Problem 1.1

Suppose the observations $T_1, \dots, T_n$ are independent random variables with T_i having density $f(t|x_i) = \frac{x_i}{\theta}\exp(-t\,x_i/\theta)$ for t>0, where x_i's are known positive quantities, and θ is an unknown positive parameter.

a) Construct a test procedure for the hypotheses $H_0: \theta = 1$ vs $H_1: \theta > 1$
b) Construct a 95% confidence interval for θ

Problem 1.2

Let $X_1, \dots, X_n$ be a random sample from $Uniform(\theta, 2\theta)$ distribution for $\theta > 0$.

a) Find the MLE of θ.
b) Is the MLE a sufficient statistics for θ? Justify your assertion.
c) Find a linear function of your MLE such that it is unbiased.
d) Let $X_{(1)}\ and\ X_{(n)}$ be the smallest and the largest order statistics, respectively. Find constants a and b such that $\hat{\theta} = aX_{(1)} + bX_{(n)}$ is an unbiased estimator of θ, and that $P\left(\frac{X_{(n)}}{2} \le \hat{\theta} \le X_{(1)}\right) = 1$. Why is $P\left(\frac{X_{(n)}}{2} \le \hat{\theta} \le X_{(1)}\right) = 1$ desirable?

Problem 1.3

Shown below are the number of galleys for a manuscript (X) and that the dollars cost of correcting typographical errors (Y) in a random sample of recent orders handled by a firm specializing in technical manuscripts.

i	1	2	3	4	5	6
X_i	7	12	4	14	25	30
Y_i	120	210	75	250	440	540

a) For each of the following three regression models, find the formula of the least squares estimator of β_0 and β_1. What is the distribution of each estimator?

Model I: $Y_i = \beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$ independently.
Model II: $Y_i = \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$ independently.
Model III: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$ independently.

b) For the given data, which model do you recommend? and why?
c) For your recommended model, estimate the probability that Y will fall between 200 and 300 for X=14.
d) Any concerns with the model assumptions for the given data? Please briefly explain and be specific.

Problem 1.4

This problem concerns discrete count data (e.g. number of insects, number of cases of a disease) collected using a completely randomized design (CRD). We know that the basic list of sources of variation for a CRD (listed in analysis of variance table format) is

Source
treatment
experimental unit within treatment [denoted unit(trt)]

We know that two standard models for count data arising from a CRD are

- "Poisson-normal": Assume the unit(trt) effect follows a $N\left(0,\sigma^2\right)$ distribution and the observations, conditional on the unit effect follow a Poisson distribution with rate parameter λ , i.e. $y \mid u \sim Poisson\left(\lambda\right)$
- "Poisson-gamma": Assume the unit(trt) effect follows a $Gamma\left(1/\phi, \phi\right)$ distribution and the observations, conditional on the unit effect follow a Poisson distribution with rate parameter λu . i.e. $y \mid u \sim Poisson\left(\lambda u\right)$

a) Write the joint p.d.f. of y and u under the "Poisson-normal" model. Hint: use the p.d.f. forms given below.
b) How would you obtain the marginal distribution of y under the "Poisson-normal" model? You do not need to write the distribution – just show how you would obtain it.
c) Show how you would obtain the expected value of y under the "Poisson-normal" model.
d) Now consider the "Poisson-gamma" model. Show that the marginal distribution of y is $\text{Negative Binomial}\left(\lambda, \phi\right)$.
e) For the negative binomial distribution of y that you derived in part (d), show that as $\phi \to 0$, its limit distribution is $Poisson\left(\lambda\right)$. Hint: use the alternative form of the negative binomial given below, but write it using $\frac{\lambda}{\lambda + k}$ instead of p .

For your convenience, here are the needed probability density functions
Hint: the notation is intentional and must be observed
y denotes the observed random variable
u denotes the unit-level effect
λ denotes the expected value of the Poisson and gamma distributions referred to in the problem
ϕ denotes the scale parameter in the gamma and negative binomial distributions

Distribution	**Probability Density Function**
$Poisson\left(\lambda\right)$	$\frac{\exp\left(-\lambda\right)\lambda^y}{y!}$

$Gamma\left(1/\phi, \phi\right)$	$\frac{1}{\Gamma\left(1/\phi\right)\phi^{1/\phi}} u^{\left(1/\phi\right)-1} \exp\left(-\frac{u}{\phi}\right)$
Negative Binomial$\left(\lambda, \phi\right)$	$\binom{y+\frac{1}{\phi}-1}{y}\left(\frac{\lambda\phi}{1+\lambda\phi}\right)^{y}\left(\frac{1}{1+\lambda\phi}\right)^{1/\phi}$
$N\left(0, \sigma^2\right)$	$\left(2\pi\sigma^2\right)^{-1/2} \exp\left(-\frac{u^2}{2\sigma^2}\right)$
alternative form of Negative Binomial	$\binom{y+k-1}{y} p^{y}\left(1-p\right)^{k}$, where $k = 1/\phi$ and $p = \frac{\lambda}{\lambda+k}$

Problem 2.1

This problem concerns a hypothetical experiment to compare treatments for their effect on the performance of a type of plant. Of interest is the number of nodules the plant develops at a critical stage of its development. If there are too few nodules, the plant's disease resistance and productivity may be compromised. An expected count of 7 nodules is considered a "magic number" above which the plant typically performs well.

The experiment is conducted in a greenhouse with eight benches. Each bench has room for three plants. Treatments 1, 2 and 3 are applied to plants on four benches (assigned at random) and treatments 4, 5 and 6 are applied to plants on the other four benches.

Researchers' initial analysis and report are as follows.

A generalized linear model assuming a Poisson distribution was used. The standard link function, $\log(\lambda_i)$, where λ_i denotes the expected count of the ith treatment, was fit to the linear predictor $\eta + \tau_i$, where τ_i denotes the ith treatment effect. SAS® PROC GLIMMIX was used to implement the analysis. Selected results are shown on the next page.

Based on the analysis, the report concluded:

There is statistically significant evidence of a difference among treatment expected counts (p=0.0058). Treatments 2, 3, 5 and 6 are the best treatments. There is no statistically significant (p>0.05) evidence of a difference among these treatments, and all have lower 95% confidence limits of their expected counts greater than 10. In addition, treatment 4 has a lower confidence limit of 7.12, which is greater than 7 but significantly less than the best treatments.

a) The report was rejected. Give two distinct reasons why.
b) Write an appropriate model you would recommend being used to analyze these data as an alternative to the analysis reported above.
c) Because these are count data, reviewers disagreed about whether a Poisson or negative binomial distribution should be used to analyze the data. The data are provided in SAS file *QE_May_2020_Day_Two_P1.sas* and in Excel file *QE_May_2020_Day_Two_P1.csv*. Given the data, which distribution would you use? Cite relevant evidence to support your choice. (HINT: recall Day One problem involving the negative binomial and Poisson distributions).
d) Consistent with your model in (b) and decision in (c), reanalyze the data. Give the program statements (R or SAS) you use to analyze the data and *relevant* output from the analysis.
e) Write a short report summarizing the results of your analysis.
f) When the report is presented, suppose the target audience for the report has limited statistical background (e.g. only an undergraduate introduction to statistics). They say that they have trouble understanding your "complicated analysis." They would like a simpler explanation – maybe even a "simpler analysis." How would you address their concern in terms they would be likely to understand?

Analysis Results for Rejected Report in Problem 2.1

Fit Statistics	
-2 Log Likelihood	207.12
AIC (smaller is better)	219.12
AICC (smaller is better)	224.07
Pearson Chi-Square	110.98
Pearson Chi-Square / DF	6.17

Type III Tests of Fixed Effects

Effect	*Num DF*	*Chi-Square*	*Pr > ChiSq*
treatment	5	16.41	0.0058

treatment Least Squares Means

	model scale				*data scale*			
treatment	*Estimate*	*Standard Error*	*Lower*	*Upper*	*Mean*	*Standard Error Mean*	*Lower Mean*	*Upper Mean*
1	2.1401	0.1715	1.8039	2.4762	8.5000	1.4577	6.0735	11.8959
2	2.7081	0.1291	2.4550	2.9611	15.0000	1.9365	11.6467	19.3188
3	2.8332	0.1213	2.5955	3.0709	17.0000	2.0616	13.4037	21.5612
4	2.2773	0.1601	1.9634	2.5911	9.7500	1.5612	7.1237	13.3446
5	2.5649	0.1387	2.2932	2.8367	13.0000	1.8028	9.9061	17.0602
6	2.7081	0.1291	2.4550	2.9611	15.0000	1.9365	11.6467	19.3188

T Grouping for treatment Least Squares Means (Alpha=0.05)

LS-means with the same letter are not significantly different.

treatment	*Estimate*		
3	2.8332		A
2	2.7081		A
6	2.7081		A
5	2.5649	B	A
4	2.2773	B	
1	2.1401	B	

Problem 2.2

A researcher is developing a new single walled carbon nano-tube sensor array capable of visualizing real time cellular Nitric Oxide signaling. The experiment was conducted at 5 concentration levels (10, 20, 30, 40, and 50 mg/L) at 4 treatments (e.g. avidin +Biotin, avidin + non-Biotin) for a total of 20 concentration by treatment combinations. Each combination was conducted on three glass slides (3 independent reps). Within each glass slide, 5 subsamples were taken. You are measuring the intensity and smoothness at each subsample. The data are available in the QE_May_2020_Day_Two_P2_data.csv file.

The researcher is interested in analyzing the effect of the treatments on the intensity and index. You are a statistical consultant giving advice to the researcher.

You need to analyze the data. You also need to write a report (please note that your report has to be written to a scientist who is only familiar with basic statistics). The report needs to include the following:

1. Description and the objectives of the study.
2. The specifications of the experimental design (make sure you include everything, and you are specific).
3. The specification of the treatment design (make sure you include everything, and you are specific).
4. A table showing sources of variation for which an appropriate analysis of these data must account.
5. The model (make sure you define all of the terms and you specify distributions where needed) and assumptions.
6. The steps of the analysis and the findings. Make sure that you interpret the necessary results. Include plots and anything that is helpful to understand the analysis and results.
7. Conclusion.
8. You also need to provide the SAS/R code in the Appendix.

Please note that the report has a strict **three-page limit**!