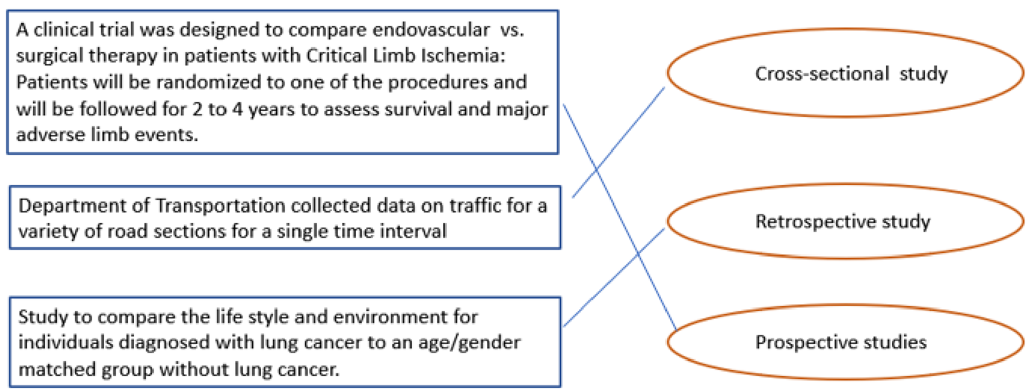
STAT3119 Midterm Exam Date: 10/24/2019

Name:__________________________________ GWID: _______________________________

1. What is the main difference between an experimental study and an observational study if we want to study the effect of various treatment levels on a certain outcome?
   5 pts treatment are randomly assigned or use of randomization for experimental study to suggest a cause/effect relationship; observational studies → association or correlation

2. Please match each specific study (from the left) to its corresponding type/category (on the right side) by drawing a line between them. 5 pts



3. In general, when you design a scientific study to investigate certain factors (treatments) on an outcome, you need to first set the sample size for your study.

   (1) The required component for the sample size calculation is: (circle only one answer) 2pts

   A. The desirable power and type I error.
   B. The variability (standard deviation) of the observations.
   C. Expected effect size (the difference that you expect to find between the treatments)
   D. All the above.

   (2) You have a limited time or fund to finish the study, therefore the power that you choose to plan you study should be: (circle only one answer) 2pts

   A. > 95% B. 80-90% C. 60-70% D. 50-60%

   (3) In this study, assume $\Delta = \max(\mu_i) - \min(\mu_i)$ is the minimum range of treatment level means for which it is important to detect. Choose the **FALSE** statement from below (circle only one answer).

   3 pts

   A. As the target effect Δ increases, you need a larger sample size.
   B. As the measurement variability σ increases, you need a larger sample size.
   C. As the required power increases, you need a larger sample size.
   D. As the required estimation precision increases, you need a larger sample size.

4. It is important to check the model assumptions first before we apply any statistical model to analyze a dataset.

   (1) Please state the two assumptions for the ANOVA model. 5 pts

   a) Constancy of variance ; b) iid normal errors

   b) Please explain how you can examine whether these assumptions are appropriate. 5 pts

   Residual analysis or statistical tests

5. Find a FALSE statement. (circle only one answer) 4 pts

   A. When we do multiple comparisons, the Type I error is inflated.
   B. Hartley test or Brown-Forsythe test is used for testing the constancy of error variance.
   C. In a one-way ANOVA, the *F*-test as the ratio of mean squares can be used to test whether the residuals are normally distributed.
   D. In a two-way ANOVA analysis with replications per treatment, the interaction effects should be tested first before we estimate the main effects of each factor separately.

6. (One-factor study) In a study of the effectiveness of different rust inhibitors, four brands (A, B, C, D) were tested. Altogether, 40 experimental units were randomly assigned to the four brands, with 10 units assigned to each brand. The higher the response value, the more effective is the rust inhibitor.

   (1) If the one-way ANOVA assumption is appropriate, please write down the two formulations for the ANOVA model: (3 pts)

   Cell means model ____ $Y_{ij} = \mu_i + \epsilon_{ij}$ ____

   Factor effects model ____ $Y_{ij} = \mu_. + \tau_i + \epsilon_{ij}$ ____

   (2) Based on the 40 observations, we can generate the following ANOVA table, please fill in the missing SS , df and MS and test statistic. (5 pts)

Source of variation	SS (sum of squares)	df	MS (mean squares)	Test statistic (F^*)
Between brands	3000	3	1000	100
Error	360	36	10	
Total	3360	39		

   c) (2 pts) State the null and alternative hypothesis in the ANOVA analysis

   H_0: $\mu1= \mu2= \mu3=\mu4$;

   H_a: Not all μ_i are equal.

   d) (2 pts) Under null H_0, what distribution the test-statistic F^* is distributed as ?
   __F(3,36)______

   e) (2 pts) The p-value of the test < 0.0001, what will be the conclusion that researcher can state based on these results? Mean response were significant differ among 4 brands

f) (2 pts) What is the estimate of the variance σ^2 _=MSE=10_________ ?

g)

Define a difference between two factor level means as $D = \mu_i - \mu_{i'}$, we can get the estimator of D from sample means $\hat{D} = \bar{Y}_{i.} - \bar{Y}_{i'.}$. What is the estimated variance of $\hat{D}$? i.e. $s^2\{\hat{D}\}$ =____MSE(1/10+1/10)=2

(2 pts)

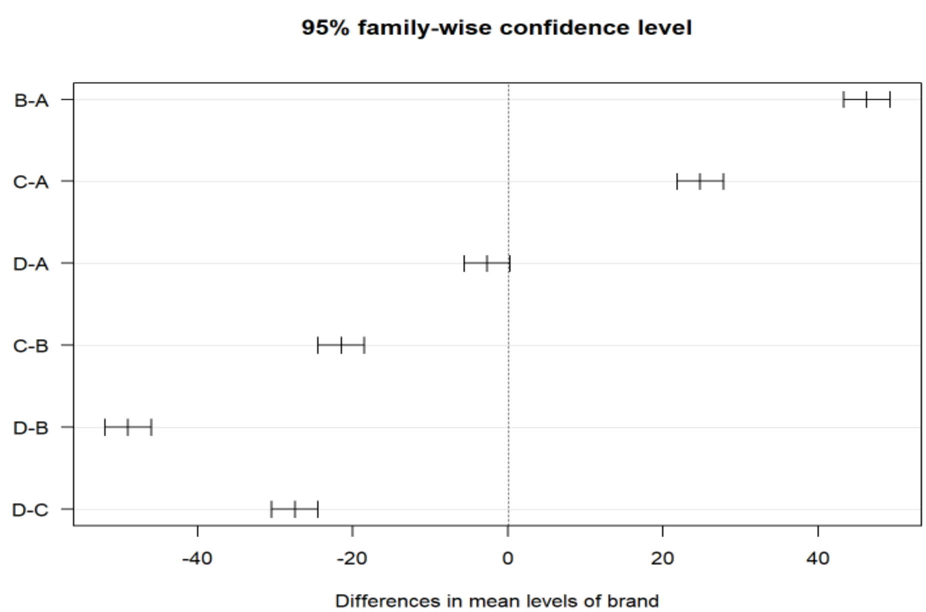
h) (2 pts) The inference of D is based on the a test statistic $\frac{\hat{D} - D}{s\{\hat{D}\}}$, which distribution does it follows? (r= number of treatment levels. n_T is the sample size)

a. Normal distrubition : $N(0, 1)$
b. F-distribution: $F(r-1, n_T-r)$
c. t-distributon : $t(n_T-r)$
d. studentized range distribution $q(r, n_T-r)$.

(9) (2 pts) If all parwise comparisons are of interest and we are controlling the overall type I error, which procedure is superior?

a. Fisher's LSD procedure
b. Bonferroni procedure
c. Scheffé procedure
d. Tukey HSD procedure

10) (3 pts) Finally, below a pairwise comparison plot for the four brands based on the Tukey's procedures:



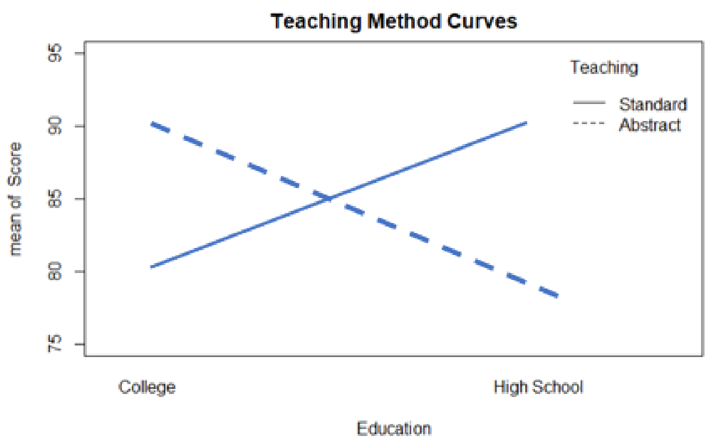
What can you conclude from these results about the differences among the four brands? [e.g. which brands(s) is most effective? Which brands differ significantly?]

B>A, C>A, B>C, B>D, C>D➔ a) brand B is best; b) brand C is the second best and better than A and D. c) A and D are not significantly different .

7. A education researcher conducted a study to examine the effects of teaching method (standard or abstract) and student's education level (high school or college) on learning of statistics. Some learning scores were collected from n= 20 students at each teaching method or education level. The summary data was given in the following table: (12 pts)

Mean learning scores			
	Student education level (*j*)		
Teaching method (*i*)	College	High school	*Row means*
Abstract	90	80	85
Standard	80	90	85
Column means	85	85	85

(1) Please draw an treatment mean plot below. 3 pts



(2) The sample means for the two education levels are the same and the sample means for the two teaching methods are the same in the Table. Does this mean there are no main effects (treatment difference) for either two teaching methods or the two education levels on the learning scores? Please explain using the summary data. 3pts

No, there are main effect. If there is interaction , the effect of one levels depends on the other and we can't look the two main effect separately.

(3) Is there the interaction effect between the teaching method and student's education level? Please explain using the summary data. 3 pts

Yes. can be seem from the the plot or data table.

(4) What would be best ANOVA model to study both factors and their interaction? Please write the corresponding factor effects model 3pts

$$Y_{ijk} = \mu.. + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

8. (Balanced two factor studies): The Castle Bakery Company supplies wrapped Italian bread to a large number of supermarkets in a metropolitan area. An experimental study was made of the effects of height of the shelf display (factor A: bottom, middle, top) and the width of the shelf display (factor B: regular, wide) on sales of this bakery's bread during the experimental period (Y, measured in cases). Twelve supermarkets, similar in terms of sales volume and clientele, were utilized in the study. The six treatments were assigned at random to two stores each according to a completely randomized design, and the display of the bread in each store followed the treatment specifications for that store. Sales of the bread were recorded below: 9 pts

	Factor B (display width)	
Factor A (display height)	B1 (regular)	B2 (wide)
A1 (bottom)	47 43	46 40
A2 (middle)	62 68	67 71
A3 (top)	41 39	42 46

The researcher found the two factor ANOVA analysis is appropriate. He calculated the following ANOVA table. The significant level is 0.05.

Analysis of Variance for Cases Sold

Source	DF	SS	MS	F	P
Height	2	1544.00	772.00	74.71	0.000
Width	1	12.00	12.00	1.16	0.323
Height*Width	2	24.00	12.00	1.16	0.375
Error	6	62.00	10.33		
Total	11	1642.00			

(1) Based on this table, what can you conclude about the interaction effects between the two factors? Use the data from the ANOVA table to explain. 3pt

Based on p-value p=0.375 , no interaction

(2) what can you conclude about the main effect of factor A (display height)? Use the data from the ANOVA table to explain. 3pt

Based on p-value p=0, significant effect for factor A

(3) what can you conclude about the main effect of factor B (display width)? Use the data from the ANOVA table to explain. 3 pt

Based on p-value p=0.323 , no effect for factor B

9. (A two-factor study with one case per treatment) An analyst in an insurance office wishes to evaluate the effects of city size and geographic region on the amount of the premium. The six cities were selected to represent different regions of the state and different sizes of cities. The following table shows the amounts of three-month premiums for a specific type and amount of coverage in a given risk category for each of the six cities, classified by size of city (factor A) and geographic region (factor B). 10 pts

(a) Premiums for Automobile Insurance Policy (in dollars)

Size of City (factor A)	Region (factor B) East ($j = 1$)	West ($j = 2$)	Average
Small ($i = 1$)	140	100	120
Medium ($i = 2$)	210	180	195
Large ($i = 3$)	220	200	210
Average	190	160	175

(1) We can test whether the 2 factors (A and B) interact by using: (circle only one answer) 2pt

a. the ANOVA table to construct the F-test based on mean squares MSAB and MSE.

b. the Kruskal- Wallis test

c. the Brown-Forsythe test

d. the Tukey test for additivity.

(2) If the test suggests the factors do not interact and the assumptions seem appropriate, we can run an ANOVA analysis. Please fill in the missing df and MS in the table below: 4 pt

Source of variation	SS	df	MS
Size of city (A)	9300	2	4650
Region (B)	1350	1	1350
Error	100	2	50
Total	10750	5	

(3) Based on the ANOVA analysis, we conclude that both factors had significant effects. Since the factor effects are additive, we can estimate the treatment means μ_{ij} . Please provide the best estimates for the following treatment means using the data table: 4 pt

$\hat{\mu}_{11}$(Small size, East region) 190+120-175 = 135

$\hat{\mu}_{32}$(Large size, West region) 160+210-175 = 195

10. In a randomized complete block design, what is the main objective to group the experiment units into blocks before randomizing the treatments within blocks ? (10 pts)

1) have more homogenous groups to reduce experimental errors /within group differences

2) remove/control confounding variables.