Impacts of Oven Temperature on Pillsbury Funfetti Cupcake Height

By Harley Clifton, Madelaine Brown, and Kyle Rutten

I. Introduction

When baking cupcakes, it is generally preferred to have cupcakes that are as tall as possible. Taller cupcakes have a cone shape which is more ideal for frosting compared to flat cupcakes. Cupcake Height serves as a stand-in for many other measures that are more descriptive, but much harder to measure such as cupcake moisture and cupcake density. Thus, this metric is a general stand-in that we seek to optimize. While each box of cupcake mix will provide a recommended baking time and temperature, each individual oven can vary greatly in how it heats up and holds its temperature. These differences in individual oven temperatures can lead to variations in Cupcake Height. We aim to identify the ideal temperature setting on Harley's oven that will result in the tallest Pillsbury Funfetti Cupcakes; therefore, making the best cupcakes possible in a given amount of time.

II. Experimental Units and Randomization

Our experimental units were each individual batch of six cupcakes. Each experimental unit was baked at an oven temperature setting of either 300°F, 325°F, 350°F, or 375°F. The temperatures were assigned to the individual batches of cupcakes on the day of baking in order from lowest to highest temperature to avoid residual heat from a prior higher temperature setting from affecting the baking process of the subsequent batch. If we were to repeat this experiment, oven temperature would ideally be assignment to our experimental units via a completely randomized design, and we would also allow the oven to cool down to room temperature before reheating to the next temperature setting.

To minimize potential confounding variables, we used the same pan, type of cupcake liners, type of eggs (Eggland's Best Classic Eggs), vegetable oil (Crisco), amount of batter in each liner (¼ cup), and type of batter (Pillsbury Funfetti). Further, we maintained the same cooling time before measuring height (5 minutes), time in the oven (19 minutes),

oven (Harley's Oven), method of measuring cupcake height (wooden skewers through the center), oven light status (on), and elevation (1461 m in Bozeman, MT) throughout this experiment. Alternative sources of potential variation which we could not feasibly control for could have stemmed from the cupcake mix itself, environmental factors, and individual variation in measurements. Variation from the cupcake mix may be attributed to the amount and location of sprinkles in each cupcake, and the exact amount of dry ingredients in each cupcake box mix. Environmental sources of variation include humidity and ambient temperature, time that the batter sat at room temperature before baking each batch, and fluctuations in oven temperature while baking. Individual variation in measurements could potentially result from the angle of skewer insertion to measure the height of each cupcake, and the reading of measurement itself. We attempted to reduce variation from these sources by having Harley mix the cake batter each time, Madelaine measured the amount of batter put into each cupcake liner with the same ¼ measuring cup and marked the height of each cupcake on a skewer, while Kyle measured all of the skewer markings against the same ruler. To control for baking inconsistencies due to individual cupcakes' location in the pan, we decided to calculate the average height of cupcakes in each batch and use this as our response variable.

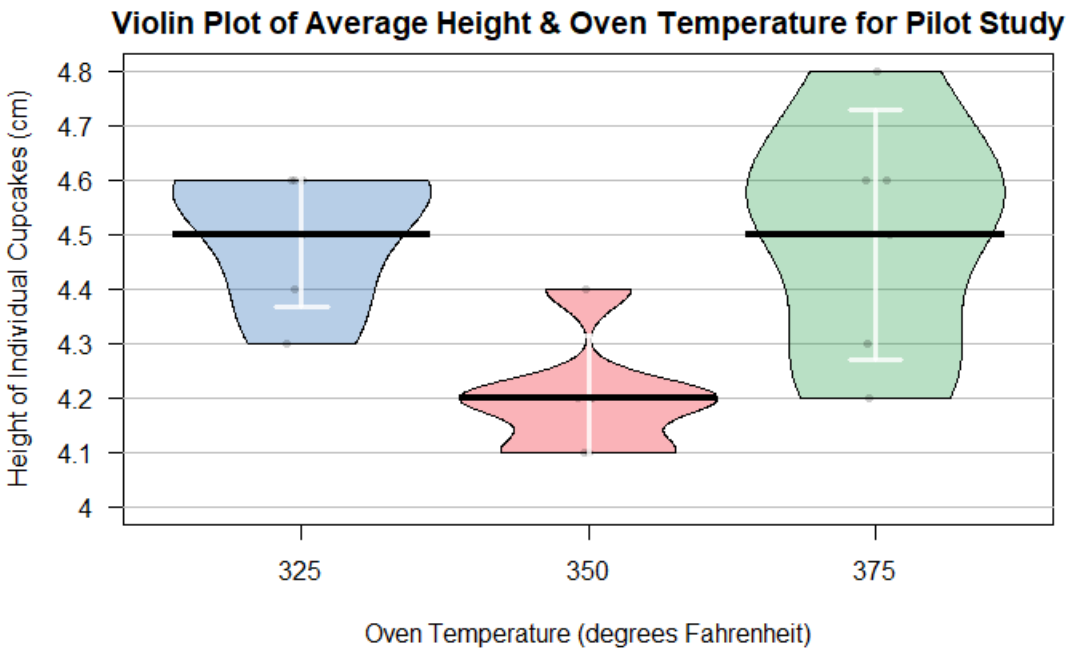
Unfortunately, we did not cover randomization until we had already run our experiment and collected the data. As a result, we did not randomize which box of cake mix got assigned to which temperature treatment, nor did we randomize the order the batches were baked in the oven. We use the first box of cake mix to obtain our pilot study measurements (first batches at 325, 350, and 375°F). For baking order in our pilot run, we decided to bake the first batch of cupcakes at the lowest oven temperature we were testing (325°F) and bake the following batches in order of increasing oven temperature (350°F and 375°F, respectively). We chose this order of treatments to prevent any residual heat from higher oven settings from causing fluctuations in actual baking temperatures. For the remaining boxes of cake mix, we followed the same baking order, baking all batches for a single treatment level before moving up to the next oven temperature setting. Since we added another oven temperature after our pilot study, our first three batches for our main study were baked at 300°F, next two at 325°F, followed by two at 350°F, and the final two

batches were baked at 375°F. Although our assignment procedure was not a completely randomized design, it allowed us to accurately test oven temperature, so it is a reasonable and valid treatment allocation nonetheless.

III. Pilot Study and Sample Size Calculations

Our pilot study consisted of baking a single batch of cupcakes (6 individual cupcakes total) at each of our three initial treatment levels: 325°F, 350°F, and 375°F. The results of which are included in **Figure 1**.

Figure 1: Violin Plots of Pilot Study Data

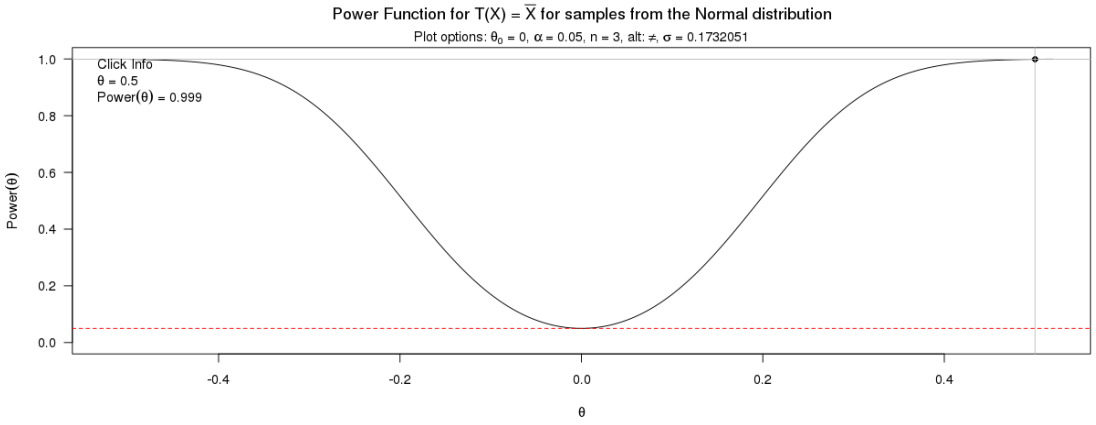


Enhanced violin plots of the average cupcake height based on oven temperature, when plotting the heights of each individual cupcake in a given batch. Using data from only the pilot study.

Based on the little variation we saw between cupcake height and oven temperature in our pilot study, and skepticism about whether Harley's oven runs hot, we decided to add a fourth treatment level of 300°F.

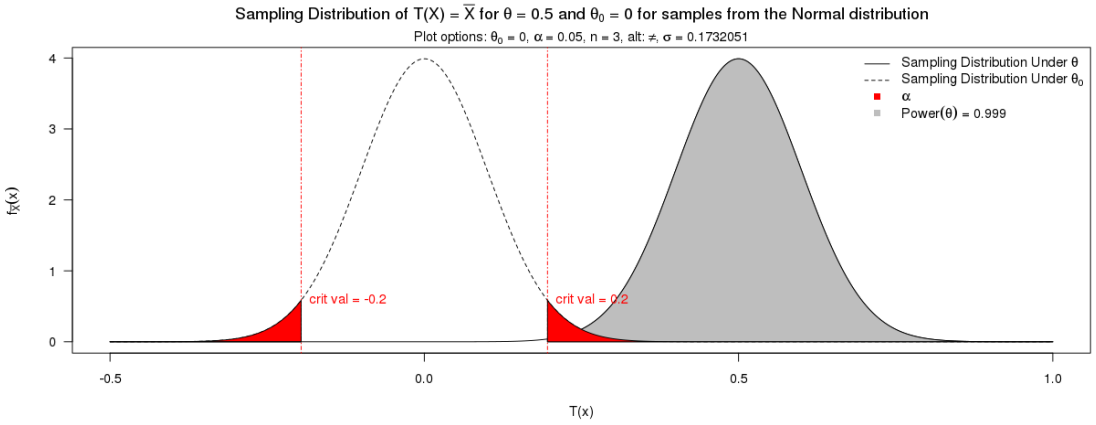
To determine the sample size needed for our full experiment, we used a Shiny app created by Christian Stratton and Jenny Green to look at theoretical power and the sampling distribution of our data as shown in **Figure 2** and **Figure 3** respectively (Stratton, 2021).

Figure 2: Power Function



This figure shows the power for for detecting differences based on various alternative hypothesis (heta) values under the specified conditions. In particular, the power for theta = 1.

Figure 3: Sampling Distribution

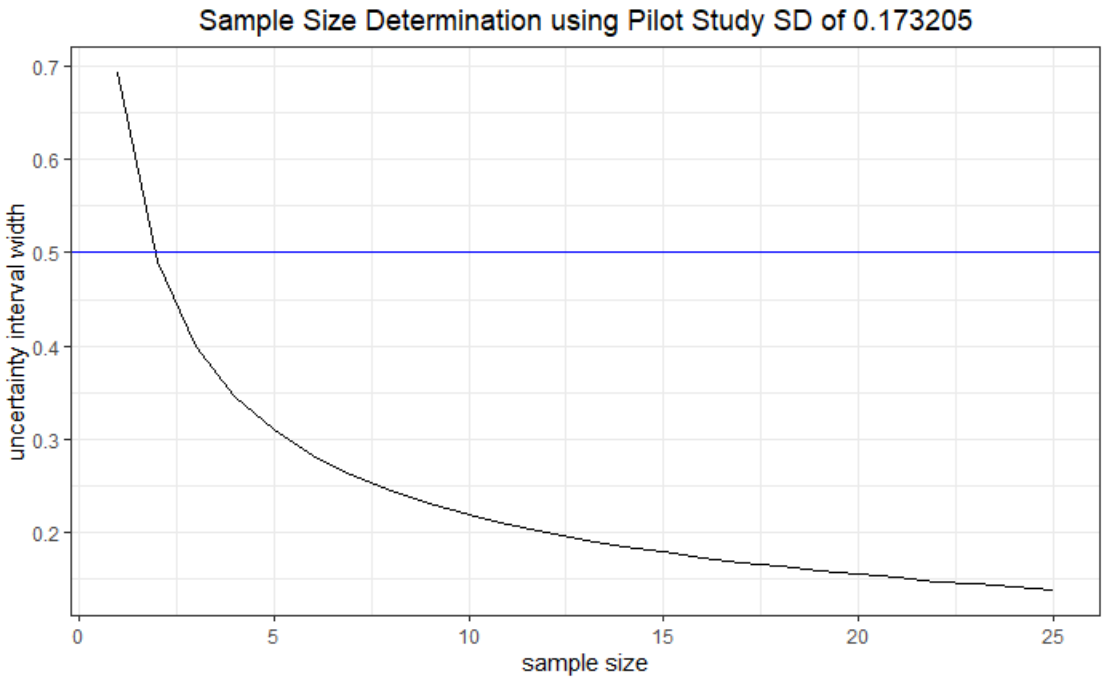


This figure shows the sampling distribution under the null and alternative hypotheses.

Using the standard deviation from our pilot study, which was 0.173205 cm, the 'ggplot2' package was used to find the number of replicates needed at each temperature setting to achieve our ideal power of 0.95. Based on the results in **Figure 4**, our minimum sample

number needed to reach this power is 2 batches of cupcakes per treatment group (Wickham, 2016).

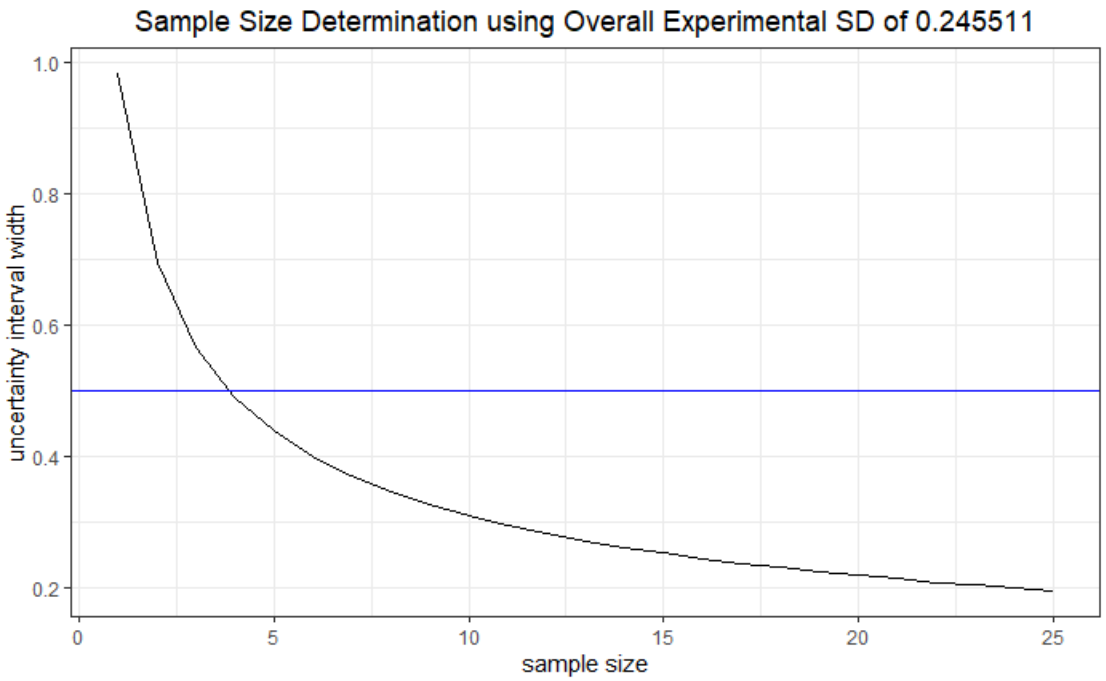
Figure 4: Sample Size Determination with Pilot Study



Sample size determination based on the standard deviation (0.173 cm) from the pilot study. You'd need at least **2** *batches at each temperature to have an estimated uncertainty interval width less than 0.5.*

After performing our experiment, we did another sample size calculation using the overall experimental standard deviation which was 0.245511 cm. The 'ggplot2' package was used to visualize the relationship between this standard deviation, power, and sample size in **Figure 5** (Wickham, 2016).

Figure 5: Retrospective Sample Size Determination with Complete Experimental Data



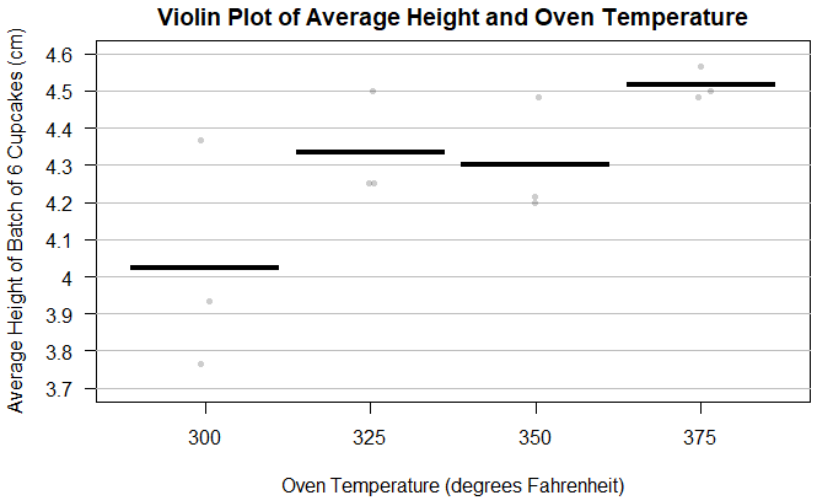
*Sample size determination, based on the standard deviation (0.246 cm) from the overall study data. You'd need at least **4** batches at each temperature level to have an estimated uncertainty interval width less than 0.5.*

In retrospect, we determined that to achieve a power of 0.95, we would have preferred to have 4 batches at each temperature level. However, time constraints prevented us from baking additional batches to achieve the ideal number of replicates at each level.

IV. Data Overview

The average cupcake heights at the four different oven temperatures were visualized in R (R Core Team, 2021) with enhanced violin plots from the 'yarrr' package as shown in **Figure 6** (Philips, 2017).

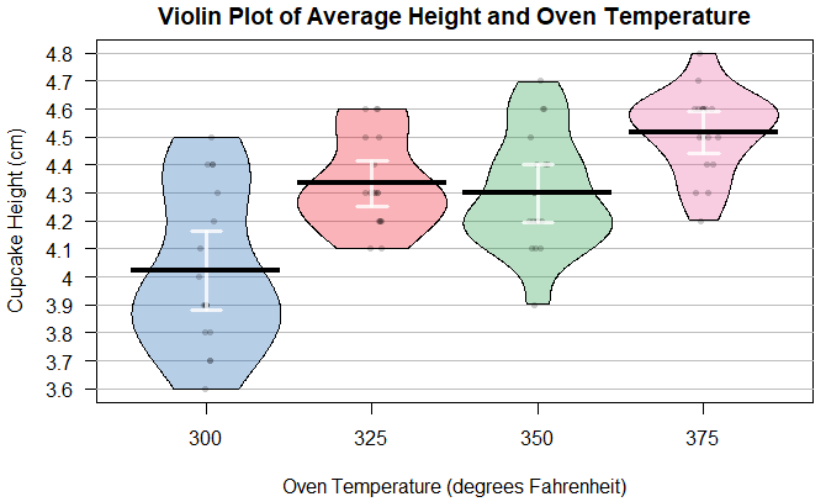
Figure 6: Violin Plot of Average Cupcake Height



Enhanced violin plots of average cupcake height based on oven temperature, when only plotting the average heights of each batch.

Since there were only three data points (average heights of each batch) at each temperature level, we created an additional violin plot with the 'yarrr' package which displays all individual cupcake height measurements instead of our batch averages in **Figure 7** (Philips, 2017).

Figure 7: Violin Plot of Individual Cupcake Heights



Enhanced violin plots of the average cupcake height based on oven temperature, when plotting the heights of each individual cupcake in a given batch.

Based on this data visualization, cupcakes baked at 375°F had the largest average height at 4.5 cm, followed by cupcakes baked at 325°F with an average height of 4.33 cm. Cupcakes baked at 350°F were the next tallest with an average height of 4.3 cm, and the shortest cupcakes, on average, were those baked at 300°F which had an average height of 4.02 cm.

V. Data Analysis and Results

In this experiment, oven temperature was our only predictor and had 4 categorical levels (300, 325, 350, and 375°F); therefore, we deemed a One-Way Analysis of Variance test appropriate for our data. First, R was used to fit a model that predicts the average height of a batch of cupcakes based on oven temperature levels, using 300 F as a baseline for comparison (R Core Team, 2021) (Model 1). The Reference Case Model (**Model 1**), is as follows:

$$\boldsymbol{\mu(AvgHeight|Temp)_i = 4.0222 + 0.3111 * I_{Temp=325} + 0.2778 * I_{Temp=350} + 0.4944 * I_{Temp=375} + \epsilon_i}$$

With indicator variables defined as follows:

- $I_{Temp=325} = 1$ is 1 when $Temperature = 325°F$ and 0 otherwise,
- $I_{Temp=350} = 1$ is 1 when $Temperature = 350°F$, and 0 otherwise, and
- $I_{Temp=375} = 1$ is 1 when $Temperature = 375°F$ and 0 otherwise.
- Additionally, $\epsilon \sim N(0, \sigma^2)$ is an error term for $i = 1, 2, \ldots, 12$ batch of cupcakes.

The same model, but in Cell Means notation (**Model 2**) is as follows:

$$\boldsymbol{\mu(AvgHeight|Temp)_i = 4.0222 * I_{Temp=300} + 4.3333 * I_{Temp=325} + 4.3000 * I_{Temp=350} + 4.5167 * I_{Temp=375} + \epsilon_i}$$

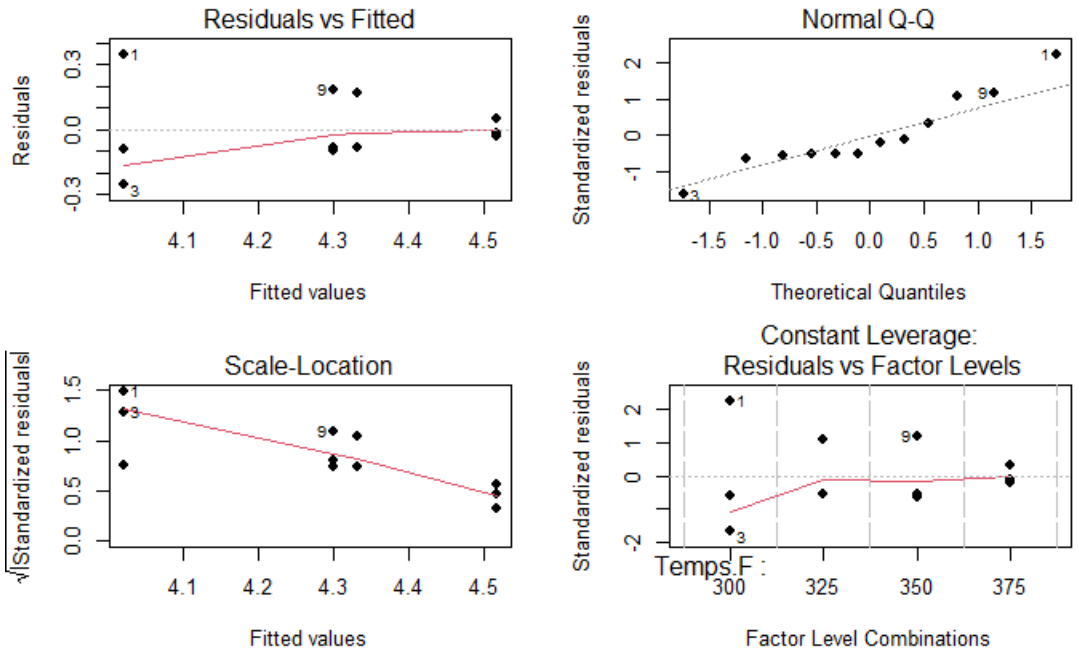
With indicator variables defined as follows:

- $I_{Temp=300} = 1$ is 1 when $Temperature = 300°F$ and 0 otherwise,
- $I_{Temp=325} = 1$ is 1 when $Temperature = 325°F$ and 0 otherwise,
- $I_{Temp=350} = 1$ is 1 when $Temperature = 350°F$, and 0 otherwise, and
- $I_{Temp=375} = 1$ is 1 when $Temperature = 375°F$ and 0 otherwise.
- Additionally, $\epsilon \sim N(0, \sigma^2)$ is an error term for $i = 1, 2, \ldots, 12$ batch of cupcakes.

This notation is useful because it explicitly states the overall average heights of the cupcakes at each temperature setting. The estimated coefficient for the intercept in the Reference Case Model represents the average batch height for cupcakes baked at 300°F in cm. The other estimated coefficients Model 1 represent the average increase in mean height of a batch of cupcakes for the other three oven temperature settings.

The multiple R-squared for Model 1 is 0.5657 (R Core Team, 2021). This means the reference case model explains 56.57% of the variation in the average Pillsbury Funfetti cupcake batch height. Although this model explains over half of the total variations in average batch height, it still leaves a lot of variation unexplained. The Standard Suite of diagnostic plots was generated for Model 1 (**Figure 8**) (R Core Team, 2021).

Figure 8: Model Diagnostics Array for Reference Case Model.



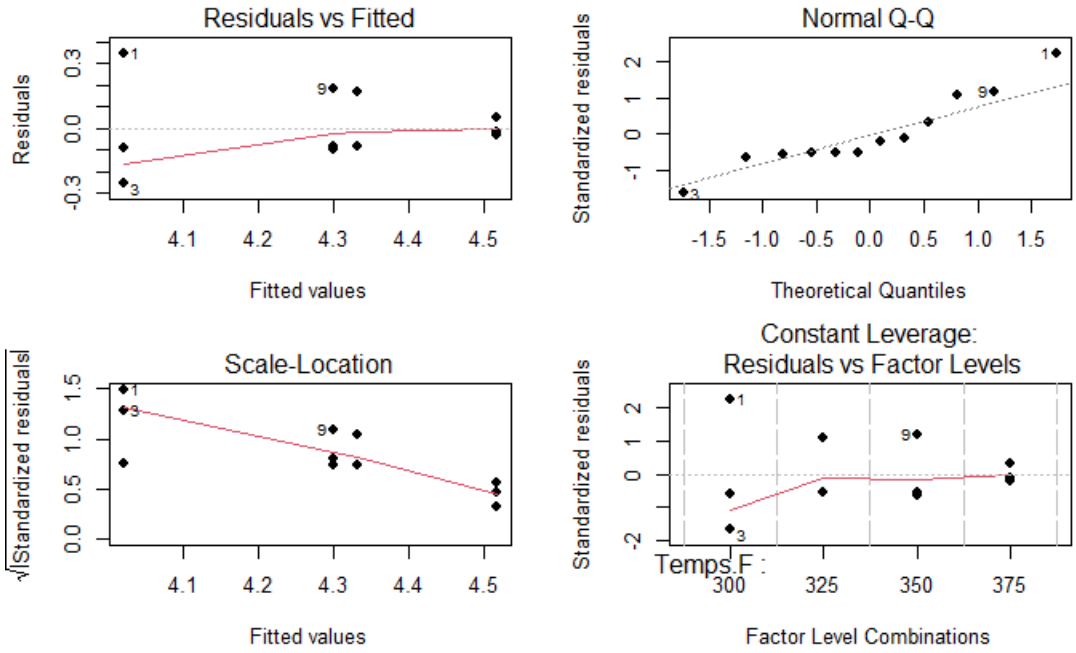
Diagnostic plots for Model 1 showing Residuals vs Fitted, Normal Q-Q, Scale-Location, and Constant Leverage: Residuals vs Factor Levels Plots.

The funneling pattern present in the Residual vs. Fitted plot indicates some issues with our constant variance assumption (**Figure 8**). Additionally, the data deviates from the 1:1 Q-Q

line in the Normal Q-Q plot suggesting a slight right skew, but not enough to indicate a large departure from normality (**Figure 8**).

The multiple R-squared for Model 2 is 0.9987 (R Core Team, 2021). Therefore, Model 2 explains 99.87% of the variation in the average Pillsbury funfetti cupcake batch height and leaves very little variation unexplained. After taking a closer look at Model 2 using standard diagnostic plots, we found a moderate funneling pattern in the Residuals vs. Leverage plot, indicating some deviation from constant variance (**Figure 9**). In the Normal Q-Q plot, there is a short right tail, indicating some right skewness, but not enough to deviate from normality (**Figure 9**).

Figure 9: Model Diagnostics Array for Cell Means Model

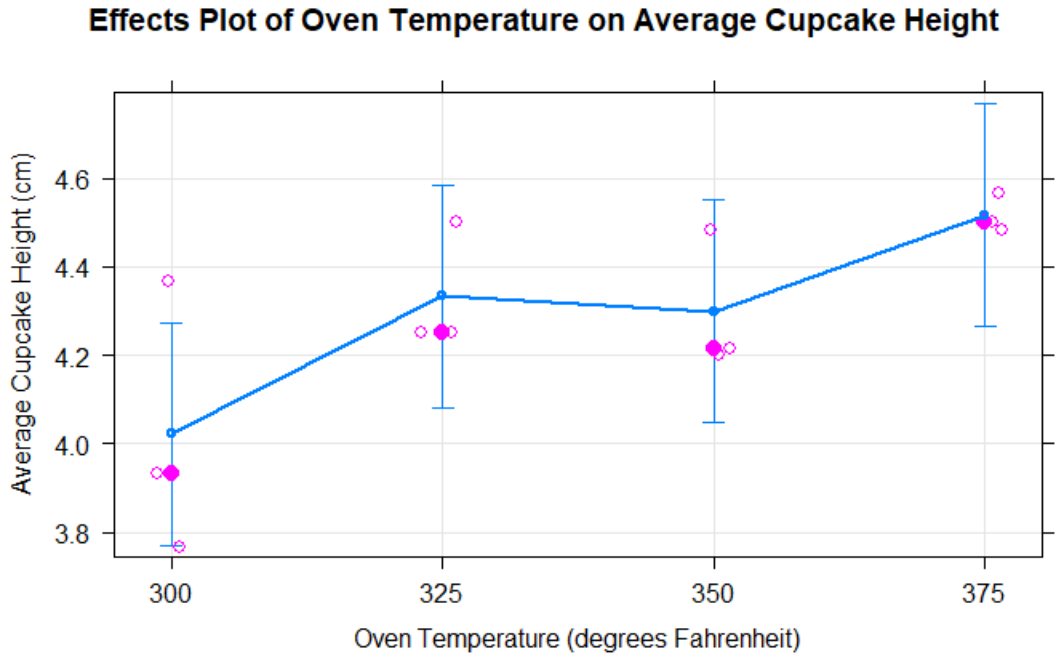


Diagnostic plots for Model 2 showing Residuals vs Fitted, Normal Q-Q, Scale-Location, and Constant Leverage: Residuals vs Factor Levels Plots.

Because Model 2 explains notably more variation than Model 1, and no noticeable change in constant variance or normality, Model 2 was chosen as our final model.

To visualize the impact of oven temperature on the average height of a batch of Pillsbury Funfetti Cupcakes, term plots were generated using the 'effects' package (Fox and Weisberg, 2019) (**Figure 10**).

Figure 10: Effects Plot



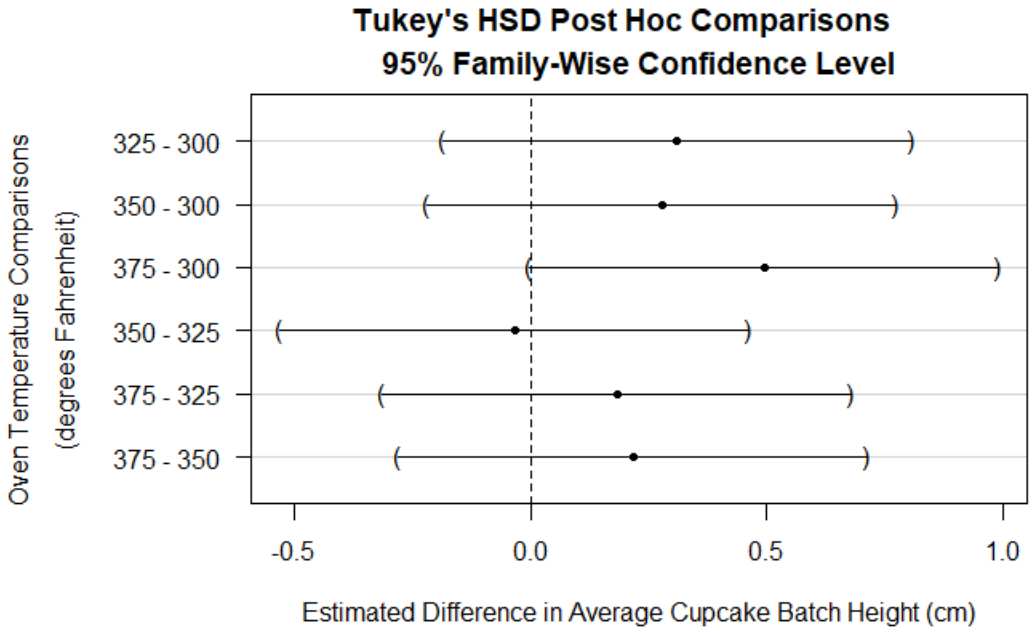
From the term plot, the most extreme differences in average batch heights occur when comparing temperatures 300°F and 375°F. According to the term plots, the mean cupcake height for treatment group 300°F was about 4.02 cm, 4.35 cm for treatment group 325°F and 4.30 cm for treatment group 350°F. Cupcakes baked at the highest temperature, 375°F, had the highest average cupcake height of about 4.50 cm. These average heights were very similar to the average heights in the enhanced violin plots in **Figure 7**.

After data visualization, a One-Way Analysis of Variance test was performed on our data (R Core Team, 2021). There was moderately strong evidence found against the null hypothesis that the mean height of a batch of Pillsbury Funfetti cupcakes was the same across all oven temperature settings ($F(4,8) = 1538.7$, p-value = 1.418×10^{-11}), suggesting that at least one temperature setting resulted in a different mean cupcake batch height (**Table 1**). Interestingly, all of the estimated coefficients for the Cell Means Model were

detected as "statistically significant". In a situation like this, it is important to keep in mind that "statistical significance" does not necessarily equate to a meaningful difference. This is exactly why we predetermined that we would consider a meaningful difference in average batch height to be 0.5 cm.

To explore the One-Way ANOVA results further, we performed Tukey's HSD post hoc comparisons to determine which temperature level(s) differed using the 'multcomp' package in R (Hothorn, 2008). The results of these comparisons are reported in **Table 2** and are displayed in **Figure 11**.

Figure 11: Tukey's HSD



After taking a closer look at the differences, it is worth noting that the largest difference in average batch height between the temperature settings occurred between 300°F and 375°F (**Table 2**). The estimated difference is 0.49444 cm. Since we previously determined a meaningful difference in average cupcake batch height would be 0.5 cm, depending on rounding choices, the difference between these two oven temperature settings could potentially be meaningful. However, **Figure 11** illustrates that we cannot come to that conclusion since the lower bound of the 95% Family Wise Confidence Interval is just below zero.

If this experiment were to be run again, the order in which the individual batches are baked in the oven should be randomized. More explicitly the 12 batches - 3 batches at each of the 4 temperature settings - should be randomly assigned to baking order in the oven. Additionally, our cupcake batter sat out at room temperature between baking each batch. If repeated, accounting for the time the batter spends sitting out could potentially help explain more of the variation in cupcake height across the treatment groups. Furthermore, the results from this experiment would provide a better estimate of the overall standard deviation to use in power calculations. Lastly, due to time constraints of our project, we made three batches of cupcakes for each treatment level; however, if we were to do this study again, we would have liked to achieve our retrospective sample size calculation.

VI. Discussion

Based on our results, there may be a meaningful difference between cupcake height when comparing baking temperatures 300°F and 375°F (4.02 cm and 4.5 cm, respectively). All other baking temperature comparisons did not result in a meaningful difference. With this information, we can conclude that baking cupcakes at 375°F will result in better cupcakes than baking at 300°F, in regards to height alone. Other factors that may contribute to cupcake satisfaction are cupcake moisture, density, and overall taste; further exploration of these qualities would require us to conduct additional experiments. When conducting our experiment, there were noticeable differences in cupcake density and texture, despite there only being marginal differences in cupcake height.

Appendix

References

C. Stratton, J. Green, and A. Hoegh. Not just normal: Exploring power with Shiny apps (2021). Technology Innovations in Statistics Education. https://shiny.stt.msu.edu/jg/powerapp/

D. Sjoberg, K. Whiting, M. Curry, J. Lavery and J. Larmarange. The gtsummary Package: Reproducible Summary Tables with the gtsummary Package (2021). The R Journal, 13(1):570-580. https://doi.org/10.32614/RJ-2021-053.

H. Wickham. The ggplot2 Package: Elegant Graphics for Data Analysis (2016). Springer-Verlag New York. https://ggplot2.tidyverse.org.

J. Fox and S. Weisberg. The effects Package: Visualizing Fit and Lack of Fit in Complex Regression Models with Predictor Effect Plots and Partial Residuals (2018). Journal of Statistical Software, 87(9):1-27. https://doi.org/10.18637/jss.v087.i09.

N. Phillips. The yarrr Package: A Companion to the e-Book "YaRrr!: The Pirate's Guide to R" (2017). R package version 0.1.5. https://CRAN.R-project.org/package=yarrr

R Core Team. R: A language and environment for statistical computing (2021). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

T. Hothorn, F. Bretz and P. Westfall. The multcomp Package: Simultaneous Inference in General Parametric Models (2008). Biometrical Journal, 50(3):346-363. http://multcomp.R-forge.R-project.org.

Model Summaries

Model 1: Reference Case Model

Oven Temperature Setting	Estimated Coefficients
(Intercept)	4.0222
Temperature = 325 F	0.3111
Temperature = 350 F	0.2778
Temperature = 375 F	0.4944

Reference Case Model Notation:

$$\mu(AvgHeight|Temp) = 4.0222 + 0.3111 * I_{Temp=325} + 0.2778 * I_{Temp=350} + 0.4944 * I_{Temp=375} + \epsilon$$

With indicator variables defined as follows:

- $I_{Temp=325} = 1$ is 1 when $Temperature = 325°F$ and 0 otherwise,
- $I_{Temp=350} = 1$ is 1 when $Temperature = 350°F$, and 0 otherwise, and
- $I_{Temp=375} = 1$ is 1 when $Temperature = 375°F$ and 0 otherwise.
- Additionally, $\epsilon \sim N(0, \sigma^2)$ is an error term.

Confidence Interval for Reference Case Model Coefficients

Created with the 'gtsummary' package in R (Sjoberg, 2021).

Characteristic	Beta	95% CI	p-value
Temps.F			0.071
300	—	—	
325	0.31	-0.05, 0.67	
350	0.28	-0.08, 0.63	
375	0.49	0.14, 0.85	

Model 2: Cell Means Model

Oven Temperature Setting	Estimated Coefficients
Temperature = 300 F	4.0222
Temperature = 325 F	4.3333
Temperature = 350 F	4.3000
Temperature = 375 F	4.5167

Cell Means Model Notation:

$$\mu(AvgHeight|Temp) = 4.0222 * I_{Temp=300} + 4.3333 * I_{Temp=325} + 4.3000 * I_{Temp=350} + 4.5167 * I_{Temp=375} + \epsilon$$

With indicator variables defined as follows:

- $I_{Temp=300} = 1$ is 1 when $Temperature = 300^{\circ}F$ and 0 otherwise,
- $I_{Temp=325} = 1$ is 1 when $Temperature = 325^{\circ}F$ and 0 otherwise,
- $I_{Temp=350} = 1$ is 1 when $Temperature = 350^{\circ}F$, and 0 otherwise, and
- $I_{Temp=375} = 1$ is 1 when $Temperature = 375^{\circ}F$ and 0 otherwise.
- Additionally, $\epsilon \sim N(0, \sigma^2)$ is an error term.

Confidence Interval for Cell Means Model Coefficients

Created with the 'gtsummary' package in R (Sjoberg, 2021).

Characteristic	Beta	95% CI	p-value
Temps.F			<0.001
300	4.0	3.8, 4.3	
325	4.3	4.1, 4.6	
350	4.3	4.0, 4.6	
375	4.5	4.3, 4.8	

Tables

Table 1: Summary of One-Way ANOVA Analysis for the Cell Means Model

	Degrees of Freedom	Sums of Squares	Mean Squared Error	F value	P-value
Oven Temperature	4	221.539	55.385	1538.7	1.418e-11
Residuals	8	0.288	0.036		

Table 2: Results of Tukey's Post Hoc Comparisons and 95% Family-Wise Confidence Intervals (Cell Means Model)

Oven Temperature Settings Being Compared	Estimated Difference in Average Batch Heights	95% CI Lower Bound	95% CI Upper Bound
325 F - 300 F	0.311111	-0.184722	0.806944
350 F - 300 F	0.277778	-0.218055	0.773611
375 F - 300 F	0.494444	-0.001389	0.990278
350 F - 325 F	-0.033333	-0.529166	0.462500
375 F - 325 F	0.183333	-0.312500	0.679166
375 F - 350 F	0.216667	-0.279166	0.712500