STAT 101B Spring 2020
Midterm Exam

Name: _____KEY_____ SID:_____

- **You are allowed to use your Notes.**
- **You are allowed to use any type of calculator including R as an arithmetic calculator.**
- **Budget you time wisely. You need time to put your answers together and upload it to ccle as pdf format.**

Q1) (10 pts.) **Answer the Following Multiple choice questions (1- 5): (2 points Each)**

1) The objective of a study was to find out whether there was any relationship between level of education (high school, four year college, and graduate work) and attitude toward pre-screening for breast cancer (i.e., going for mammograms)

	sum of square	degrees of freedom	mean square	F	P
between group	1815.068	2	907.534	?	?
within group	51912.079	509	101.988		
total	53727.147	511			

Given the above data, what is the best answer?
Select one:

- a. Reject the null hypothesis and conclude that the higher the level of education the more positive the attitude toward seeking prescreening for breast cancer.
- b. Reject the null and conclude there is a relationship between level of education and seeking prescreening for breast cancer.
- c. Fail to reject the null and conclude that there is no relationship between level of education and attitude toward prescreening for breast cancer.
- d. Fail to reject the null and conclude that high school, four-year college, and graduates have a similar attitude toward prescreening for breast cancer.

2) 138 students were taught mathematics through cooperative learning. They were pre-tested and post-tested on their knowledge of fractions and the following data was reported.
pretest mean = 39.6
posttest mean = 59.41
r (correlation between pre and post data) = 0.46
95% CI = 15.8, 23.72
What is the best answer?
Select one:

- a. The statistical method used was the paired sample test of the mean. We need the p value to decide if the null was rejected.
- b. The one sample test of the mean was used to test the null that gain on fraction is zero and we are 95% confident that the students gained between 16% to 24% on their knowledge of fractions.
- c. The statistical method used to analyze the data was the two sample test of the mean and the null was not rejected.
- d. The statistical method used was the paired sample test of the mean and the null was rejected.

3) A researcher conducts One-Way ANOVA to find out whether there is any relationship between major (engineering, mathematics, statistics, and computer science) and annual income upon graduation from college. She collects this data on a random sample of 400, (100 from each discipline). What does SS_{Within} show?

Select one:

A) the average annual income of each group minus the mean annual income for all subjects squared, added, and multiplied by 100.
B) The annual income of each subject minus the mean of income for his group squared and added.
C) annual income for each subject minus the average income for all 400 subjects squared and added.
D) The variance for all the annual income of all the 400 subjects multiplied by 399.

4) When conducting a one-way ANOVA, the _____ the between-treatment variability is when compared to the within-treatment variability, the _____ the value of F_{DATA} will be tend to be.
a. smaller, larger
b. smaller, smaller
c. larger, larger
d. smaller, more random
e. larger, more random

5) When the *k* population means are truly different from each other, it is likely that the average error deviation:
a. is relatively large compared to the average treatment deviations
b. is relatively small compared to the average treatment deviations
c. is about equal to the average treatment deviation
d. differs significantly between at least two of the populations
e. none of the above

Multiple Choice Answers:

Q1 Part #	**1**	**2**	**3**	**4**	**5**
Answer	**B Full credit , A partial credit**	**D**	**B**	**B or C**	**B**

Q2) Case Study: An article in the *Journal of Testing and Evaluation* (Vol. 16, no.2, pp. 508-515) investigated the effects of cyclic loading "Frequency" and environmental conditions on fatigue crack growth at a constant 22 MPa stress for a particular material. The head of the data from this experiment are shown below (the response is crack growth rate). (40 points)

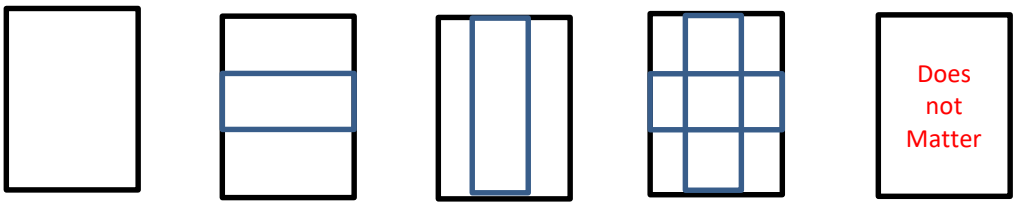
Giving a final answer may not guarantee you a full credit. You need to show your work.

```
> head(Q2Data)
# A tibble: 6 x 3
  Frequency Environment `Crack Growth`
1       0.1 Air                   2.08
2       0.1 Air                   2.81
3       0.1 Air                   2.71
4       0.1 Air                   2.24
5       1   Air                   2.38
6       1   Air                   2.06
> dim(Q2Data)
[1] 36  3
> table(Q2Data$Environment)

     Air      H20 Salt H20
      12       12       12
> table(as.factor(Q2Data$Frequency))

0.1   1  10
 12  12  12
```

A) (I) Create a Factor diagram for this experiment:



A (II) List the sources of variation in this study along with their corresponding labels and their degrees of freedom.

Source of Variation	**d.f.**
Factor 1: Frequency	2
Factor 2: Environment	2
Interaction: Frequency X Environment	4
Residual	27

(B) Suppose you have decided to block by Frequency. Not considering the interaction between the two factors. Which of the two above designs (two factors vs "one factor one blocking factor) have more power ($Fixing\ \alpha$)? Why?

The two factors design has more power than one factor and one blocking factor.
Since more of the SST is explained in the two factor design, we have less SSE which gives a smaller , *so* we get more power.

(C) How large a sample size would be needed for comparing the average crack growth of two environment levels only (Air and H20). Use α = 0.05, to detect a difference between means of 2.5 or more with Power = 0.90, $given\ \sigma = 3.301666$

$$n = \frac{2 * 3.301666^2 (1.96 + 1.28)^2}{2.5^2} = 36.61898 \approx 37\ for\ each\ group.$$

A total of 2*37 = 74 samples. Any number really close is ok.

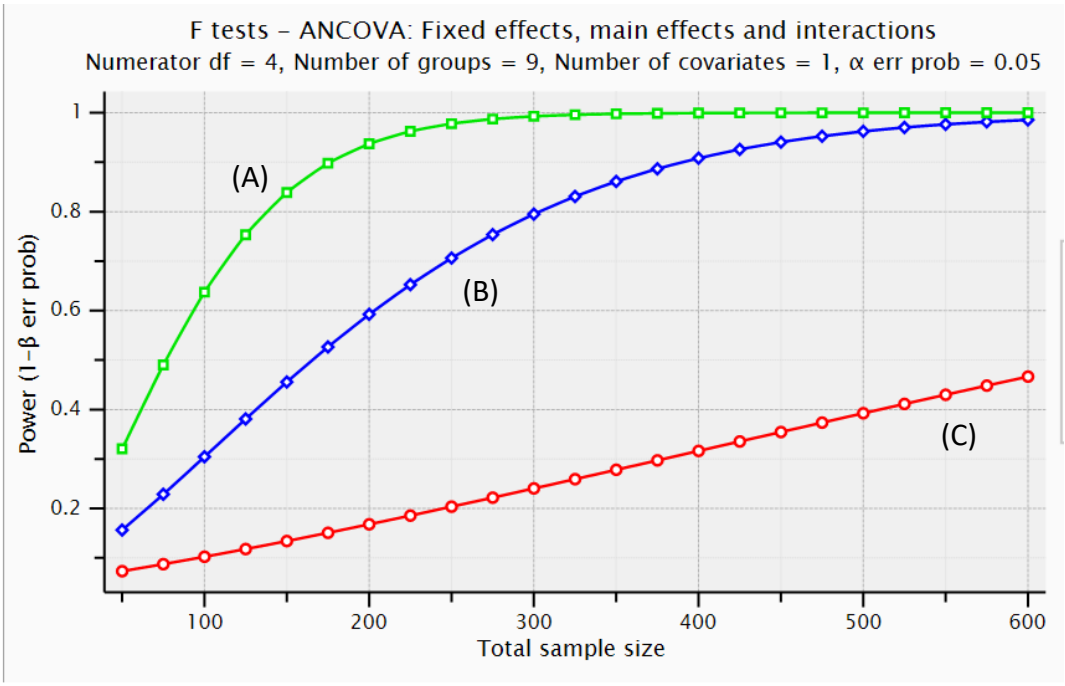
```
> pwr.t.test(d=2.5/3.301666,power=0.90,type="two.sample")

     Two-sample t test power calculation

              n = 37.64002
              d = 0.7571935
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

(D) (6 pts.) Match the following Effect sizes to their correspondence Plot:



Effect Size	Plot
0.10	(A)
0.20	(B)
0.30	(C)

(E) What would be the group size needed using effect size = 0.20 with a $power = 0.80$ (use the plot above)

For Effect Size = 0.20 and Power = 0.80 then n = 300

Q3) Given the following statistical summaries for the study in Q2 (Averages in a balanced design 4 observations in each cell). (50 Points)

Giving a final answer may not guarantee you a full credit.
You need to show your work.

Given that: `> model.tables(Q2aov)`

```
Tables of effects
 Frequency
Frequency
    0.1       1      10
 1.3606  1.0205 -2.3811
 Environment
Environment
     Air      H20 Salt H20
 -1.8806   1.0978   0.7828
 Frequency:Environment
         Environment
Frequency Air     H20     Salt H20
      0.1 -1.3181  0.7041  0.6140
      1   -0.9885  0.5281  0.4605
      10   2.3066 -1.2322 -1.0744
```

Given that $\sigma_{Crack\ Growth}$ =3.301666 and $\bar{y}$ ... = 4.294722

A) (10 pts.) Find the marginal averages for each level of the two factors:
$Frequency: i = 1, 2, 3$; $and\ Environment: j = 1, 2, 3$. i= 1 represents frequency = 0.1
Find also the average per combination for the missing ones only in the output below
Calculate the following:

$\bar{y}i..:$ $\bar{y}_1..=$ $\bar{y}_2..=$ $\bar{y}_3..=$

```
> aggregate(formula=Q1Data$`Crack Growth`~Q1Data$Frequency,FUN=mean)
  Q1Data$Frequency Q1Data$`Crack Growth`
1              0.1               7.660833
2              1.0               3.109167
3             10.0               2.114167
```

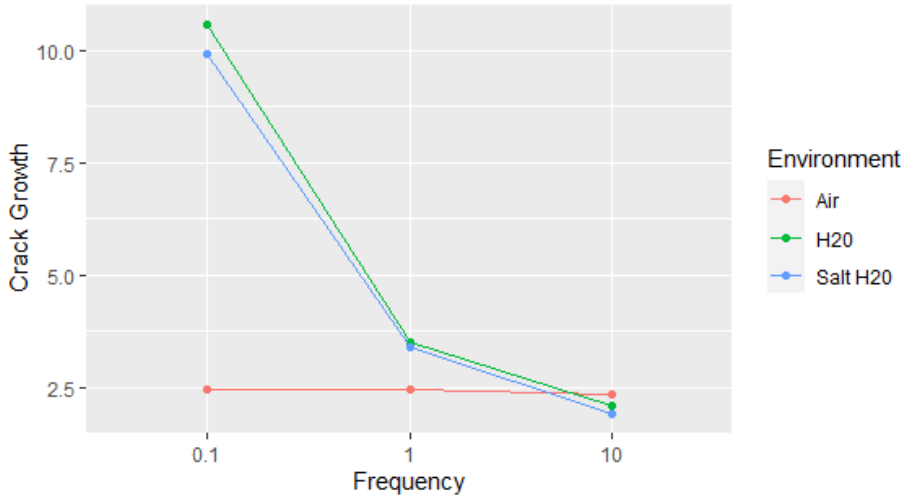
$\bar{y}.j.:$ $\bar{y}_{.1.} =$ $\bar{y}_{.2.} =$ $\bar{y}_{.3.} =$

```
> aggregate(formula=Q1Data$`Crack Growth`~Q1Data$Environment,FUN=mean)
  Q1Data$Environment Q1Data$`Crack Growth`
1                Air               2.414167
2                H20               5.392500
3           Salt H20               5.077500
```

$\bar{y}ij.:$ $\bar{y}_{11.} =$ $\bar{y}_{12.} =$ $\bar{y}_{13.} =$

```
> aggregate(formula=`Crack Growth`~Frequency+Environment,FUN=mean)
  Frequency Environment Crack Growth
1       0.1         Air       2.4600
4       0.1         H20      10.5900
7       0.1    Salt H20       9.9325
```

B) (10 pts.) Create an Interaction plot using Frequency as your x-axis. What can you conclud e from your graph?



C) (10 pts.) Find $SS_{Total.}$, $SS_{Frequncy}$, $SS_{Environment}$ and $SS_{Frequncy\ X\ Environment}$
Given that $SS_{Residuals}$ = 5.42

$SS_{Total.}$ = 35*(3.301666)^2 = 381.535

$SS_{Frequncy}$ = 4*3*[(7.660833-4.294722)^2 + (3.109167-4.294722)^2 + (2.114167-4.294722)^2] = 209.89

$SS_{Environment}$ = 4*3*[(2.414167-4.294722)^2 + (5.392500-4.294722)^2 + (5.077500-4.294722)^2] = 64.25

$SS_{Frequncy\ X\ Environment}$ = $SS_{Total.}$ - $SS_{Frequncy}$ - $SS_{Environment}$ - $SS_{Residuals}$ = 101.97

D) (20 pts.) Create a complete ANOVA table representing the sources of variations, the sum of squares, dfs, mean sum of squares and the corresponding F-values. (estimate your P-values)

```
> summary(Q1aov)
                    Df      Sum Sq      Mean Sq F value       Pr(>F)
Frequency            2      209.89       104.95   522.4      < 2e-16 ***
Environment          2       64.25        32.13   159.9     1.08e-15 ***
Frequency:Environment 4     101.97        25.49   126.9      < 2e-16 ***
Residuals           27        5.42         0.20
Total               35     381.535       10.901
```