STATISTICS 642 - FINAL EXAM

Tuesday May 5, 2020

Student's Name _____

Student's Email Address _____

INSTRUCTIONS :

(1) The exam consists of 9 pages of questions including the title page and 21 pages of STAT 642 Tables.

(2) You have exactly **2 hours** to complete the exam and 20 minutes for downloading exam and uploading solutions.

(3) **Exam Period:** A period of 2 hours and 20 minutes Starting at 1:00 p.m. CDT May 5 and Concluding at 3:20 p.m. CDT May 5.

(4) You MUST upload your solutions to the exam by 3:20 p.m. CDT Tuesday May 5.

(5) The Exam will be Zoom proctored. You will have two options for composing your solutions:

- You can print out the exam, write out your solutions on the exam booklet, then upload pages 1-6 of the exam booklet to eCampus.
- Alternatively, you can just write your solutions on blank sheets of paper, scan it into a single pdf file, then upload the pdf file to eCampus just as you would with homework solutions.

- See General Information in our eCampus site for methods for scanning.

(6) Include with your uploaded solutions your signature and email address.

(7) **Make sure your solutions are readable. Past submissions have not been dark enough and hence very difficult to read.**

(8) Do not discuss or provide information to anyone concerning the questions on this exam or your solutions until I post the solutions to the exam.

(9) You may use the following:

- Calculator
- Summary Sheets - **8-pages,** 8.5"**x**11"**, (you may write/type/cut-paste anything on both sides of the 8 sheets**)
- The STAT 642 Exam Tables

(10) Do not use any other written material except for your summary sheets and STAT 642 Exam Tables. Do not communicate with anyone during the exam. The solutions must be just your own work.

I attest that I spent no more than 2 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature_____

Problem I. (25 points) For the following experiment, provide the requested information:

The EPA designed a study to investigate the impact of excess nitrogen and air pollution on plant growth in wetland areas. The researchers selected four levels of nitrogen (N_1-N_4), three types of particulates (P_1-P_3) found in air pollution, and two exposure times (E_1, E_2). The experiment was conducted in an artificial setting within a greenhouse. Eight trays were used in the study with each tray holding three artificial wetlands. All of the artificial wetlands receive a standard set of seeds to start growth. Four of the trays are placed on a table located at the north end of the greenhouse and the other four trays are placed on a table located at the south end of the greenhouse. Separately for each table, the four trays are randomly assigned to the four levels of nitrogen. Within each tray, the three wetlands are randomly assigned to the three types of particulates. Each wetland is then split in half with one half randomly assigned to be low exposure time, E1; the other half to the high exposure time, E2. At the end of 8 months, the researcher determined the amount of biomass in each wetland. The weight (kg) of biomass are given here with notation: Tray (TR), Nitrogen (N), Particulate (P), and Exposure (E).

Table 1								Table 2							
		P3		P2		P1				P2		P1		P3	
TR	N	E1	E2	E2	E1	E1	E2	TR	N	E2	E1	E1	E2	E2	E1
TR1	N3	87.2	88.8	70.4	75.7	75.9	80.6	TR5	N2	78.2	80.5	65.1	68.3	65.3	66.6
TR2	N1	87.2	88.8	70.4	75.7	75.9	80.6	TR6	N4	79.8	85.2	57.6	61.4	58.5	61.6
TR3	N4	87.2	88.8	70.4	75.7	75.9	80.6	TR7	N1	82.4	83.1	50.5	54.0	51.6	54.7
TR4	N2	87.2	88.8	70.4	75.7	75.9	80.6	TR8	N3	75.5	78.7	39.0	43.9	41.9	45.1

1. Type of Randomization: CRD, RCBD, LSD, Split-Plot, Crossover, etc.

2. Type of Treatment Structure: single factor, crossed, nested, fractional, etc.

3. Identify each of the Factors as being Fixed or Random:

4. Describe the Experimental Units and Measurement Units:

5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures

Problem II (30 points.) A structural engineer is studying the strength of aluminum alloy purchased from the three largest vendors. Each vendor submits the alloy in standard-sized bars, either 1, 2, or 3 inches. The processing of different sizes of bar stock from a common ingot involves different forging techniques, and this factor may be important. Furthermore, the bar stock is forged from ingots made in different batches. Each vendor submits two test specimens of each size bar stock from 3 batches. The strength of each bar is determined and is reported in the following table. The three vendors are the only vendors under consideration, the batches are randomly selected from the Vendors production output.

		Vendor								
		I			II			III		
Batch		1	2	3	4	5	6	7	8	9
Bar Size	1 Inch	12.30	13.46	12.35	13.01	13.46	13.15	12.47	12.75	13.24
		12.59	14.00	12.06	12.63	13.92	13.20	12.96	12.68	13.15
	2 Inches	13.16	13.29	12.50	12.74	13.84	13.46	12.73	12.60	13.92
		13.00	13.62	12.39	12.68	13.75	13.57	12.64	12.65	13.64
	3 Inches	12.87	13.46	12.73	12.47	13.62	13.36	13.01	12.80	13.19
		12.92	13.82	12.15	12.15	13.28	13.42	12.62	12.71	13.23

1. Complete the following AOV table. Note: The Mean Squares (MS) are provided in the table not the Sum of Squares and the following notation is used: S=Bar Size, V=Vendor, B=Batch.

SOURCE	DF	MS	EMS
S		0.1263	
V		0.4424	
S*V		0.0594	
B(V)		1.6702	
S*B(V)		0.0919	
ERROR		0.0404	

2. At the $\alpha = 0.05$ level, evaluate the effect of Vendor and Bar Size on the strength of the aluminum alloy.

3. Using the numeric values of the MS's given above and your EMS's, find the values of the following quantities where y_{ijkl} is the strength of bar l of Size j made using material from Batch k of Vendor i's aluminum:

$$y_{ijkl} = \mu + \tau_i + a_{k(i)} + \gamma_j + (\tau\gamma)_{ij} + b_{jk(i)} + e_{ijkl} \text{ with } i = 1, 2, 3;\ j = 1, 2, 3;\ k = 1, 2, 3;\ l = 1, 2$$

Estimate the standard error of the estimated difference in the mean strength of two Vendors:

Problem III. (9 points) A process engineer is designing a study to investigate how to decrease the overhead in the manufacture of a new product. Six factors A, B, C, D, E, and F were identified for investigation. It was decided to use 2 levels (L or H) of each of these factors. The study design was a fractional factorial with 16 runs using the generators

$I_1 = ABDE = +$ and $I_2 = ACE = +$ to generate the 16 treatments to be used in the study.

1. For each of the following treatments, check YES if the treatment will appear in the experiment, otherwise check NO.

   i. $(A, B, C, D, E, F) = (H, L, H, H, L, H)$ _____YES _____NO

   ii. $(A, B, C, D, E, F) = (H, L, H, L, H, H)$ _____YES _____NO

2. What is the resolution of this design? Justify your answer.

3. What effects which must be assumed to be negligible in order that the data from the experiment will provide an estimate of the interaction between Factors C and F.

Problem IV. **(36 points) Place your answer to each of the MULTIPLE CHOICE questions in the space provided. Make sure to use UPPER case letters: A, B, C, D, E and select ONLY one answer for each question.**

Name _____

(1.)_____

(2.)_____

(3.)_____

(4.)_____

(5.)_____

(6.)_____

(7.)_____

(8.)_____

(9.)_____

(10.)_____

(11.)_____

(12.)_____

_____(1). Consider a design with equal replication of t treatments in a completely randomized experiment. An examination of the residuals yielded a p-value = .3476 from the Shapiro-Wilks test. What can you conclude about the model conditions?

A. constant variance of the residuals does not appear to be violated

B. constant variance of the residuals appears to be violated

C. normal distribution of the residuals does not appear to be violated

D. normal distribution of the residuals appears to be violated

E. the p-value is not a good indicator of whether model conditions are valid

_____(2). A RCBD with three factors: F_1-fixed, F_2-random, F_3-fixed, was conducted. The experimenter obtained the following results from the AOV F-tests: $F_1 * F_2 * F_3$ is not significant, $F_1 * F_2$- significant, $F_1 * F_3$-not significant, $F_2 * F_3$-significant, and F_1, F_2, F_3 are all significant. She then decides to determine if there are pairwise differences in the levels of F_1. Which of the following would be the most appropriate approach to answering her questions.

A. Tukey's HSD applied to the levels of F_1 separately at all combinations of (F_2, F_3).

B. Tukey's HSD applied to the levels of F_1 averaged over all combinations of (F_2, F_3).

C. Tukey's HSD applied to the levels of F_1 separately at the levels of F_2, averaged over the levels of F_3.

D. Tukey's HSD applied to the levels of F_1 separately at the levels of F_3, averaged over the levels of F_2.

E. None of the above would be appropriate because the design was not a CRD

_____(3). A covariate was measured along with the responses in a completely randomized design with a single factor having 5 levels. The p-value from the AOV reveals significant evidence that the slopes of the 5 treatment lines are different. A comparison of the 5 treatments

A. cannot be conducted because there is an interaction between the covariate and treatment which violates the conditions for analysis of covariance.

B. could be made using Tukey's HSD on the sample treatment means.

C. could be made using Tukey's HSD on the adjusted treatment means.

D. could be made using Tukey's HSD on the sample treatment means at specified values of the covariate.

E. could be made using Tukey's HSD on the adjusted treatment means at specified values of the covariate.

_____(4). An entomologist designs an experiment to evaluate the effectiveness of five Dose levels of a pesticide to control fire ants. She randomly selects 100 1-acre plots of land and randomly assigns 20 plots to each dose level. Next, she randomly selects 15 fire ant hills in each plot and records the weight, W, of fire ants killed after two weeks of treatment. The scientist runs the following code in SAS to analyze her data:

```
PROC GLM;

CLASS DOSE PLOT;

MODEL W = DOSE PLOT(DOSE);

RANDOM PLOT(DOSE)/TEST;

LSMEANS DOSE/PDIFF ADJUST=TUKEY;
```

She then uses the output from LSMEANS to group the five Doses according to the mean weight of fire ants killed. The conclusions reached using the SAS output will be incorrect because

A. the LS means are biased due to using the wrong df for the error term.

B. the calculation of $\widehat{SE}(\hat{\mu}_i)$ is incorrect, SAS only considers $\hat{\sigma}^2_{PLOT(DOSE)}$ and not $\hat{\sigma}^2_e$ in the calculation.

C. the calculation of $\widehat{SE}(\hat{\mu}_i)$ is incorrect, SAS only considers $\hat{\sigma}^2_e$ and not $\hat{\sigma}^2_{PLOT(DOSE)}$ in the calculation.

D. multiple comparisons should not be made because the PLOT factor has random levels.

E. There is no problem in the analysis because a Tukey adjustment was made to the p-values.

_____(5). The response variable y_{ij} is observed in a RCBD experiment with five treatments and ten blocks. The researcher fits the model and conducts a residual analysis. The normal probability plot indicates a very right skewed distribution for the residuals. The transformation $y^*_{ij} = 1/\sqrt{y_{ij}}$ is suggested by Box-Cox procedure. What would be the most easily interpreted analysis of this data?

A. Run an AOV on the original data. With 50 data values, the central limit theorem applies.

B. Run the Friedman test even though the residuals are very right skewed.

C. Run an AOV on the transformed data.

D. The analyses given in A., B., and C. are all appropriate analyses

E. The analyses given in A., B., and C. are all inappropriate analyses

_____(6). An experiment was conducted as a randomized complete block (RCBD) with 10 blocks and a factorial treatment structure, Factor A with 3 levels and Factor B with 4 levels. The homogeneity of variances condition can be evaluated by

A. applying the B-F-L test to the 12 treatment variances because the data is from 12 populations.

B. a box plot of the 120 residuals.

C. it is not necessary to have equal variances in a RCBD because the heterogeneity in the EU's has been controlled by the blocking.

D. it is not necessary to have equal variances because, with 120 observations, the central limit theorem eliminates the need for equal variances.

E. none of the above

_____(7). A veterinarian wants to investigate $t = 10$ treatments for controlling heartworms in puppies. Her consulting statistician determines that she will need $r = 9$ replications per treatment. There is enormous variation in the effectiveness of the treatment so the veterinarian wants to use groups of homogeneous puppies and decides to use litters of puppies as her blocking variable. Most litters contain fewer than 10 puppies so she decides to use a BIBD, with at most 9 puppies per litter. Which of the following combinations of b litters and k puppies per litter would yield the most effective design?

A. $b = 10$ and $k = 9$

B. $b = 18$ and $k = 5$

C. $b = 30$ and $k = 3$

D. $b = 45$ and $k = 2$

E. All of the above are equally effective because $\lambda = k - 1$

_____(8). An experiment was conducted as a CRD with 6 reps of the 4 levels of a single factor F_1. There were measurements taken on each of the 24 experimental units (EU) at time points t_1, t_2, t_3, t_4, t_5. The analysis of the data was conducted as a CRD split-plot design with F_1 as the whole plot treatment and Time of measurement as the split-plot treatment. The p-values from the F-tests for the main effect of Time and the interaction between Time and F_1 are only approximations to the true p-values because

A. the correlation between the 5 measurements on each EU may not satisfy compound symmetry.

B. the order in which the 5 measurements are taken on each EU was not randomized.

C. the 5 measurements on each EU are subsamples and hence not a true factor.

D. the 5 measurements on each EU are not independent.

E. the 5 measurements on each EU should be modeled as the levels of a random effect.

_____(9). In a RCBD with 5 blocks and a factorial treatment structure consisting of a qualitative factor F_1 at 2 levels and a quantitative factor F_2 at 4 levels, the researcher wants to know if the linear trend in the mean responses across the levels of F_2 are different for the two levels of F_1. Which of the following contrasts would address this question?

A. $L = \mu_{11} + \mu_{12} + \mu_{13} + \mu_{14} - \mu_{21} - \mu_{22} - \mu_{23} - \mu_{24}$

B. $L = -3\mu_{11} - \mu_{12} + \mu_{13} + 3\mu_{14} - 3\mu_{21} - \mu_{22} + \mu_{23} + 3\mu_{24}$

C. $L = -3\mu_{11} - \mu_{12} + \mu_{13} + 3\mu_{14} + 3\mu_{21} + \mu_{22} - \mu_{23} - 3\mu_{24}$

D. $L = \mu_{11} - 3\mu_{12} + 3\mu_{13} - \mu_{14} - \mu_{21} + 3\mu_{22} - 3\mu_{23} + \mu_{24}$

E. none of the above because the design was not a CRD

_____(10). A RCBD experiment was run with 3 blocks, 25 EU's were randomly assigned to each of the 6 levels of a treatment factor, and each EU was measured at 5 specified locations on the EU.

Let $y_{ijk\ell}$ be the measurement from ℓth EU receiving treatment j in block i at location k on the EU.

$y_{ijkl} = \mu + b_i + \tau_j + d_{j\ell} + \gamma_k + (\tau\gamma)_{jk} + e_{ijkl}$, with $i = 1, \ldots, 3$; $j = 1, \ldots, 6$; $k = 1, \ldots, 5$; $l = 1, \ldots, 25$;

where μ, τ_j γ_k, and $(\tau\gamma)_{jk}$ are fixed population parameters and b_i, $d_{j\ell}$, and $e_{ijk\ell}$ are random variables with $N(0, \sigma_b^2)$, $N(0, \sigma_d^2)$, and $N(0, \sigma_e^2)$ distributions, respectively. Which one of the follow statements best describes the correlation structure of the random variables in the model?

A. b_i's are correlated, $d_{j\ell}$'s are correlated, and $e_{ijk\ell}$'s are correlated

B. b_i's are independent, $d_{j\ell}$'s are independent, and $e_{ijk\ell}$'s are correlated

C. b_i's are independent, $d_{j\ell}$'s are correlated, and $e_{ijk\ell}$'s are correlated

D. b_i's are correlated, $d_{j\ell}$'s are independent, and $e_{ijk\ell}$'s are correlated

E. b_i's are independent, $d_{j\ell}$'s are independent, and $e_{ijk\ell}$'s are independent

_____(11). In a crossover design evaluating 5 treatments, there may be a strong positive correlation between the 5 responses from the same experimental unit. You recalled that the actual power of the AOV F-test will now be greater than the power of the AOV F-test when the observations are independent. However, the negative impact of positive correlation on the F-test is

A. the probability of a Type I will be higher than expected under no correlation

B. the probability of a Type I will be less than expected under no correlation

C. the probability of a Type II will be higher than expected under no correlation

D. the probability of a Type II will be less than expected under no correlation

E. The power of the F-test will in fact be decreased because the assumptions of the AOV F-test have been violated

_____(12). A study of the interaction between two factors, Factor A with 4 fixed levels and Factor B with 2 fixed levels, was designed with 5 experimental units randomly assigned to each of the treatments. The FDA requires a power of at least .90 for an α=.01 test whenever there is one or more pairs of treatments having a difference of at least 6 units in their mean responses. From previous studies, it was determined that the variation in responses is approximately $\sigma_e = 1.86$. Which one of the following is the closest approximation to the power of the test? Show your calculations.

A. Power ≤ 0.30.

B. $0.30 <$ Power ≤ 0.50

C. $0.50 <$ Power ≤ 0.70

D. $0.70 <$ Power ≤ 0.90

E. Power > 0.90