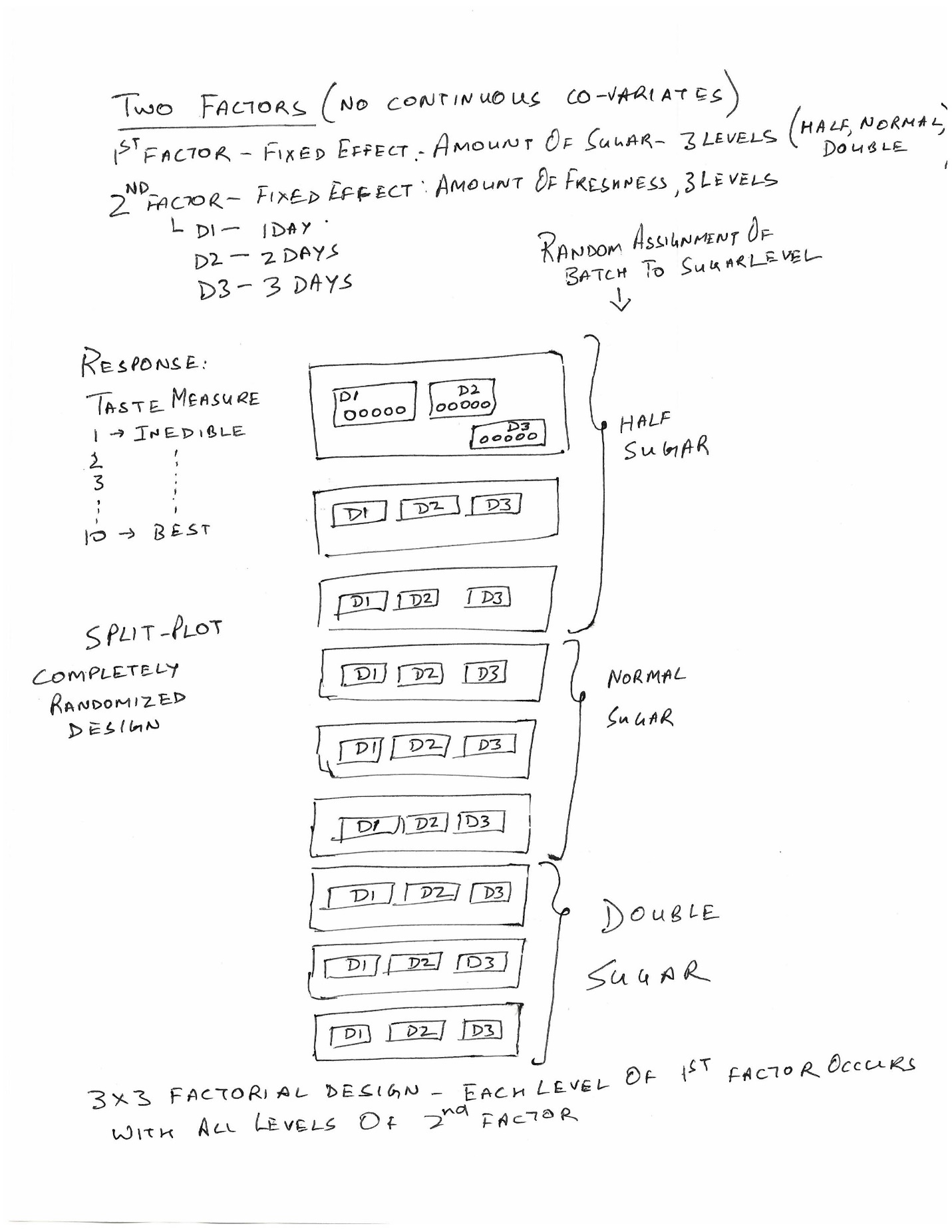


Figure 1 - Study Diagram

**Research and Statistical Questions**

* Request for help in analyzing the data.
* Analyze the main effects of the factors: amount of sugar and amount of freshness.
* Analyze the interaction between the two factors.

The null hypothesis: states that there are no differences in taste preferences (or that they are all equal) among the group means.

**H0**: μ1=μ2= μ3=μ4

**Ha**: One of the means is different

Let’s look at the null hypothesis in a little bit of context. In the given scenario we have the following:

* Between subjects factor – here the amount of sugar is the between subjects factor.
* Within subjects factor – we have the amount of freshness as the within-subjects factor

The study we are undertaking will have 3 null hypotheses detailed below:

* Means of the levels of the within-subjects factors are equal. It implies that the taste preferences will be the same for the different days of freshness.
* Means of the levels of the between-subjects factors are equal. It implies that the taste preferences will be the same for the different amounts of sugar.
* No interaction effect between amount of sugar and amount of freshness.

**EDA**

The structure of the data indicates that we have the following type of research:

* The study is a split plot in CRD. It is a 3X3 factorial design because each level of treatments of first factor occurs with all levels of 2nd factor treatment. There are no continuous covariates.
* In a completely randomized design, treatment levels or combinations are assigned to experimental units at random.
* We have a split plot study here with the main plot treatment being the sugar levels.
* The main plot treatment (sugar level) is applied to the batch of cookies that is the main plot.
* The subplots treatment is the amount of freshness.
* We will consider cookies to be homogenous since all of them are baked with a similar process. Their variation is assumed to be negligible and contained within the random error.
* Both the factors are considered to be fixed effect.
* The levels of the different variables are

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **Taste** | 9 | 2 3 4 5 6 7 8 9 10 |
| **Day** | 3 | 1 2 3 |
| **Sugar** | 3 | double half regular |
| **Batch** | 9 | 1 2 3 4 5 6 7 8 9 |

## Box plots of Sugar vs Taste

When plotting the box plot of taste grouped by sugar level in Minitab, we get:

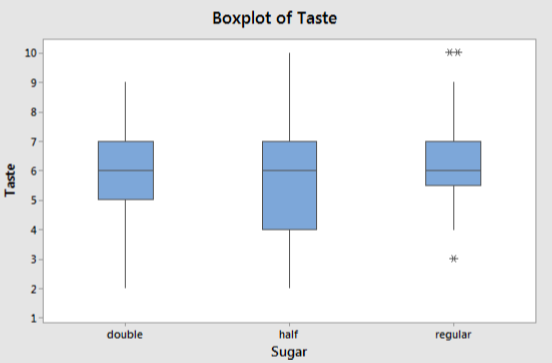


Figure 2 - Boxplot of Sugar vs Taste

This boxplot reveals to us that the means are fairly consistent at all three levels of sugar data. There does not seem to be any outliers or oddities about the plot.

## Box plots of Freshness (days) vs Taste

When plotting the box plot of taste grouped by sugar level in Minitab, we get:

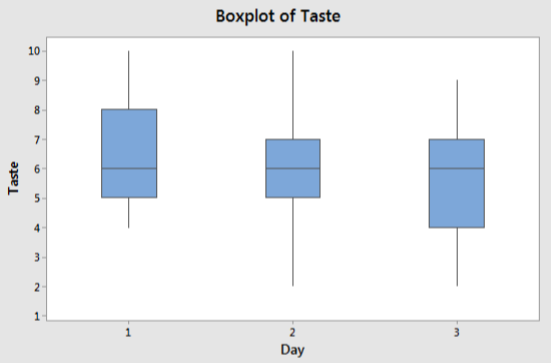


Figure 3 - Boxplot of Freshness vs Taste

This boxplot reveals to us that the means are fairly consistent at all three days of freshness data. There does not seem to be any outliers or oddities about the plot.

## Interaction Plot

It is meaningful to look at the interaction plot to analyze the interaction between the level of sugar and freshness factors. Using minitab we plot the interaction and get the following output:

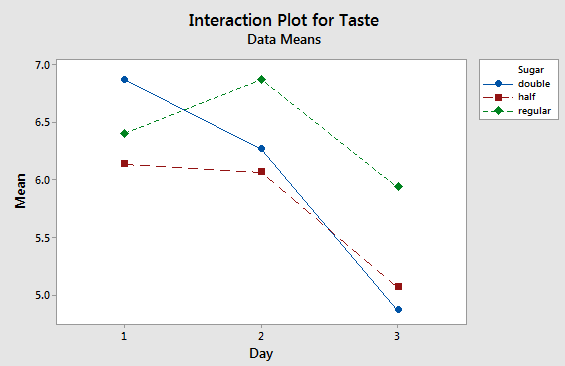


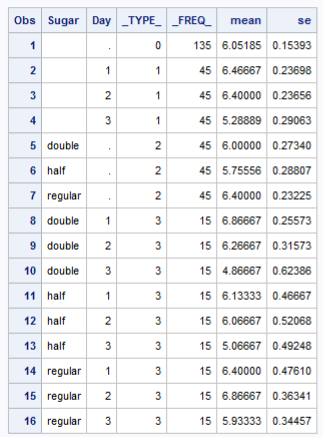
Figure 4 - Interaction of the factors

The interaction while is expected looks suspicious. The double and half sugars are going almost parallel. We will see what results we get from the model but the interaction is suspect at this point.

## Means: Grand and Treatment specific

Let’s look at the various means of the scores grouped by treatment:

Table 1 - Treatment and grand means



The following code was used to generate the above table:

proc summary data=baking;

class Sugar Day;

var Taste;

output out=output1 mean=mean stderr=se;

run;

proc print data=output1;

title 'Mean output';

run;

# Statistical Analysis

As a first step lets try an ANOVA with a simple split-plot model where we will not include the interaction between Day\*Batch(Sugar)

The model code is:

proc mixed data=baking method=type3 plots=all;

class AvgTaste Day Sugar $ Batch;

model AvgTaste=Sugar Day Sugar\*Day;

random Batch(Sugar);

store abc123; /\*Stores results for the next procedure (abc123 is name I give)\*/

title 'ANOVA of baking Data';

run;

We get the following results:

Table 2 - Model for both factors and their interaction

| **Type 3 Analysis of Variance** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **Expected Mean Square** | **Error Term** | **Error DF** | **F Value** | **Pr > F** |
| **Sugar** | 2 | 9.525926 | 4.762963 | Var(Residual) + 15 Var(Batch(Sugar)) + Q(Sugar,Day\*Sugar) | MS(Batch(Sugar)) | 6 | 1.32 | 0.3348 |
| **Day** | 2 | 39.392593 | 19.696296 | Var(Residual) + Q(Day,Day\*Sugar) | MS(Residual) | 120 | 6.78 | 0.0016 |
| **Day\*Sugar** | 4 | 9.451852 | 2.362963 | Var(Residual) + Q(Day\*Sugar) | MS(Residual) | 120 | 0.81 | 0.5190 |
| **Batch(Sugar)** | 6 | 21.644444 | 3.607407 | Var(Residual) + 15 Var(Batch(Sugar)) | MS(Residual) | 120 | 1.24 | 0.2899 |
| **Residual** | 120 | 348.622222 | 2.905185 | Var(Residual) | . | . | . | . |

We notice here that the residual has very high degrees of freedom. The reason is that there are replicate values (scores from 5 cookies for each sugar and day combination), which can be used to estimate the residual at the subplot level. Had there been no replicates, day and batch interaction are assumed to be insignificant and used to estimate the residual. Here we will therefore include the interaction Day\*Batch(Sugar) as in the following code:

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Sugar Day Sugar\*Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

Now we get the following results

Table 3 - Model for both factors and their interaction

| **Type 3 Analysis of Variance** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **Expected Mean Square** | **Error Term** | **Error DF** | **F Value** | **Pr > F** |
| **Sugar** | 2 | 9.525926 | 4.762963 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + 15 Var(Batch(Sugar)) + Q(Sugar,Day\*Sugar) | MS(Batch(Sugar)) | 6 | 1.32 | 0.3348 |
| **Day** | 2 | 39.392593 | 19.696296 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + Q(Day,Day\*Sugar) | MS(Day\*Batch(Sugar)) | 12 | 7.93 | 0.0064 |
| **Day\*Sugar** | 4 | 9.451852 | 2.362963 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + Q(Day\*Sugar) | MS(Day\*Batch(Sugar)) | 12 | 0.95 | 0.4685 |
| **Batch(Sugar)** | 6 | 21.644444 | 3.607407 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + 15 Var(Batch(Sugar)) | MS(Day\*Batch(Sugar)) | 12 | 1.45 | 0.2739 |
| **Day\*Batch(Sugar)** | 12 | 29.822222 | 2.485185 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) | MS(Residual) | 108 | 0.84 | 0.6074 |
| **Residual** | 108 | 318.800000 | 2.951852 | Var(Residual) | . | . | . | . |

The results indicate that we have significant results only for the factor Day. Since the interaction is insignificant, we proceed to remove the interaction term from the model.

## Model without the interaction term

The model code is as follows:

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Sugar Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

From this model we get the following results:

Table 4 - Model for both factors without their interaction

| **Type 3 Analysis of Variance** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **Expected Mean Square** | **Error Term** | **Error DF** | **F Value** | **Pr > F** |
| **Sugar** | 2 | 9.525926 | 4.762963 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + 15 Var(Batch(Sugar)) + Q(Sugar) | MS(Batch(Sugar)) | 6 | 1.32 | 0.3348 |
| **Day** | 2 | 39.392593 | 19.696296 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + Q(Day) | MS(Day\*Batch(Sugar)) | 16 | 8.02 | 0.0039 |
| **Batch(Sugar)** | 6 | 21.644444 | 3.607407 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + 15 Var(Batch(Sugar)) | MS(Day\*Batch(Sugar)) | 16 | 1.47 | 0.2503 |
| **Day\*Batch(Sugar)** | 16 | 39.274074 | 2.454630 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) | MS(Residual) | 108 | 0.83 | 0.6476 |
| **Residual** | 108 | 318.800000 | 2.951852 | Var(Residual) | . | . | . | . |

We find here that the sugar is not significant and therefore take the next step of removing Sugar from the model and only including the Day.

## Model with only Day factor

We get the following code:

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

With this model we get the following results:

Table 5 - Model for only freshness factor

| **Type 3 Analysis of Variance** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **Expected Mean Square** | **Error Term** | **Error DF** | **F Value** | **Pr > F** |
| **Day** | 2 | 39.392593 | 19.696296 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + Q(Day) | MS(Day\*Batch(Sugar)) | 16 | 8.02 | 0.0039 |
| **Batch(Sugar)** | 8 | 31.170370 | 3.896296 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) + 15 Var(Batch(Sugar)) | MS(Day\*Batch(Sugar)) | 16 | 1.59 | 0.2052 |
| **Day\*Batch(Sugar)** | 16 | 39.274074 | 2.454630 | Var(Residual) + 5 Var(Day\*Batch(Sugar)) | MS(Residual) | 108 | 0.83 | 0.6476 |
| **Residual** | 108 | 318.800000 | 2.951852 | Var(Residual) | . | . | . | . |

Here we see that the Day factor is significant. Let’s look at the residuals.

## Residuals

The residuals plots are:

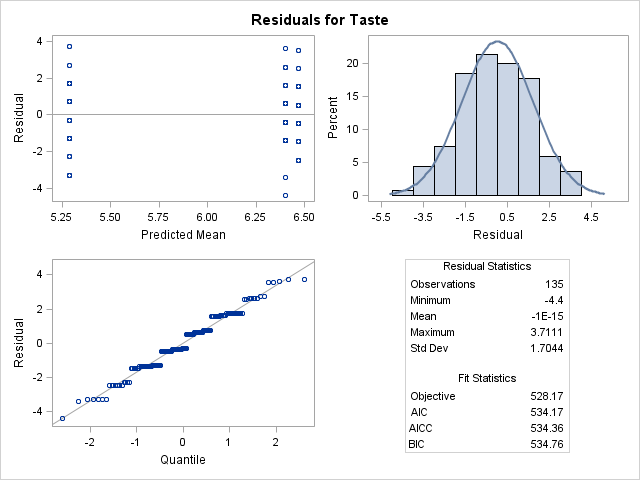


Figure 5 - Residual analysis

Here we see:

* There are two vertical lines around predicted mean=5.25 and =6.5. Our normal expectation is that there is no apparent pattern in the data.
* In the quantile vs residual plot, we see a linear trend but the step wise pattern looks suspicious.
* The bell curve does look normal.

Let’s try the conditional residuals:

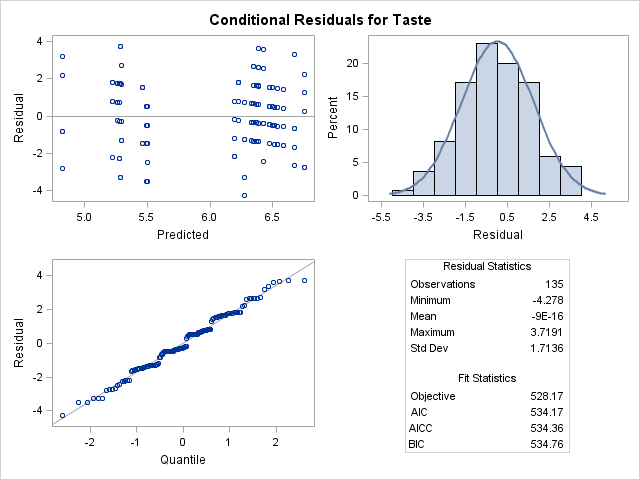


Figure 6 - Conditional residual analysis

We can see from these plots:

* The plot of the residuals mostly indicates that there’s no visible trend in the residuals, no outliers, and in general, no changing variance.
* The Q-Q plot is a normal probability plot – The plot is almost linear and the assumption of normally distributed residuals holds.

We see somewhat desirable results from the residual analysis.

## LS-Means

We see the following output from the LS-Means

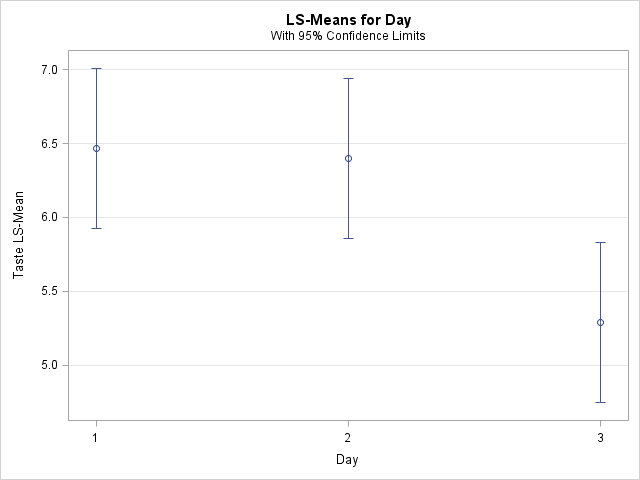


Figure 7 - LS Means output

Table 6 - LS Means table

| **Differences of Day Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer** | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Day** | **\_Day** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** | **Adj P** | **Alpha** | **Lower** | **Upper** | **Adj Lower** | **Adj Upper** |
| **1** | **2** | 0.06667 | 0.3303 | 16 | 0.20 | 0.8426 | 0.9778 | 0.05 | -0.6335 | 0.7669 | -0.7856 | 0.9189 |
| **1** | **3** | 1.1778 | 0.3303 | 16 | 3.57 | 0.0026 | 0.0069 | 0.05 | 0.4776 | 1.8780 | 0.3255 | 2.0300 |
| **2** | **3** | 1.1111 | 0.3303 | 16 | 3.36 | 0.0039 | 0.0104 | 0.05 | 0.4109 | 1.8113 | 0.2588 | 1.9634 |

Table 7 - LS Means grouping

| **Tukey-Kramer Grouping for Day Least Squares Means (Alpha=0.05)** | | |
| --- | --- | --- |
| **LS-means with the same letter are not significantly different.** | | |
| **Day** | **Estimate** |  |
| **1** | 6.4667 | A |
|  |  | A |
| **2** | 6.4000 | A |
|  |  |  |
| **3** | 5.2889 | B |

**Recommendations**: We have seen that this case study required a split plot analysis. We also saw that the interaction between the day and sugar are not significant. The only significant factor is freshness.

**Appendix A – SAS Code**

data baking;

input Taste Day Sugar $ Batch;

datalines;

7 1 double 1

8 1 double 1

6 1 double 1

6 1 double 1

7 1 double 1

5 2 double 1

8 2 double 1

5 2 double 1

7 2 double 1

6 2 double 1

2 3 double 1

9 3 double 1

2 3 double 1

7 3 double 1

9 3 double 1

4 1 half 2

4 1 half 2

7 1 half 2

5 1 half 2

6 1 half 2

3 2 half 2

2 2 half 2

5 2 half 2

7 2 half 2

5 2 half 2

8 3 half 2

4 3 half 2

7 3 half 2

4 3 half 2

2 3 half 2

5 1 regular 3

6 1 regular 3

4 1 regular 3

6 1 regular 3

10 1 regular 3

7 2 regular 3

8 2 regular 3

6 2 regular 3

7 2 regular 3

6 2 regular 3

6 3 regular 3

6 3 regular 3

7 3 regular 3

5 3 regular 3

7 3 regular 3

8 1 double 4

8 1 double 4

7 1 double 4

6 1 double 4

8 1 double 4

7 2 double 4

6 2 double 4

9 2 double 4

5 2 double 4

5 2 double 4

6 3 double 4

5 3 double 4

2 3 double 4

4 3 double 4

3 3 double 4

6 1 half 5

10 1 half 5

4 1 half 5

9 1 half 5

6 1 half 5

9 2 half 5

5 2 half 5

8 2 half 5

6 2 half 5

7 2 half 5

5 3 half 5

6 3 half 5

6 3 half 5

4 3 half 5

2 3 half 5

7 1 regular 6

8 1 regular 6

4 1 regular 6

4 1 regular 6

9 1 regular 6

9 2 regular 6

8 2 regular 6

5 2 regular 6

10 2 regular 6

7 2 regular 6

4 3 regular 6

7 3 regular 6

7 3 regular 6

4 3 regular 6

7 3 regular 6

6 1 double 7

8 1 double 7

5 1 double 7

6 1 double 7

7 1 double 7

6 2 double 7

7 2 double 7

5 2 double 7

7 2 double 7

6 2 double 7

3 3 double 7

7 3 double 7

6 3 double 7

5 3 double 7

3 3 double 7

5 1 half 8

6 1 half 8

7 1 half 8

8 1 half 8

5 1 half 8

5 2 half 8

6 2 half 8

6 2 half 8

9 2 half 8

8 2 half 8

4 3 half 8

5 3 half 8

4 3 half 8

7 3 half 8

8 3 half 8

7 1 regular 9

6 1 regular 9

7 1 regular 9

5 1 regular 9

8 1 regular 9

6 2 regular 9

7 2 regular 9

6 2 regular 9

6 2 regular 9

5 2 regular 9

6 3 regular 9

7 3 regular 9

6 3 regular 9

7 3 regular 9

3 3 regular 9

;

proc mixed data=baking method=type3 plots=all;

class AvgTaste Day Sugar $ Batch;

model AvgTaste=Sugar Day Sugar\*Day;

random Batch(Sugar);

store abc123; /\*Stores results for the next procedure (abc123 is name I give)\*/

title 'ANOVA of baking Data';

run;

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Sugar Day Sugar\*Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Sugar Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

proc mixed data=baking method=type3 plots=all;

class Taste Day Sugar Batch;

model Taste=Day;

random Batch(Sugar) Day\*Batch(Sugar);

store abc123;

title 'ANOVA of baking Data';

run;

ods html style=statistical sge=on;

proc plm restore=abc123;

lsmeans Sugar / adjust=tukey plot=meanplot cl lines;

lsmeans Day / adjust=tukey plot=meanplot cl lines;

lsmeans Sugar\*Day / adjust=tukey plot=meanplot cl lines;

/\* The lsmeans statement here prints out the model fit means, performs the Tukey

mean comparisons, and plots the data. \*/

ods exclude diffplot;

run; title; run;