STAT 135 Concepts of Statistics **Name:** ______________________
Summer 2021 Lec 001
Practice Final Exam
09:00 - 11:00 AM, 08/10/2021
Time Limit: In class

Instructions: This exam is open book. It contains 4 pages (including this cover page) and 5 questions. Total of points is 105.

1. Start with the problems that you find the easiest. If you are stuck, move on to the next question.
2. Write clearly and show you work. Partial credits will be assigned for attempting the questions except for the bonus question. Only a *complete and correct proof* for the bonus question will be acknowledged.
3. The bCourses submission is due at **12:00 AM**. Make sure you leave ample time to scan and submit you answers. *Late submissions will be penalized.*

Grade Table (for GSI use only)

Question	Points	Score
1	15	
2	21	
3	30	
4	34	
5	5	
Total:	105	

1. (15 points) During phase III trial of the Pfizer/BNT 162b2 vaccine, some vaccine recipients were asked to complete diaries of their symptoms during the 7 days after vaccination. Table 1 summarizes the number of recipients who experienced diarrhea:
   (a) (10 points) Scientists want to see whether different age groups after the 2nd dose have equal probabilities of having diarrhea of a certain severity. State appropriate null and alternative hypotheses and perform an appropriate test at $\alpha = 0.05$;
   (b) (5 points) Report the relative frequencies table, and comment on the discrepancies among the groups.
2. (21 points) A county environmental agency suspects that the fish in a particular polluted lake have elevated mercury levels (mg/kg). In order to test her suspicions, she samples stripped bass in an unpolluted lake and compares the results to stripped bass sampled in the lake that is polluted:

	Pfizer-BioNTech Dose 2		
	Aged 12-15	Aged 18-55	Aged > 55
No diarrhea	1038	1879	1523
Mild	59	179	114
Moderate	6	36	21
Severe	0	4	2

Table 1: Diarrhea among vaccine recipients — Mild: 2 to 3 loose stools in 24 hours; moderate: 4 to 5 loose stools in 24 hours; severe: 6 or more loose stools in 24 hours. Data from CDC.gov.

```
unpolluted <- c(0.144,0.681,0.758,0.367,0.433,0.159,0.094,0.409,
       0.092,0.133,0.666,0.391,0.006,0.625,0.285)
polluted <- c(0.769,0.451,0.202,0.437,0.026,0.645,0.995,0.700,0.944,
       0.408,0.884,0.590,0.137,0.093,0.319,0.292,0.012)
```

(a) (8 points) Under the Normal population assumption, test at $\alpha = 0.05$ whether the polluted lake has a higher mercury level.

(b) (8 points) Test the same hypothesis using a non-parametric method step by step.

(c) (5 points) Report the p-values from both tests. Are they very different?

3. (30 points) A researcher ran a two-factor experiment to compare the growth of 3 different species (Factor A) under different fertilizer treatments in a greenhouse (Factor B). He assigned combinations of fertilizer and species levels to 72 pots to have 6 replications in the greenhouse. This would be a referred to as 3×4 factorial treatment design. The data is in Table 2 (You can read this data set via copying and pasting in `R`).

   (a) (6 points) Make a Treatment Mean Plot to visually examine whether there is interaction between the two factors;

   (b) (10 points) Obtain the analysis of variance table by hand. Does any one source account for most of the total variability? Explain;

   (c) (6 points) Test whether or not the two factors interact; use $\alpha = 0.05$. State the alternatives, rejection region, and conclusion. What is the p-value of the test?

   (d) (8 points) Use Tukey and Bonferroni to test all pairwise comparisons for factor B at $\alpha = 0.05$, and visualize the two sets of confidence intervals. What level of factor B produces the slowest growth?

4. (34 points) A researcher wanted to understand the determinants of having a high body mass index (BMI) in adults (age ≥ 20), and conducted a survey with 432 individuals to collect their average dietary caloric intakes ($kcal$), ages and their BMI. Read in the BMI data by doing

   ```
   BMI=read.table("BMI.txt", sep=" ", header=TRUE)
   ```

   (a) (7 points) Fit the linear regression model:

$$\text{bmi}_i = \beta_0 + \beta_1 \text{kcal}_i + \beta_2 \text{age}_i + \epsilon_i, \epsilon \sim N(0, \sigma^2).$$

Factor A	Factor B			
	Control	**F1**	**F2**	**F3**
Species 1	21.0	32.0	22.5	28.0
	19.5	30.5	26.0	27.5
	22.5	25.0	28.0	31.0
	21.5	27.5	27.0	29.5
	20.5	28.0	26.5	30.0
	21.0	28.6	25.2	29.2
Species 2	23.7	30.1	30.6	36.1
	23.8	28.9	31.1	36.6
	23.7	34.4	34.9	37.1
	22.8	32.7	30.1	36.8
	22.8	32.7	30.1	36.8
	24.4	32.7	25.5	37.1
Species 3	25.1	28.4	22.8	32.8
	22.6	26.4	23.2	34.3
	24.5	27.8	26.4	33.3
	23.7	26.7	23.8	31.9
	22.6	25.3	25.4	32.6
	23.9	25.9	22.7	30.6

Table 2: Greenhouse data set

Examine pairwise scatter-plots, model residuals vs fitted values, and QQ-plots of residuals. Describe the violations of the model assumptions that are evident in the residual analysis.

(b) (7 points) To tackle the violation of the normality assumption, we apply the Box-Cox transformation to the response variable. Box-Cox transformation is a widely used method to find some constant λ such that the transformed Y^λ is as close to normally distributed as possible. (When $\lambda = 0$, the transformation becomes $\log(Y)$.) For the variable bmi, the best λ is 1.89. Try and fit the model

$$\text{bmi}_i^{1.89} = \beta_0 + \beta_1 \text{kcal}_i + \beta_2 \text{age}_i + \epsilon_i, \epsilon \sim N(0, \sigma^2).$$

Examine pairwise scatter-plots, model residuals vs fitted values and QQ-plots of residuals. Describe any violations of the modeling assumptions that are evident in the residual analysis. Has this model improved upon the model in (a)?

(c) (9 points) To fix the non-linearity that you might have observed in (a) and (b), we add the polynomial powers of the two predictors and fit the linear regression model

$$\text{bmi}_i^{1.89} = \beta_0 + \beta_1 \text{kcal}_i + \beta_2 \text{kcal}_i^2 + \beta_3 \text{age}_i + \beta_4 \text{age}_i^2 + \epsilon_i, \epsilon \sim N(0, \sigma^2).$$

Report the final model fit and the estimates of all model parameters. Interpret the estimated model by describing in words the relationship between BMI, caloric intake, and age as you have modeled them. List a few evident improvements compared to the previous models. Include any plots you think are helpful.

(d) (6 points) Report 95% exact confidence intervals of β_1 and β_3 for the model in (c);

(e) (5 points) The researcher now wants to make predictions about the BMI of any individual aged 30 with $5kcal$ caloric intake. Construct a 95% confidence interval for the mean BMI of a individual of the same age and caloric intake, and also construct a 95% prediction interval for the BMI of this individual. (Try not to use `predict()` to obtain the intervals.) Compared to the average BMI of the 432 individuals in the survey, is the predicted BMI relatively high?

5. (5 points) *(Bonus question) Let $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, \sigma^2)$, and suppose that the prior distribution on θ is $N(\mu, \tau^2)$. (Here we assume that σ^2, μ and τ^2 are all known.) Prove that the posterior distribution of θ is also normal, with mean and variance given by

$$\begin{aligned} E(\theta|\mathbf{X}_n) &= \frac{\tau^2}{\tau^2 + \sigma^2/n}\bar{X}_n + \frac{\sigma^2/n}{\sigma^2/n + \tau^2}\mu, \\ \mathrm{Var}(\theta|\mathbf{X}_n) &= \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}. \end{aligned}$$

Comment on the contributions of data information and prior information to the posterior distribution as $n \to \infty$.