STA 9797
Fall 2021
Homework # 1
Erik Carrion

Question 1

Looking at the histograms of Total Cholesterol for the different types of chocolate we see that Total Cholesterol for the Dark Chocolate treatment is unimodal and slightly right skewed. Total Cholesterol for the Milk Chocolate group appears uniformly distributed, and Total Cholesterol for the White Chocolate group looks slightly left skewed. Looking at the relative positions of the distributions, white chocolate is further to the right than both dark and milk chocolate, suggesting Total Cholesterol for the white chocolate group is higher than the other two groups.

Looking to the means and standard deviations of the three groups: Dark Chocolate (mean = 196.98, Var = 108.76) has a lower mean than Milk Chocolate (Mean = 205.14, Var= 119.58) and White Chocolate (Mean = 223.31, Var = 318.00). Whether these differences are significant, however, is to be determined. We also note that the variance of Total Cholesterol for the White Chocolate Group is higher than for Dark Chocolate and Milk Chocolate.

Question 2

To determine if we should employ a fixed effects or random effects design, we must look to the research question, how the study was formulated and executed, the treatments, and finally the intended inference.

The researcher asks, "Does the type of chocolate you eat have an effect on your total cholesterol level?" To determine this, 45 people were randomly selected and assigned to one of three groups: Dark Chocolate, Milk Chocolate, or White Chocolate. For 30 days each subject ate a level dosage of 1.7 ounces of their given type of chocolate and at the end of the study their cholesterol was measured for comparison among the groups.

Nowadays you can find many different types of chocolate – each formulated in one way or another. They are differentiated by country of origin, cultivation (organic or not), their percentage of dark chocolate, additives such as nuts, sea salt, or spices – the list goes on as manufacturers compete fiercely to set themselves apart in the marketplace. Given all of this, the choice of Dark, Milk, or White Chocolate represents a sample from the total population of chocolate types. Furthermore, if we consider that one company's dark chocolate may be different from another's due to any number of reasons such as the competency of the farmer growing the chocolate, efficiency in manufacturing, proprietary roasting methods, etc., then the chocolate chosen for the study represents a sample from a population of chocolate manufacturers. Given these considerations, it might be reasonable to employ a random effects model.

However, if we consider that the researcher is not interested in the effects of a specific brand's chocolate (Hershey's vs. Cadbury) or a specific sub-type of chocolate (i.e. dark chocolate vs. sea salted milk chocolate), but rather only in the effects of the chocolate chosen as treatments, we would be guided towards a fixed effects model. Since the researcher is only interested in the difference in Total Cholesterol for these three specific types of chocolate one could argue that Dark Chocolate, Milk Chocolate, and White Chocolate represent the totality of all levels in the population and it is thus fully represented in the study, and we should therefore employ a fixed effects model.

STA 9797
Fall 2021
Homework # 1
Erik Carrion

Question 2 – Model Discussion

The Fixed Effects Model yields an F-Statistic of 14.97 which has a p-value of < .0001 indicating we can reject the null hypothesis that the group means are equivalent. We can see this visually in the side-by-side box plots. While the means for the Dark Chocolate and Milk Chocolate groups are similar, the White Chocolate group not only has a larger mean but a larger variance as well. This is confirmed by Levene's Test which yields an F-Value of 4.58 and a p-value of .0158 and so we reject the null hypothesis of equal variances among the three groups.

This could threaten our ANOVA analysis as it relies on the homogeneity of variance. However, ANOVA is relatively robust to minor deviations from this assumption, so we are not overly concerned for the Milk Chocolate group's larger variance. A more pressing concern might be the data's departure from normality. Assessing such a departure, however, is outside the scope of the current analysis.

The model yields an R^2 of .416 which might be sufficient for an observational study, but for a study where the subjects are randomly chosen and assigned a treatment, we would hope for slightly better performance. Unfortunately, the study was limited in its scope and design as it did not have a control group. Therefore, we don't know if people who ate no chocolate of any type for 30 days fared any better.

Question 3

When looking to compare individual groups, our main concern is to avoid drawing conclusions based on a nominal significance level as doing so would lead to an increase in Type 1 error and so any such analysis must compensate for this risk. With limited data, few groups, and the ANOVA results in hand, we can avoid this risk through visual inspection. However, if the differences are subtle, thus requiring a group-by-group comparison, we can do so utilizing Tukey's Studentized Range (given we have a balanced design) which corrects for the family wise error rate.

In performing the Tukey test, we have a critical value of 3.436. If we were performing a two tailed t test at a significance level of .05, our critical value would be approximately 2.0 and so we have a much higher bar to pass. This test confirms our visual inspection which concluded that the Dark Chocolate & Milk Chocolate group were not significantly different and that the White Chocolate group was significantly different from the other groups.

Question 4

Hershey's should investigate consumer sentiment surrounding dark chocolate and its health benefits and study the relationship of those sentiments to sales of Dark Chocolate. Using the results of this study they could then investigate if it is worthwhile to invest in a new line of Dark Chocolates that are positioned as a healthier alternative to the competition.

Appendix A: SAS Output

Type = Dark

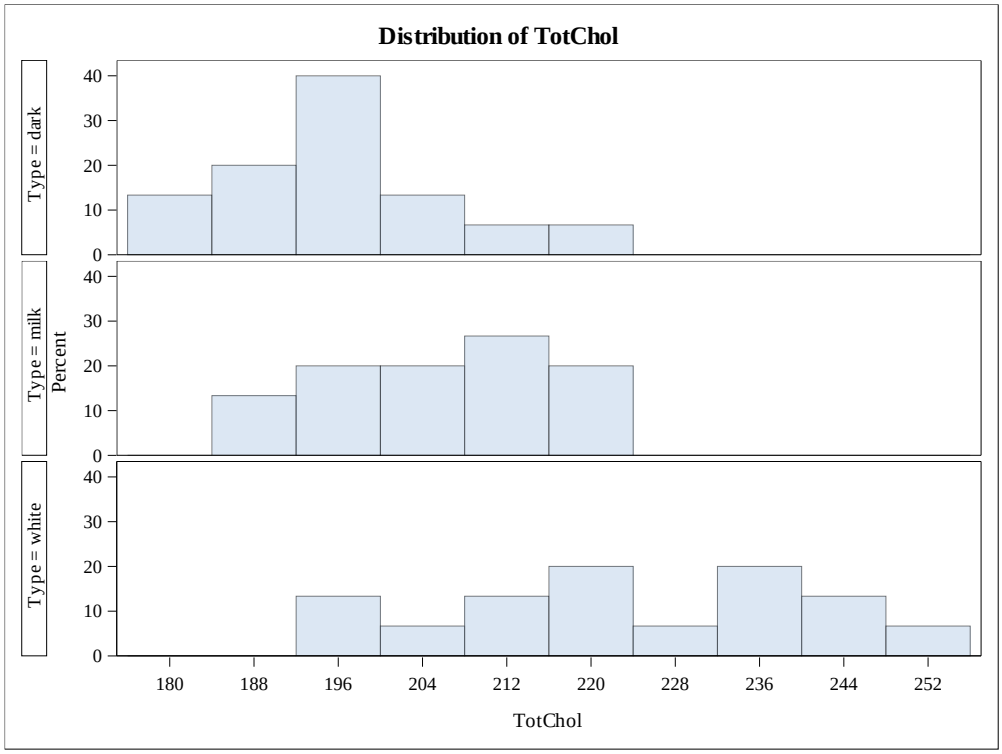
Type	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
dark	15	Subject	Subject	15	8.0000000	4.4721360	1.0000000	15.0000000
		TotChol	TotChol	15	196.980000	10.428614	182.70000	220.500000

Type = Milk

Type	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
milk	15	Subject	Subject	15	23.0000000	4.4721360	16.0000000	30.0000000
		TotChol	TotChol	15	205.1366667	10.9351247	188.1000000	221.3500000

Type = White

Type	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
white	15	Subject	Subject	15	38.0000000	4.4721360	31.0000000	45.0000000
		TotChol	TotChol	15	223.3133333	17.8324991	195.7000000	253.6500000



STA 9797
Fall 2021
Homework # 1
Erik Carrion

Effect of Chocolate Type on Total Cholesterol - Fixed Effects Design

The ANOVA Procedure

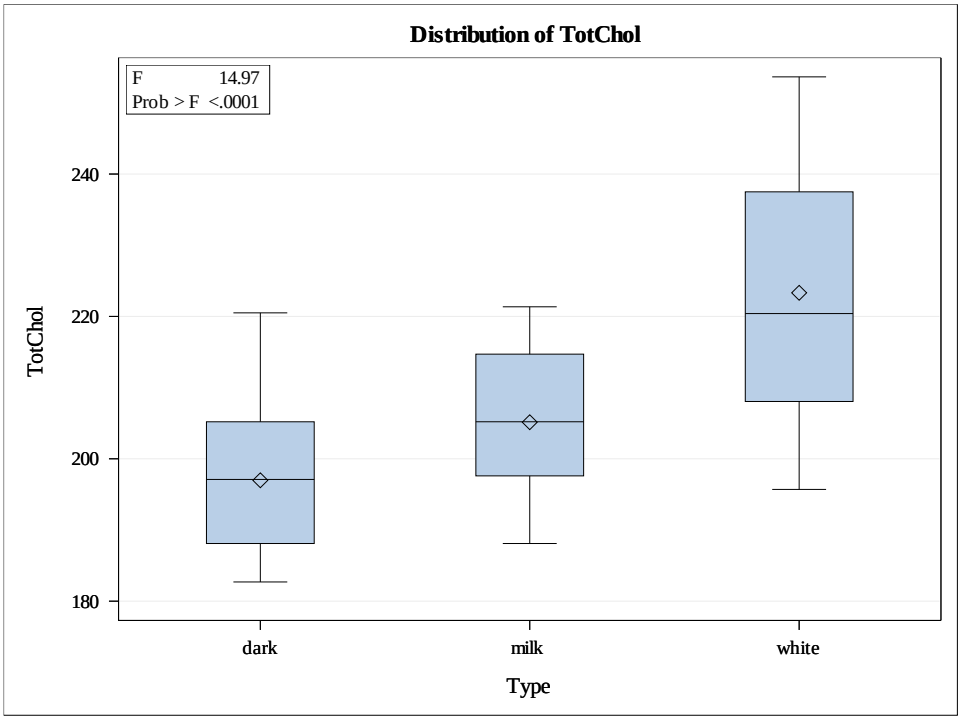
Class Level Information		
Class	**Levels**	**Values**
Type	3	dark milk white

Number of Observations Read	45
Number of Observations Used	45

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5451.83433	2725.91717	14.97	<.0001
Error	42	7648.63367	182.11033		
Corrected Total	44	13100.46800			

R-Square	Coeff Var	Root MSE	TotChol Mean
0.416156	6.473063	13.49483	208.4767

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Type	2	5451.834333	2725.917167	14.97	<.0001



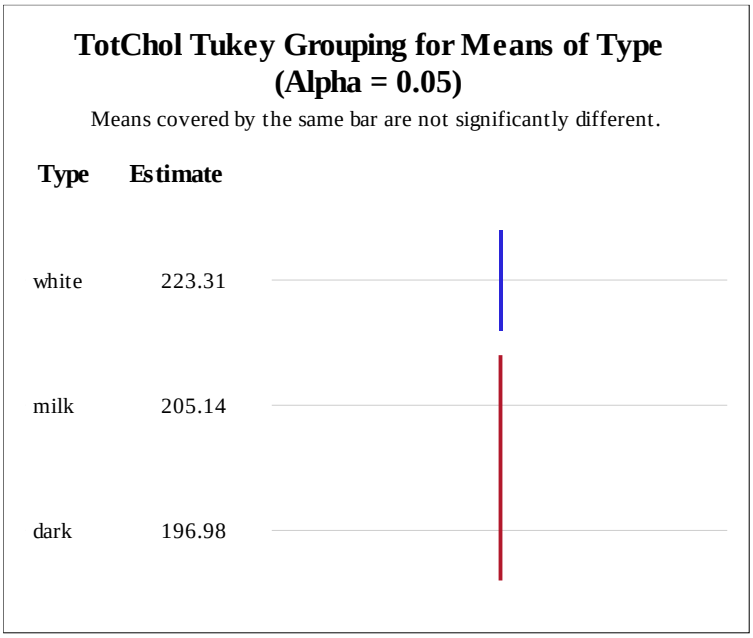
Levene's Test for Homogeneity of TotChol Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Type	2	362688	181344	4.58	0.0158
Error	42	1661861	39568.1		

Tukey's Studentized (HSD) Test for TotChol

Note: This test controls the Type I experimentwise error, but it generally has a higher Type II error rate than REGWQ

Alpha	0.05
Error Degrees of Freedom	42
Error Mean Square	182.1103
Critical Value of Studentized Range	3.43580
Minimum Significant Difference	11.972

STA 9797
Fall 2021
Homework # 1
Erik Carrion



Appendix B: SAS Code

```
data _null_;
      rc=dlgcdir("Z:\OneDrive - Smart City Real Estate\Personal\Baruch\S5\STA 9797\HW");
      put rc=;
   run;
proc import file = "chocolate.xlsx" dbms = xlsx out = chocolate;
run;

proc print data = chocolate;
run;
ods rtf file = "STA 9797 HW 1.rtf";
ods graphics on;
proc means data = chocolate;
       by Type;
       class Type;
       run;
proc univariate data = chocolate;
       class Type;
       var TotChol;
       histogram TotChol / nrows = 3;
       ods select histogram;
       title "Distribution of Total Cholesterol by Chocolate Type";
       run;
proc anova data = chocolate;
       class Type;
       model TotChol = Type;
       means Type / tukey hovtest=levene;
       title "Effect of Chocolate Type on Total Cholesterol - Fixed Effects Design";
       run;
ods graphics off;
ods rtf close;
```