

**Instructions:** Follow all guidelines for homework submission when answering the following questions. All hypotheses should be stated using parameters, when possible. All inferences should be carried out using  $\alpha = 0.05$  unless otherwise stated.

1. The Umbrella Corporation is attempting to grow large amounts of a rare, but horrible organism known as the T-virus. They have run an initial experiment in which there are two primary factors (A: incubation time; B: culture medium), each at two levels. The response variable is the size of the resulting virus colony. The factorial experiment was replicated six times, and we plan to block on replicate. The responses printed in the table below are ordered top to bottom, from replicate 1 to replicate 6. (22 pts)

	Time	Culture Medium	
		1	2
6		24	26
		22	22
		27	29
		22	28
		24	30
		29	25
12		35	33
		40	29
		36	31
		39	36
		39	33
		37	31

- a) Compute the values of the four symbols below. (4)

symbol	I	A	B	AB	Value
(1)	+	-	-	+	148
<i>a</i>	+	+	-	-	162
<i>b</i>	+	-	+	-	226
<i>ab</i>	+	+	+	+	193

- b) Use Minitab to estimate the factor effects, and present a table of output displaying these values. Demonstrate that you know how these values are obtained by manually calculating the effect of factor A (incubation time). (4)

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		30.292	0.530	57.15	0.000	
Blocks						
1		-0.79	1.19	-0.67	0.514	1.67
2		-2.04	1.19	-1.72	0.105	1.67
3		0.46	1.19	0.39	0.704	1.67
4		0.96	1.19	0.81	0.431	1.67
5		1.21	1.19	1.02	0.324	1.67
A	9.250	4.625	0.530	8.73	0.000	1.00
B	-1.750	-0.875	0.530	-1.65	0.120	1.00
A*B	-3.750	-1.875	0.530	-3.54	0.003	1.00

$$A = (-148 + 162 - 226 + 193)/4 = -4.75$$

- c) Provide the ANOVA table for the factorial model (blocking on replicate), and give a summary statement about the presence of significant main effects / interactions. (4)

#### Analysis of Variance

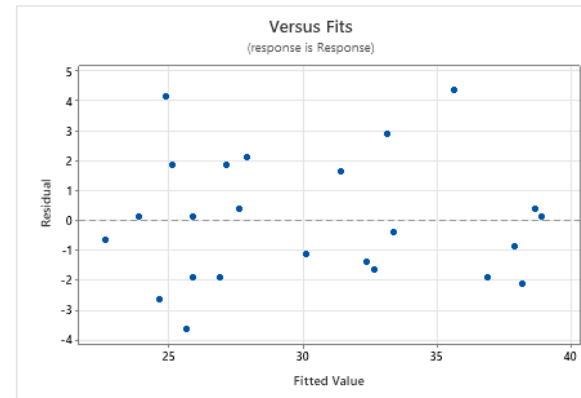
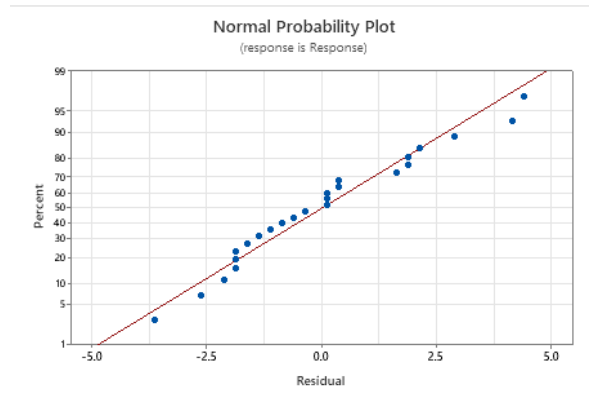
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	8	645.83	80.729	11.97	0.000
Blocks	5	29.71	5.942	0.88	0.517
Linear	2	531.75	265.875	39.44	0.000
A	1	513.37	513.375	76.15	0.000
B	1	18.37	18.375	2.73	0.120
2-Way Interactions	1	84.38	84.375	12.52	0.003
A*B	1	84.38	84.375	12.52	0.003
Error	15	101.13	6.742		
Total	23	746.96			

There is significant effects from factor A as well as the interaction between AB. Blocking doesn't seem to be very significant.

- d) By what principle does culture medium remain in the model? (2)

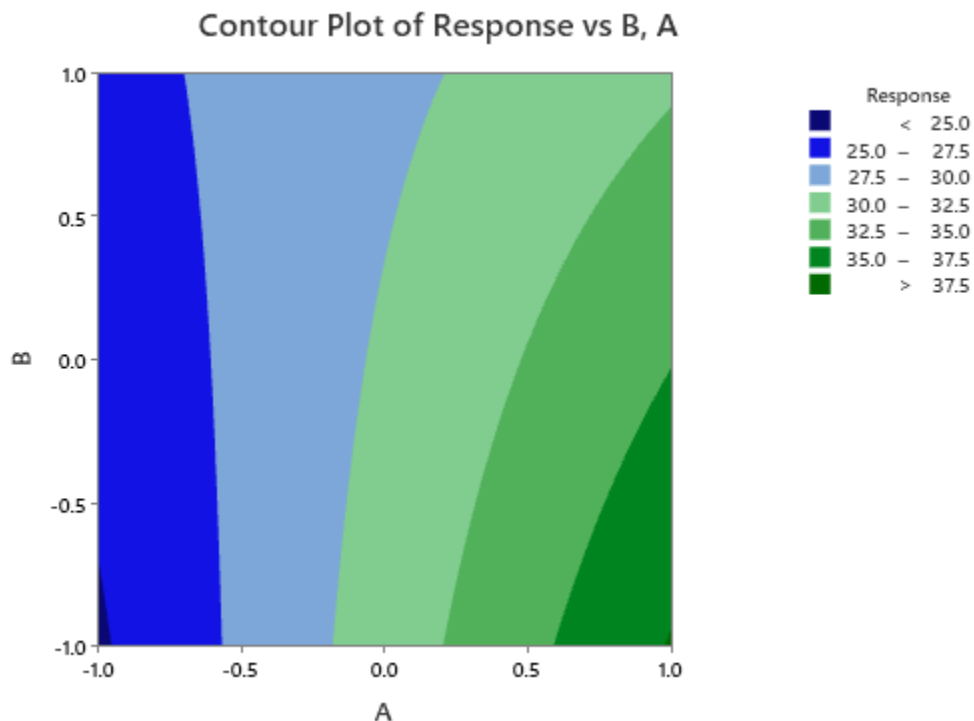
Hierarchy principle

- e) Evaluate the model assumptions using residual diagnostics. (4)



Pretty good looking residuals. Normality and constant variance assumption isn't violated.

- f) The board of directors at the Umbrella Corporation is expecting you to show them a **contour plot**, as opposed to other commonly used plots – you do not want to disappoint them. Present the contour plot, and suggest advantageous levels of incubation time and culture medium. (4)



Low level for culture medium and high level for incubation time

- g) State the value of  $SSE$  from the current model. What would the  $SSE$  have been if we had chosen not to employ blocking?

101.13. Error variance would be higher

2. The data for this problem come from an article titled “Electrochemical Degradation of Distillery Spent Wash Using Catalytic Anode: Factorial Design of Experiments” in Chemical Engineering Journal. The goal of the experiment was to determine how to reduce the chemical oxygen demand (COD) by as much as possible. The response variable was recorded as percent COD removal. The factors were A: electric current density (14 mA/cm<sup>2</sup> or 42 mA/cm<sup>2</sup>), B: dilution percent (10% or 30%), reaction time (2 hrs or 5 hrs), and D: pH (4 or 9). Due to budget constraints, only a single full replicate of the experiment could be performed. The data appear in the file cod.txt (UBLearns). (18 pts)

a) State the name of the design to be used. (1)

2<sup>4</sup> full factorial design

b) Fit the full factorial model, and provide a table containing estimates of the factor effects. Highlight the three largest (in absolute value) effects. (2)

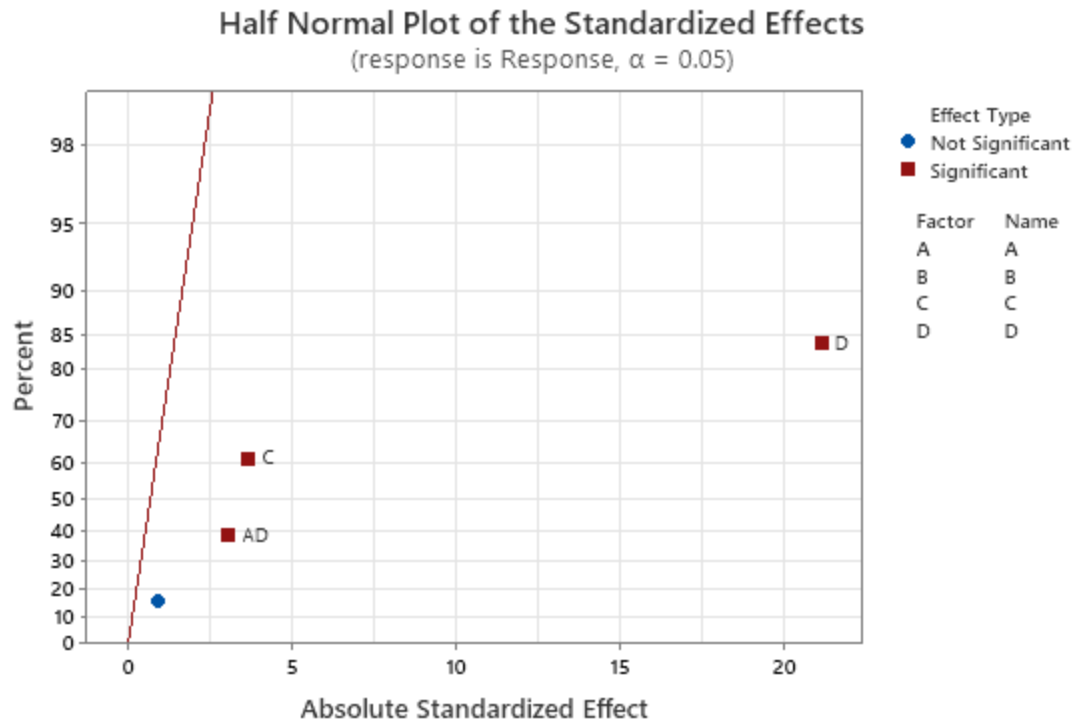
### Coded Coefficients

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		49.99	*	*	*	
A	1.4304	0.7152	*	*	*	1.00
B	2.610	1.305	*	*	*	1.00
C	5.547	2.774	*	*	*	1.00
D	32.26	16.13	*	*	*	1.00
A*B	0.5529	0.2764	*	*	*	1.00
A*C	0.9354	0.4677	*	*	*	1.00
A*D	-4.619	-2.310	*	*	*	1.00
B*C	-0.4679	-0.2339	*	*	*	1.00
B*D	-1.2959	-0.6479	*	*	*	1.00
C*D	2.449	1.225	*	*	*	1.00
A*B*C	-2.090	-1.045	*	*	*	1.00
A*B*D	1.7784	0.8892	*	*	*	1.00
A*C*D	-0.9966	-0.4983	*	*	*	1.00
B*C*D	-0.5459	-0.2729	*	*	*	1.00
A*B*C*D	-1.0016	-0.5008	*	*	*	1.00

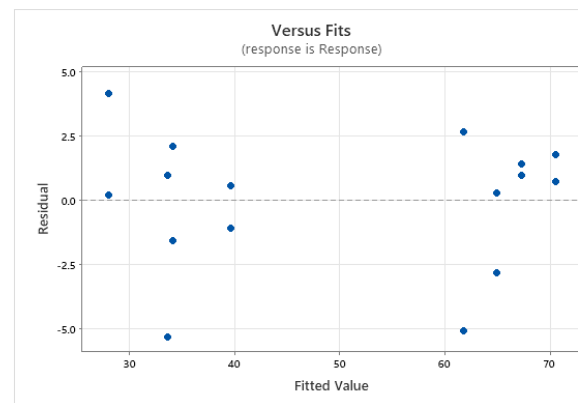
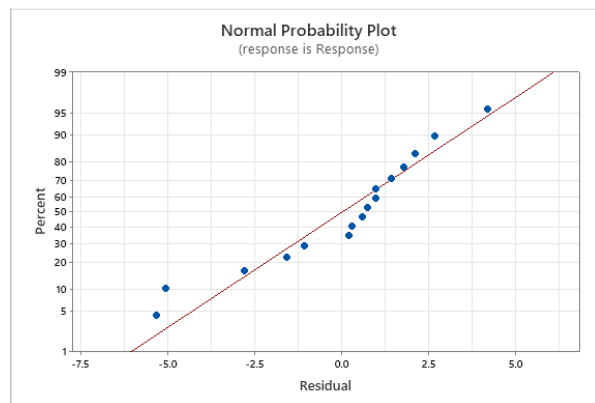
c) Explain why the F statistics are missing from the initial ANOVA table. (2)

It's a single replicate

d) Refine the model, and present the half-normal plot that corresponds to the most parsimonious model. (4)



e) Conduct a brief residual analysis. (2)



Normality assumption and variance assumption isn't violated.

f) State which, if any, of the four primary factors are absent from the refined model. (1)

B was absent

g) Build a new worksheet in Minitab that will be used to analyze a  $2^3$  factorial design, taking advantage of what we noted in part (f). After performing appropriate sorting and pasting the response variable into the new worksheet, provide a screenshot of the first 3 rows of this worksheet. (2)

### Coded Coefficients

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		49.986	0.747	66.94	0.000	
A	1.430	0.715	0.747	0.96	0.366	1.00
C	5.547	2.774	0.747	3.71	0.006	1.00
D	32.262	16.131	0.747	21.60	0.000	1.00
A*C	0.935	0.468	0.747	0.63	0.549	1.00
A*D	-4.619	-2.310	0.747	-3.09	0.015	1.00
C*D	2.449	1.225	0.747	1.64	0.140	1.00
A*C*D	-0.997	-0.498	0.747	-0.67	0.523	1.00

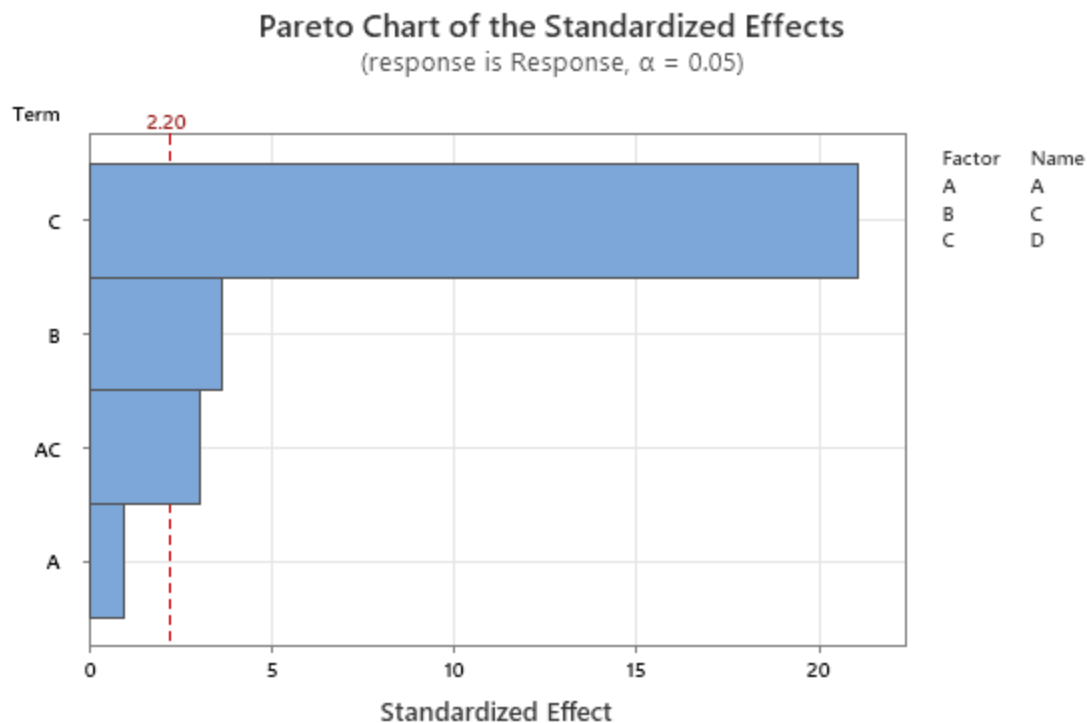
### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.98700	98.41%	97.01%	93.63%

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	7	4411.33	630.19	70.63	0.000
Linear	3	4294.52	1431.51	160.44	0.000
A	1	8.18	8.18	0.92	0.366
C	1	123.08	123.08	13.80	0.006
D	1	4163.25	4163.25	466.62	0.000
2-Way Interactions	3	112.84	37.61	4.22	0.046
A*C	1	3.50	3.50	0.39	0.549
A*D	1	85.35	85.35	9.57	0.015
C*D	1	23.99	23.99	2.69	0.140
3-Way Interactions	1	3.97	3.97	0.45	0.523
A*C*D	1	3.97	3.97	0.45	0.523
Error	8	71.38	8.92		
Total	15	4482.70			

- h) Refine the  $2^3$  factorial model and again provide a half-normal plot. State whether you have reproduced the plot from part (d). (4)



Besides the names being slightly different, it's the same

3. The owner of a massive apple orchard commissioned a study investigating juice sales, which he has interest in maximizing. The study tracked the quantity of juice sold (in gallons) across six months of the year and six grocery store locations. The primary factor was the type of container that the apple juice was sold in, labeled i-vi. The data appear below. (24 pts)

Month	Store					
	1	2	3	4	5	6
1	i = 46	vi = 64	v = 123	ii = 76	iii = 35.8	iv = 155.2
2	iii = 58.5	ii = 43.5	vi = 120.4	v = 79	iv = 66.9	i = 128.3
3	iv = 66.7	iii = 28	ii = 86	i = 64	vi = 63.5	v = 106.5
4	ii = 47.5	iv = 66.6	i = 95	iii = 83	v = 75.5	vi = 167.8
5	vi = 76.5	v = 43.5	iii = 108	iv = 104.5	i = 65	ii = 185.8
6	v = 61.5	i = 38.6	iv = 134.5	vi = 106.5	ii = 71.5	iii = 171

- a) Which design should be used to analyze this data set, capable of blocking on both month and store? (1)

Randomized Complete Block Design

- b) Ensure that you have properly entered the data by reproducing the sample grand mean as  $\bar{Y}_{...} = 86.4889$ . (1)

4	C5	C6
onse		
46.0		86.4889

Store result in variable: C6  
 Expression:  
 MEAN('Response')

c) Fit the appropriate model in Minitab, and provide the ANOVA table. (4)

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Month	5	3386	677.3	4.01	0.011
Store	5	46618	9323.6	55.16	0.000
Container	5	3493	698.6	4.13	0.010
Error	20	3381	169.0		
Total	35	56878			

d) Use the output above to address the main study question. Provide the hypotheses, test statistic, p-value, and conclusion in context. (4)

$H_0$ : Mean quantity of juice sold all the same

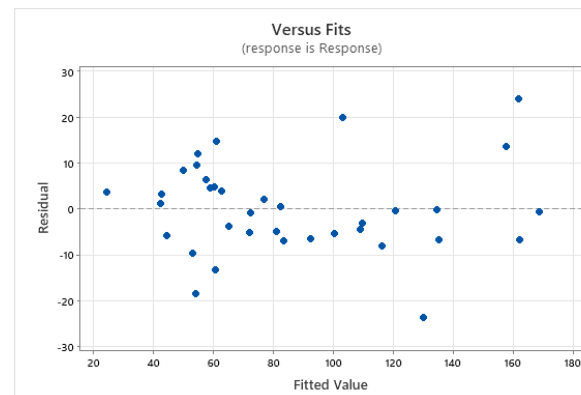
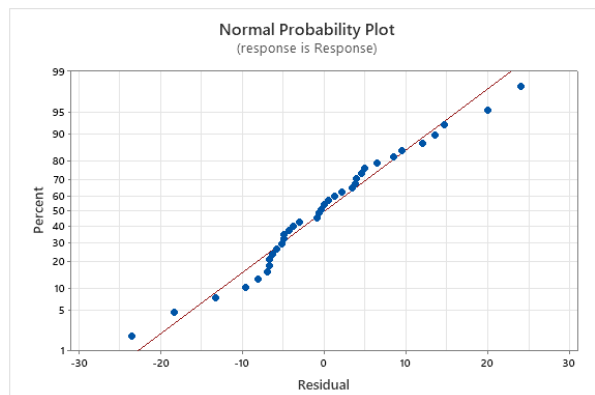
$H_a$ : Mean quantity of juice sold not the same

Test statistic: 4.13

P-value: 0.010

Conclusion: Reject  $H_0$ . Container matters for how much juice is sold.

e) Briefly assess the model assumptions. (4)



Both normality and constant variance assumption isn't violated



- f) Ignore the blocking factors, and fit a one-way ANOVA model to address the same study question as the original model. Provide the ANOVA table, then carry out the test of interest. (6)

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Container	5	3493	698.6	0.39	0.850
Error	30	53385	1779.5		
Total	35	56878			

H<sub>0</sub>: Mean quantity of juice sold all the same

H<sub>a</sub>: Mean quantity of juice sold not the same

Test statistic: 0.39

P-value: 0.85

Conclusion: Fail to reject H<sub>0</sub>. It doesn't matter what the container the juice is sold in

- g) Compare the results of part (d) and part (f). Explain what causes the discrepancy. (4)

There must be a significant interaction between store and container sold or month and container sold or all 3.

4. With supply chain interruptions and shortages of electronic components worldwide, we need to discover how the semiconductor manufacturing process can be optimized. Five factors will be investigated: aperture setting (factor A; small vs large), exposure time (factor B; 1 vs 2 minutes), development time (factor C; 30 vs 45 seconds), mask dimension (factor D; small vs large), and etch time (factor E; 14 vs 15 minutes). Materials are expensive, and a single replicate of the experiment was performed in two blocks, using the ABCDE interaction to determine block assignments. The data appear in the file semiconductor.txt (UBLearns). (20 pts)

- a) The data file does not contain a column for Blocks. Paste the data into a Minitab worksheet, and create a column for Blocks according to the rule described above. Note that this should be done in the original worksheet, i.e. not a worksheet created using the DOE menu. Provide a screenshot of the first three rows of this original worksheet. (4)

	A	B	C	D	E	Response		Blocks
1	-1	-1	-1	-1	-1	7	-1	1
2	-1	-1	-1	-1	1	8	1	2
3	-1	-1	-1	1	-1	8	1	2

- b) Use Minitab's DOE menu to create a design for a single replicate of a  $2^5$  design using two blocks. After appropriately sorting both worksheets, paste in the responses from the original worksheet, and fit the initial model. Which estimated factor effects are largest in absolute value (give the top four)? (4)

### Coded Coefficients

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		30.53		*	*	*
Blocks						
1		0.09375		*	*	* 1.00
A	11.812	5.906		*	*	* 1.00
B	33.94	16.97		*	*	* 1.00
C	9.688	4.844		*	*	* 1.00
D	-0.8125	-0.4062		*	*	* 1.00
E	0.4375	0.2188		*	*	* 1.00
A*B	7.937	3.969		*	*	* 1.00
A*C	0.4375	0.2187		*	*	* 1.00
A*D	-0.06250	-0.03125		*	*	* 1.00
A*E	0.9375	0.4688		*	*	* 1.00
B*C	0.06250	0.03125		*	*	* 1.00
B*D	-0.6875	-0.3438		*	*	* 1.00
B*E	0.5625	0.2813		*	*	* 1.00
C*D	0.8125	0.4062		*	*	* 1.00
C*E	0.3125	0.1562		*	*	* 1.00
D*E	-1.1875	-0.5938		*	*	* 1.00
A*B*C	-0.4375	-0.2188		*	*	* 1.00

A*B*D	0.3125	0.1562	*	*	*	1.00
A*B*E	-0.18750	-0.09375	*	*	*	1.00
A*C*D	-0.4375	-0.2188	*	*	*	1.00
A*C*E	0.3125	0.1562	*	*	*	1.00
A*D*E	0.8125	0.4062	*	*	*	1.00
B*C*D	0.4375	0.2187	*	*	*	1.00
B*C*E	0.9375	0.4688	*	*	*	1.00
B*D*E	0.18750	0.09375	*	*	*	1.00
C*D*E	-0.8125	-0.4063	*	*	*	1.00
A*B*C*D	-0.06250	-0.03125	*	*	*	1.00
A*B*C*E	0.18750	0.09375	*	*	*	1.00
A*B*D*E	0.9375	0.4688	*	*	*	1.00
A*C*D*E	-0.3125	-0.1562	*	*	*	1.00
B*C*D*E	-0.9375	-0.4687	*	*	*	1.00

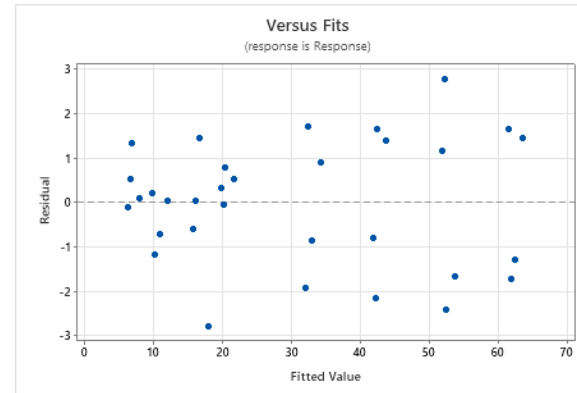
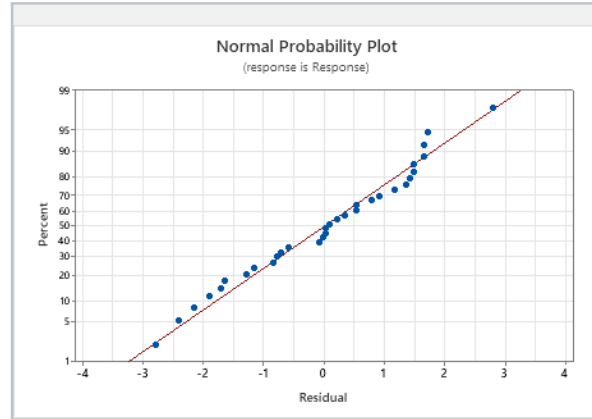
- c) Neither the table of effect estimates nor the ANOVA table includes a row for ABCDE. Explain why this is the case. (2)

Block is confounded with ABCDE

- d) Refine the model in stages until you have achieved the most parsimonious model, and provide the final model's ANOVA table. (4)

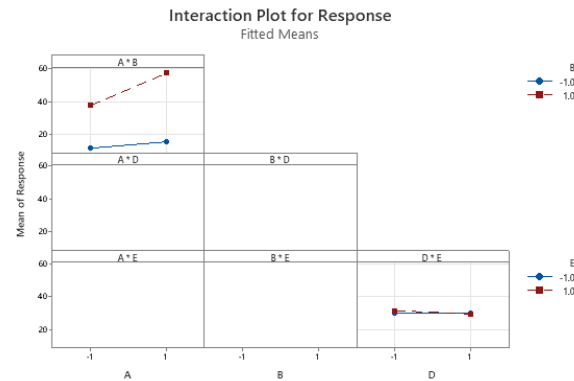
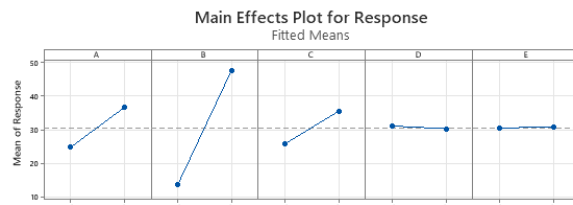
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	8	11603.5	1450.44	551.69	0.000
Blocks	1	0.3	0.28	0.11	0.747
Linear	5	11087.9	2217.58	843.48	0.000
A	1	1116.3	1116.28	424.59	0.000
B	1	9214.0	9214.03	3504.67	0.000
C	1	750.8	750.78	285.57	0.000
D	1	5.3	5.28	2.01	0.170
E	1	1.5	1.53	0.58	0.453
2-Way Interactions	2	515.3	257.66	98.00	0.000
A*B	1	504.0	504.03	191.71	0.000
D*E	1	11.3	11.28	4.29	0.050
Error	23	60.5	2.63		
Total	31	11664.0			

- e) Carry out a brief residual analysis based on the final model. (2)



Normal Assumption is fine, constant variance is alright.

- f) Use main effect plots / interaction plots to determine how to maximize semiconductor yield. (4)



A, B, C, E at the high level and D at the low level. D at the high level with E at the low level. A at the high level with B at the high level

5. Lumber has been in short supply, due to mill closures and people's general unwillingness to work. Freshly-cut lumber contains notable amounts of moisture, and must be dried prior to being sold. A study was done to investigate how to minimize moisture content of newly milled lumber. Planks were cut from four species of tree (1:loblolly pine, 2:shortleaf pine, 3:yellow poplar, and 4:red gum). The location (1:central, 2:distal, or 3:proximal) from which the lumber was cut was also recorded (this refers to whether a plank was cut from the center of a log, far from the center, or near the center, respectively). The planks were then dried using one of two methods (1:rapid or 2:slow). Five replicates of the full factorial experiment were obtained. The data appear in the file tree\_moisture.txt (UBLearns). (24 pts)

- a) Use mathematical symbols to write the statistical model for the full factorial experiment using three primary factors and blocking on replicate. Provide the range of possible values for all subscripts, and state any model assumptions. (4)

$$Y_{ijkl} = \mu + \tau_i + \beta_j + \gamma_k + (\tau\beta)_{ij} + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + (\tau\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

for  $i = 1, 2, 3, 4$  ;  $j = 1, 2, 3$  ;  $k = 1, 2$   
 $l = 1, \dots, n$   $\epsilon_{ijkl} \sim N(0, \sigma^2)$

- b) Ensure that you have successfully read in the data by producing the sample grand mean moisture content as 1137.23. (1)

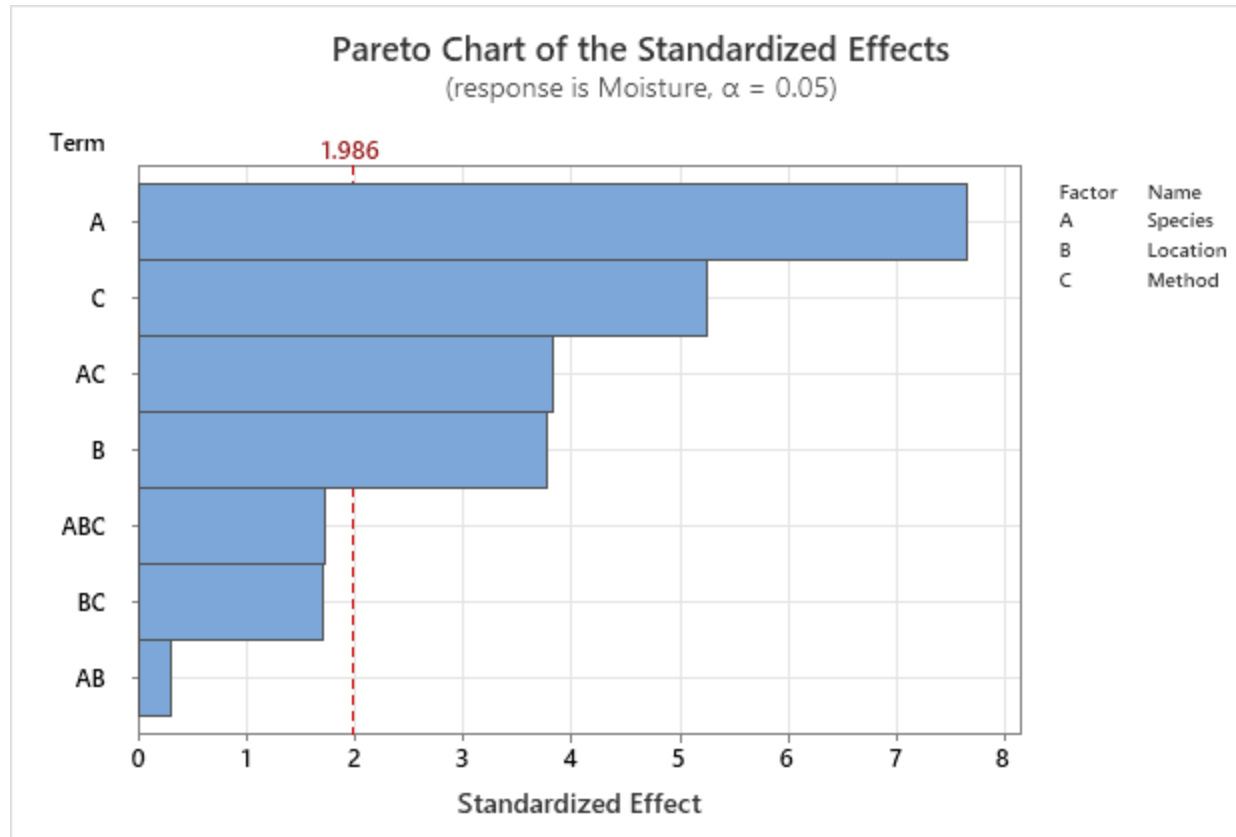
	C6	C7	C8
e			
4		1137.23	

Store result in variable: C7

Expression:

MEAN('Moisture')

- c) Fit the full model using the DOE section of Minitab, and present the Pareto chart. Explain how this plot is used to determine which terms are considered to have a significant effect. (3)

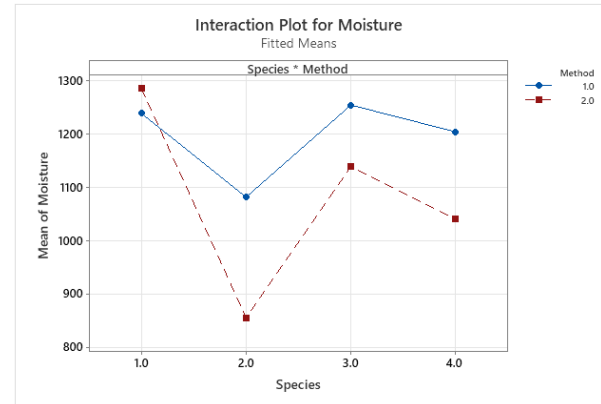
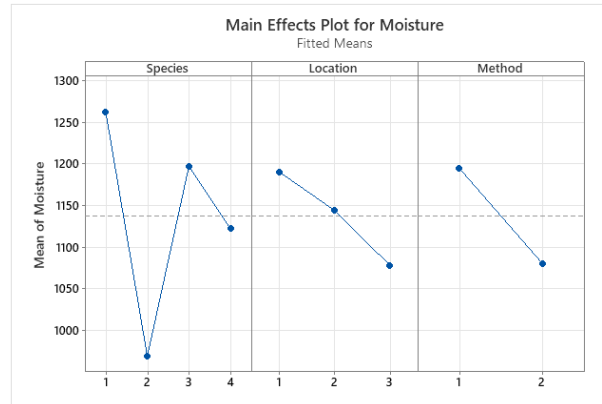


Anything to the right of the red dotted line is considered to have a significant effect.

- d) Refine the model in stages until you have achieved the most parsimonious model.  
Provide the ANOVA table from the final model. (4)

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	13	2519218	193786	12.89	0.000
Blocks	4	131863	32966	2.19	0.075
Linear	6	2081403	346901	23.07	0.000
Species	3	1432590	477530	31.75	0.000
Location	2	254589	127294	8.46	0.000
Method	1	394224	394224	26.21	0.000
2-Way Interactions	3	305952	101984	6.78	0.000
Species*Method	3	305952	101984	6.78	0.000
Error	106	1594069	15038		
Total	119	4113287			

- e) Provide interaction plots corresponding to the terms that remain in the model. Interpret these plots to identify how moisture content can be minimized. (4)



To minimize moisture, use shortleaf pine in proximal and drying it slowly.  
(Species: 2, Location: 3, Method: 2). Interaction between species and method is the same.

- f) After seeing the interaction plot, a lumber executive notices that the slow method of drying planks seems to produce lower moisture content across all tree species but one. He asks whether the mean moisture content for rapid-dried loblolly pine is any different from mean moisture content for slow-dried loblolly pine. Obtain Tukey's tests for all pairwise differences between treatment means involving species and method. Carry out the test suggested by the executive, stating the hypotheses, test statistic, p-value, and conclusion in context. (6)

H<sub>0</sub>: Rapid dried loblolly pine has the same mean moisture as slow-dried loblolly pine

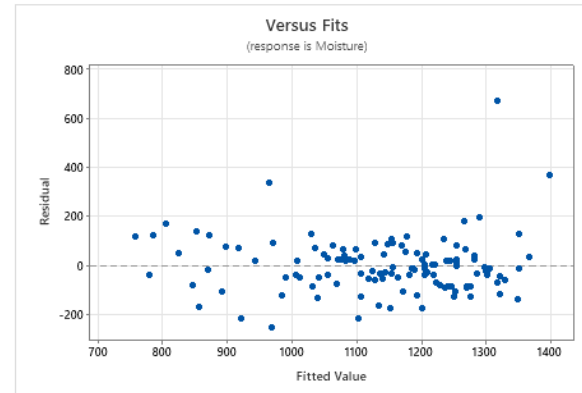
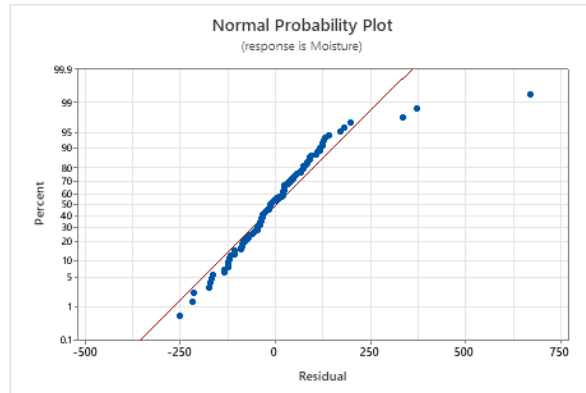
H<sub>a</sub>: Rapid dried loblolly pine has different mean moisture as slow-dried loblolly pine

Test statistic: 1.02

p-value: 0.971

conclusion: Fail to reject H<sub>0</sub>. There is not enough evidence that suggest rapid dried loblolly pine mean moisture is different than slow dried loblolly pine.

- g) Briefly address the model assumptions using the refined model. (2)



Normality and constant variance assumption isn't violated.

6. Decades ago, during the cold war with Russia, aerial photography was a critical intelligence tool, giving insight into the military capabilities of enemy nations. Suppose that engineers at a Rochester-based photography company were tasked with learning how to improve the clarity of aerial photographs. Before digital photography, physical photographs had to be printed and developed using a complex process involving a chemical bath. Blurry photographs were considered useless, and the response variable in this study, which the US government wished to maximize, was a measure of clarity. Five factors were investigated, each having two levels; for convenience, we will refer to them using letters A through E. Development of these classified photographs was extraordinarily expensive at the time, and the lab could not afford to perform the full factorial experiment. The data are located in the file photos22.txt (UBLeads). (22 pts)

- a) How many runs would be required to fit the full factorial model? What is the value of  $N$  in this study? (2)

32 runs required

$N = 16$

- b) Give a concise name for the design that we intend to use. (1)

$\frac{1}{2}$  fraction of a  $2^k$  Design

- c) Fit an initial version of the model, and present a table that displays estimates of the factor effects. Highlight the three largest factor effects (absolute value) in one color, and the three smallest effects using a different color. (2)



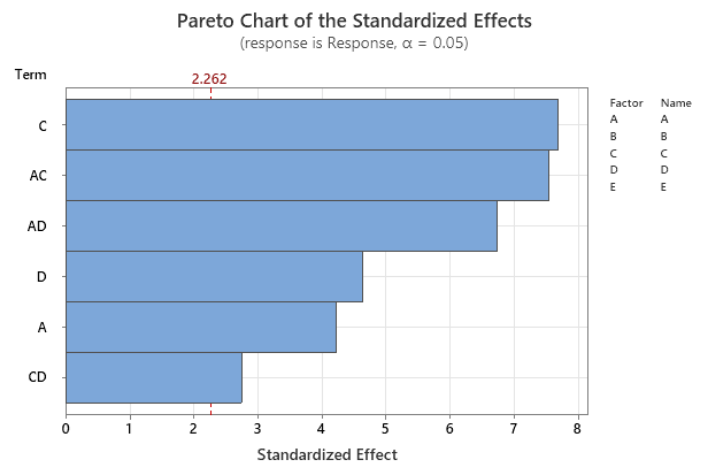
## Coded Coefficients

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		2109		*	*	*
A	110.20	55.10		*	*	1.00
B	27.10	13.55		*	*	1.00
C	200.2	100.1		*	*	1.00
D	-120.65	-60.32		*	*	1.00
E	-15.225	-7.612		*	*	1.00
A*B	-18.500	-9.250		*	*	1.00
A*C	196.53	98.26		*	*	1.00
A*D	-175.50	-87.75		*	*	1.00
A*E	-39.77	-19.89		*	*	1.00
B*C	15.875	7.937		*	*	1.00
B*D	-16.650	-8.325		*	*	1.00
B*E	-6.275	-3.137		*	*	1.00
C*D	-71.58	-35.79		*	*	1.00
C*E	-37.35	-18.68		*	*	1.00
D*E	35.52	17.76		*	*	1.00

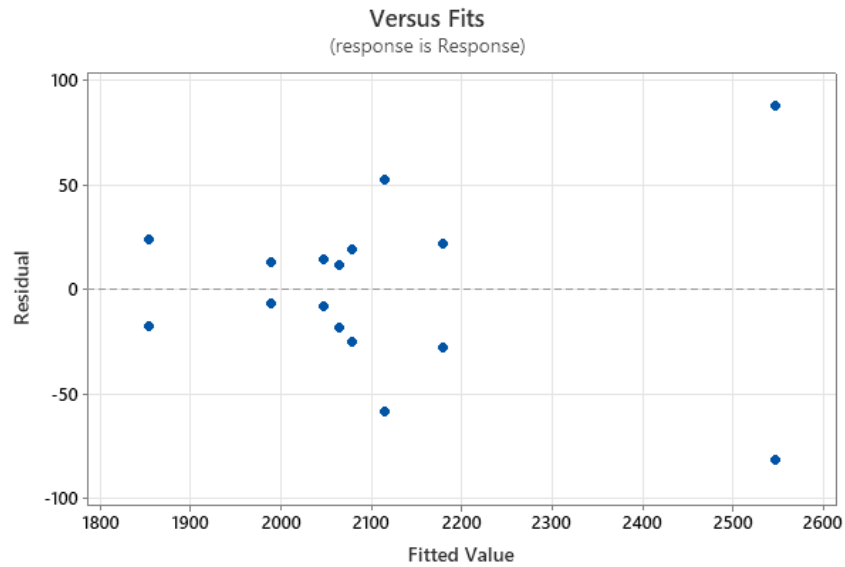
- d) Conduct the process of refining the model. Take care here (i.e. take it slow), as the Pareto chart is built on a small number of runs. Present the ANOVA table and Pareto chart for the final model. (4)

## Analysis of Variance

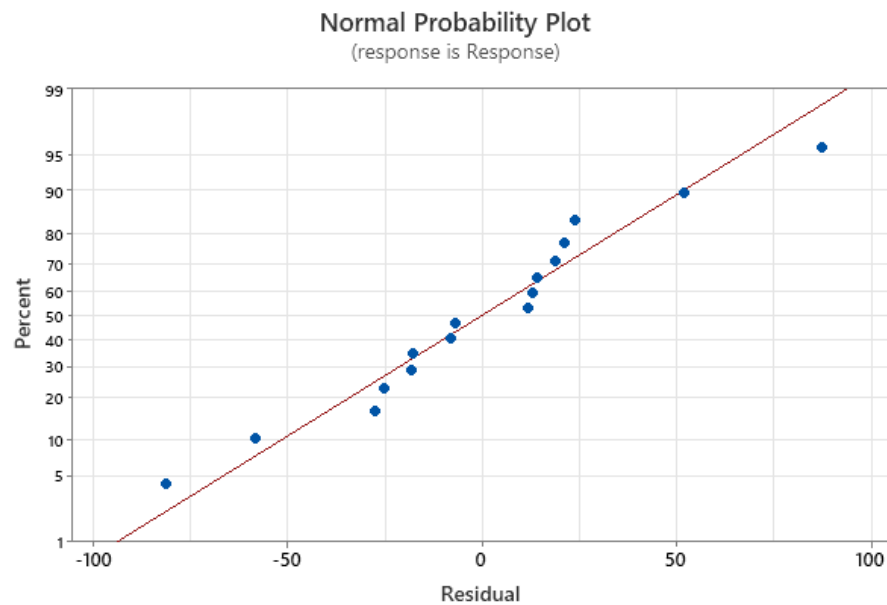
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	6	565343	94224	34.66	0.000
Linear	3	267162	89054	32.76	0.000
A	1	48576	48576	17.87	0.002
C	1	160360	160360	58.99	0.000
D	1	58226	58226	21.42	0.001
2-Way Interactions	3	298181	99394	36.56	0.000
A*C	1	154488	154488	56.83	0.000
A*D	1	123201	123201	45.32	0.000
C*D	1	20492	20492	7.54	0.023
Error	9	24465	2718		
Total	15	589808			



- e) Briefly analyze the residuals from the final model. (2)



Doesn't look the best but not terrible. Variance assumption not violated.



Normality assumption is good.

- f) Use Minitab's response optimizer to suggest advantageous levels of the factors involved in your final model. Write a succinct statement that instructs the government how they can optimize image clarity. (3)

A, B, C, at high level and D, E at the low level

Dear Government,

My team have worked long and hard to analyze the data provided to us and have come up with the best way to produce clear aerial photographs. You should put A, B, C, settings at the high level and D, E at the low level.

- g) Reflecting on your initial estimates of the factor effects, notice that there are two main effects that seem to have less of an impact on photograph clarity than other effects. Suppose these factors had been disregarded outright, and consider analyzing the data in light of the three remaining factors in photos22.txt, using the  $2^3$  factorial design. How many replicates of this experiment do we possess? What term does the textbook use to describe this situation? (2)

Great factorial design

- h) Fit the full  $2^3$  factorial model using the three active primary factors. Provide the initial ANOVA table. (2)

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	7	565501	80786	26.59	0.000
Linear	3	267162	89054	29.31	0.000
A	1	48576	48576	15.99	0.004
C	1	160360	160360	52.78	0.000
D	1	58226	58226	19.16	0.002
2-Way Interactions	3	298181	99394	32.71	0.000
A*C	1	154488	154488	50.85	0.000
A*D	1	123201	123201	40.55	0.000
C*D	1	20492	20492	6.74	0.032
3-Way Interactions	1	158	158	0.05	0.826
A*C*D	1	158	158	0.05	0.826
Error	8	24307	3038		
Total	15	589808			

- i) Refine the  $2^3$  factorial model, and present the ANOVA table corresponding to the most parsimonious model. How does this model compare to the model you reported in part (d)? (4)

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	6	565343	94224	34.66	0.000
Linear	3	267162	89054	32.76	0.000
A	1	48576	48576	17.87	0.002
C	1	160360	160360	58.99	0.000
D	1	58226	58226	21.42	0.001
2-Way Interactions	3	298181	99394	36.56	0.000
A*C	1	154488	154488	56.83	0.000
A*D	1	123201	123201	45.32	0.000
C*D	1	20492	20492	7.54	0.023
Error	9	24465	2718		
Lack-of-Fit	1	158	158	0.05	0.826
Pure Error	8	24307	3038		
Total	15	589808			

It's the same ANOVA table.