ECHE 313 Test 2

Write your name: KEY

The exam layout is such that all problems add up to 100 points. This point total is meant to guide you in time allocation for this test and will be adjusted such that this exam is 25% of your total class grade. You may use class materials, class notes, formula sheet, homework, book, and calculator (with no internet access) to aid in the completion of the exam. No computers or pads allowed. The exam is designed to be completed in 1hr and 15 min. All of the necessary tables are provided in the back of your book. Please show your work and all steps necessary to arrive at your answer. Where applicable, it is best to write out the general formals you want to use first, then write in your subbed in values so it is clear. Also, when reporting numbers, keep the same number of decimals as the data you are working with unless there are specific directions on how many sig figs to report.

Good luck!

Premise: You are a process engineer working for a food company that makes tofu. Your company implements lean and six sigma practices. <u>Unless otherwise stated, your company uses α=0.05.</u>

1) **(10 points) Deem the following statements as True or False by circling either T or F:**
    a. The shape of the F-distribution becomes more normal the lower the degrees of freedom are: T or F
    b. Randomization of experiments help eliminate nuisance variables: T or F
    c. The assumption "errors and data are independently distributed" is the constant variance assumption: T or F
    d. Empirical models are always mechanistically relevant: T or F
    e. The method of least squares is used to estimate a linear regression equation: T or F
    f. You should use R^2 to determine if your linear regression is significant: T or F
    g. R^2 adjusted is used in multiple linear regression to prevent inflation of R^2: T or F
    h. A process that is operating in the presence of assignable causes is said to be in statistical control: T or F
    i. The only time a process is considered "out of control" is when a point lands outside of the LCL or UCL: T or F
    j. Specification limits are different than the lower control limit (LCL) and upper control limit (UCL): T or F

2) **(48 points total) Due to a recent E. Coli outbreak in another company, your company wishes to improve its pasteurization process to ensure there is no chance of a potential outbreak occurring. To pasteurize your products, the tofu packages are heated to 180°F for 50 min then quickly cooled. To test if the process is effective, a microbiology team cultures samples of the tofu and counts number of harmful pathogens. Your team wishes to investigate the impact of heating time on the number of cultured pathogens after the process. Here is the data your team gathered:**

		Number of Pathogens in Sample					
	Length of Time (min)	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Average
1	50	12.0	16.0	5.0*	8.0	6.0	10.5 or 9.4
2	75	1.0	0.0	2.0	0.0	3.0	1.2 (mistake in table)
3	100	0.0	1.0	0.0	1.0	0.0	0.4
4	125	1.0	0.0	1.0	0.0	0.0	0.4

a. (3 points) List the parameter you are interested in and the *name* of the hypothesis test you would use.

Parameter of Interest: multiple means of pathogen count

Name of Test: single-factor ANOVA

b. (3 points) Write out the correct null hypothesis for this test (be specific and consider the data set – don't generalize)

$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$

Length of time is the # (see chart above)

c. (2 points) Write out the correct alternative hypothesis for this test

$H_1: \tau_i \neq 0$ for at least one i $(i = 1\text{-}4)$

d. (7 points) Write the formula(s) needed to calculate the test statistic, and calculate its value given that SSE is 92 (report test statistic to 3 sig figs)

$$F_0 = \frac{\frac{SS_{Treatments}}{(a-1)}}{\frac{SS_E}{a(n-1)}}$$

$$SS_{Treatments} = n \sum_{i=1}^{a} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= 5\left[(9.4-2.85)^2 + (1.2-2.85)^2 + (0.4-2.85)^2 + (0.4-2.85)^2\right]$$

$$= 288.15$$ (or 364.7 w/ 10.5 and 2.74 as $\bar{y}_{i.}$ and $\bar{y}_{..}$)

$\bar{y}_{..}$ = grand mean

$$= \frac{12.0 + 16.0 + \ldots + 0.0}{20}$$

$$= 2.85$$

*also will accept 3.13

$$F_0 = \frac{\frac{288.15}{3}}{\frac{92}{16}} = \boxed{16.7}$$

or w/ 364.7 would be 21.1

e. (3 points) Describe what the test statistic represents (specifically the numerator and denominator of the test statistic)

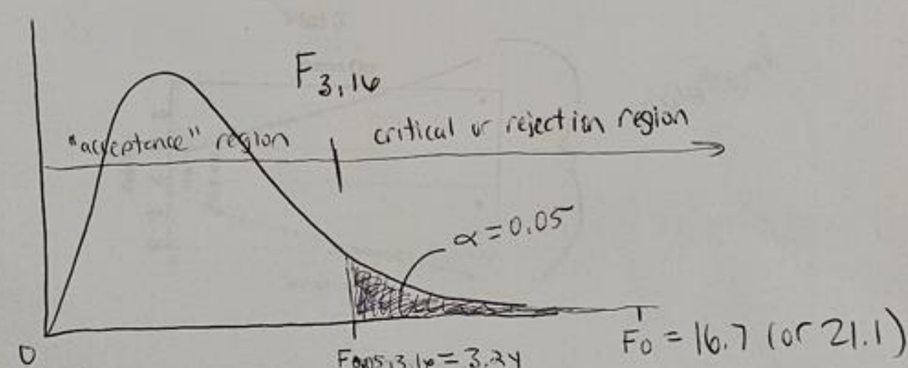
The numerator represents variance coming from treatments (signal), and the denominator represents variance coming from random error (or noise).

Side note: That is why if the test stat is large, it is more likely treatments have an impact.

f. (3 points) Define your general rejection criteria using fixed significance level testing and report your critical value(s) to 3 sig. figs.

Reject if $F_0 > F_{\alpha, a-1, a(n-1)}$ or specifically if $F_0 > F_{0.05, 3, 16} = 3.24$

g. (10 points) Sketch out your reference distribution making sure to fully define the distribution you are sketching on with the correct label, drawing it with the correct shape and where zero is. In addition, label 1) acceptance region, 2) the rejection region (also known as the critical region), 3) α, 4) the critical value(s) and 5) the test statistic.



h. (3 points) Assume your assumptions for this test are not violated. State your conclusion and remember to word your conclusion appropriately in the context of the problem statement.

Reject the H_0. We have enough evidence to say that heating time has a significant impact on pathogen count.

i. (3 points) Describe when it is appropriate to do post-hoc testing and if it is appropriate in this case.

You can proceed with post hoc testing only after rejecting the null hypothesis of ANOVA.

It is appropriate to do post hoc testing in this case!

j. (3 points) Write out the general formula for and calculate the residual for the point with a (*) in the data table. Report 2 significant figures.

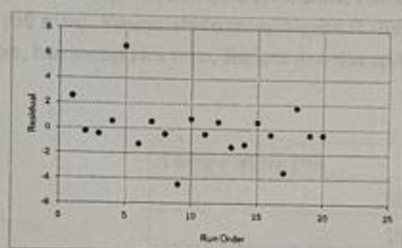
$$\epsilon_{ij} = y_{ij} - \bar{y}_i = 5.0 - 9.4 = -4.4$$

or

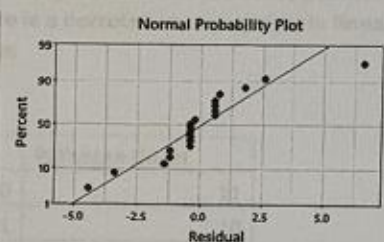
10.5 or -5.5

k. (8 points) If these are the residual plots:

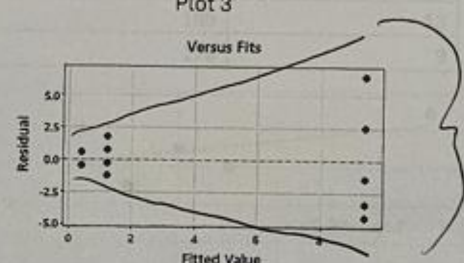
Plot 1



Plot 2



Plot 3



violation!

a) State what assumptions are being checked for each plot

Plot 1: errors and observations are independently distributed

Plot 2: errors and observations are normally distributed

Plot 3: each treatment has the same variance (equal variance assumption)

b) State if any assumptions are violated and if your conclusions are still valid

The equal variance assumption is violated (Plot 3)

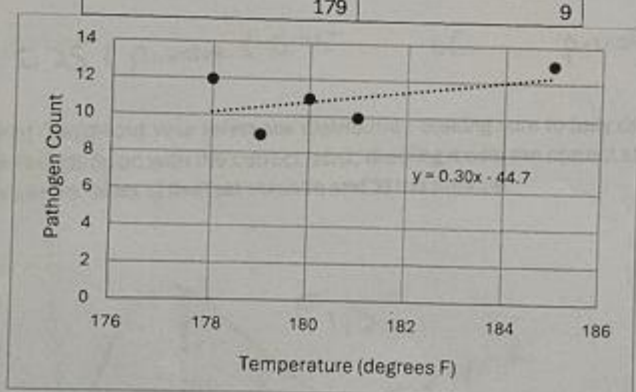
Our conclusions are not valid

c) Describe what you should do if the assumptions are violated

We should transform the data and re-run ANOVA, checking assumptions

3) (25 points) Your team has some past data of colony count (response variable, y) and pasteurization temperature (independent variable, x) all taken for the same pasteurization duration (50 min). You wish to investigate if there is a correlation, using simple linear regression, between the two. Below are the data:

Temperature (°F)	Pathogen Count
180	11
181	10
178	12
185	13
179	9



You have already performed least squares regression to find that the slope (B1) is 0.30, and the intercept is (B0) is 44.7.

a) (3 points) Write out the null and alternative hypothesis needed to test if the linear model is significant

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

b) (5 points) Write the formula(s) needed to calculate the test statistic, and calculate its value given that SST is 10.0 and SSE is 7.2 (Report test statistic to 3 sig figs)

$$F_0 = \frac{SSR/1}{SSE/n-2}$$

$$SS_T = SS_R + SS_E \Rightarrow SS_R = SS_T - SS_E$$

$$SS_R = 10.0 - 7.2$$

$$SS_R = 2.80$$

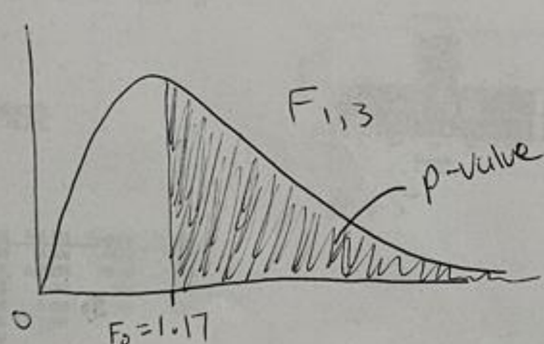
$$F_0 = \frac{2.80/1}{7.2/(5-2)} = \boxed{1.17}$$

c) (3 points) State your rejection criteria using the p-value method and find the range of the p-value using the tables in the back of your book.

Reject if p-value $< \alpha = 0.05$

$0.25 <$ p-value < 0.75 or p-value > 0.25

d) (7 points) Sketch out your reference distribution making sure to fully define the distribution you are sketching on with the correct label, drawing it with the correct shape and where zero is. In addition, label 1) the test statistic and 2) the p-value.



e) (3 points) State your conclusions in the context of the problem

Fail to reject Ho.
we do not have enough evidence to
claim there is a significant liner relationship
between temp and Pathogen count

f) (4 points) Write the formula for and calculate the R^2 value. Describe in words what the R^2 value represents, and why R^2 shouldn't be used to claim if a model is significant.

$$R^2 = \frac{SS_R}{SS_T} = \frac{2.80}{10.0} = 0.28$$

28% of the error can be accounted for by this model.
we can't use R^2 to claim if the model is significant because
there is no "threshold" value, or in other words it is not a hypothesis
test. we can't quantify our degree of confidence that
what we claim is true

4) **(6 points) You found another set of data which had temperature, time and pathogen count data. You want to be able to generate a model from this data. You decide to run multiple linear regression (simple terms) and get the following output.**

Temperature (°F)	Pasteurization Time (min)	Storage Time (weeks)	Pathogen Count (y)
180	50	2	10
200	40	10	8
220	50	4	5
240	40	6	0
150	30	2	15
130	50	6	22

Regression Equation

Pathogen Count = 39.49 - 0.1876 Temperature (°F) - 0.0975 Length of Time
- 0.261 Storage Time (weeks)

Coefficients

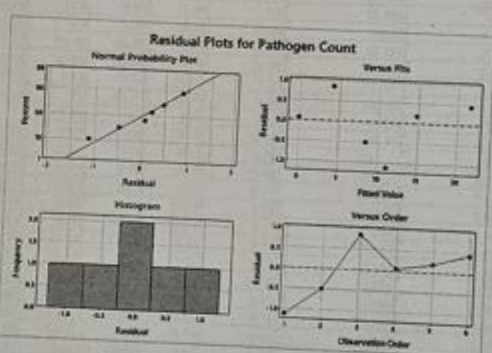
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	39.49	3.43	11.50	0.007	
Temperature (°F)	-0.1876	0.0123	-15.21	0.004	1.07
Length of Time	0.0975	0.0611	1.60	0.251	1.00
Storage Time (weeks)	0.261	0.170	1.54	0.264	1.07

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.11400	99.17%	97.92%	83.98%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	295.518	98.506	79.38	0.012
Temperature (°F)	1	287.275	287.275	231.49	0.004
Length of Time	1	3.165	3.165	2.55	0.251
Storage Time (weeks)	1	2.938	2.938	2.37	0.264
Error	2	2.482	1.241		
Total	5	298.000			



$\alpha = 0.05$

a) (2 points) Can you report a final model? Why or why not?

Not yet; some of the terms are not signifigant

b) (2 points) If you can report a final model report what it is here, if you can't report a final model, what are your next steps *specifically* for this data set to get to a final model?

We need to remove the insignifigant terms and re-run the new model.

c) (2 points) Let's pretend that in problem 3 you didn't find a significant relationship between temperature and pathogen count, but in problem 4 (this problem) you did. Why might this occur?

The temp range is much different is problem 3 vs. 4

Models are only valid in the range tested!

5) (11 points total) Your company wants to establish an Xbar and R control chart for the pasteurization process after some changes were made to reduce variability. There are 25 preliminary samples, each of size 10.

Sample #	Repeat 1	Repeat 2	Repeat 3	Repeat 4	Repeat 5	Repeat 6	Repeat 7	Repeat 8	Repeat 9	Repeat 10	Average (Xbar)	Range (R)
1	10.87	9.23	10.15	10.51	11.42	8.6	10.79	9.31	11.04	9.97	10.189	2.82
2	12.48	8.95	10.01	11.47	9.19	9.57	9.46	9.62	10.97	8.91	10.063	3.57
3	10.04	8.32	10.35	9.99	10.87	10.66	10.53	11.49	10.2	10.42	10.287	3.17
4	11.35	8.66	9.98	9.75	11.21	8.5	10.01	9.87	11.37	9.91	10.061	2.87
5	9.46	8.89	9.73	10.76	9.91	11.18	9.83	10.28	9.62	10.47	10.013	2.29
6	11.05	10.12	10.25	11.33	9.82	10.08	11.5	9.87	11.04	9.6	10.466	
7	9.85	10.28	10.72	10.56	10.65	9.51	9.4	9.99	10.28	10.06	10.13	1.32
8	11.24	8.89	10.81	9.58	9.81	9.77	10.49	10.01	9.85	11.21	10.166	2.35
9	9.33	9.84	9.74	10.34	10.13	9.59	10.94	10.66	9.71	10.09	10.037	1.61
10	11.05	10.99	9.81	10.66	9.96	10.15	9.53	9.74	10.29	9.79	10.197	1.52
11	10.12	9.47	9.94	9.95	10.66	9.53	11	10.47	10.29	9.56	10.099	1.53
12	9.89	10.29	9.77	10.35	10.66	10	9.56	9.7	9.72	9.92	9.986	1.1
13	11.2	10.06	11.12	9.93	9.65	10.28	9.83	10.08	9.79	10.56	10.25	1.55
14	9.97	9.75	11.02	10.56	10.33	10.42	9.86	9.51	9.83	9.88	10.113	1.51
15	11.05	10.88	9.46	9.95	10.14	10.16	9.79	9.82	10.23	10.06	10.154	1.59
16	10.21	9.95	10.84	10.52	9.75	9.67	9.92	9.83	9.65	9.97	10.031	1.19
17	11.42	11.03	10.47	9.91	9.6	10.48	9.73	9.91	10.56	9.91	10.302	1.82
18	9.82	9.71	9.5	9.58	10.37	9.83	9.77	10.27	10.14	9.71	9.87	0.87
19	10.56	10.49	10.29	9.72	10.36	10.16	10.09	10.22	9.95	9.73	10.157	0.84
20	10.66	9.52	10.84	10.17	10.33	10.74	9.78	9.64	9.62	9.78	10.108	1.32
21	9.74	9.86	9.76	10.13	9.74	10.18	10.06	10.36	9.94	10.49	10.026	0.75
22	10.25	9.84	9.78	9.63	9.96	9.82	10.08	9.92	10.46	9.92	9.966	0.83
23	9.99	9.97	9.58	9.79	10.11	9.67	10.43	10.06	10.21	9.71	9.952	0.85
24	10.52	9.84	9.99	10.07	9.62	9.91	10.17	9.94	9.75	9.63	9.944	0.9
25	9.93	10.47	9.91	10.01	10.44	10.21	9.93	10.47	9.61	9.85	10.083	0.86
											252.65	40.93
											Sum of all Xbars	Sum of all Rs

a. (2 points) Show how you would get the range for sample 6

R = Largest value − Smallest value = $11.5 - 9.6 = \boxed{1.90}$

b. (2 points) If you wanted to compare the variance between the process before the changes, to the variance after the changes, what is the name of the statistical test you would recommend?

2-variance test

c. (7 points) Write out the general formulas for and calculate the centerline as well as the three sigma control limits that should be used on the Xbar and R control charts. $m=25 \quad n=10$

R-Chart: $\bar{R} = \dfrac{\text{Sum of all R's}}{\text{\# of samples}} = \dfrac{40.93}{25} = 1.637$

$UCL = D_4 \bar{R} = (1.777)(1.6372) = \boxed{2.91}$ (2 decimals) line data

centerline: $R = 1.631 = \boxed{1.63}$

$LCL = D_3 \bar{R} = (0.223)(1.6372) = \boxed{0.37}$

X-bar Chart: $\bar{\bar{x}} = \dfrac{\bar{x}\text{'s (all added)}}{m} = \dfrac{252.65}{25}$

$UCL = \bar{\bar{x}} + A_2\bar{R} = 10.106 + (0.308)(1.6372) = \boxed{10.61}$ 10.106

centerline: $\bar{\bar{x}} = \boxed{10.11}$

$LCL = \bar{\bar{x}} - A_2\bar{R} = 10.106 - (0.308)(1.6372) = \boxed{9.60}$