Final Review

1. Draw and properly label the four steps in the statistics lifecycle. Generally speaking, the information we covered in the chapters this semester was Sampling Methodology (Not in the Text), Descriptive Statistics (Ch 1), Probability Concepts (Ch 2), Discrete RV's (Ch 3), Continuous RV's (Ch 4), Point Estimation (Ch 6), Confidence Intervals (Ch 7), One-Sample Hypothesis Testing (Ch 8), and Two-Sample Hypothesis Testing (Ch 9). Group these into where they fall along the statistics lifecycle.
2. *Survey Design* – Suppose you are working for a start-up as the new Data Analyst and are given access to the company's SQL server. This contains info about all 7 clients, including client ID, revenue per client over the last year, number of contracts with that client over the past year, and number of employees that client has. The data is as follows:

Client ID	Revenue	Number of Contracts	# of Employees
1000001	$198,000	12	3
1000002	$154,832	8	2
1000003	$167,130	4	2
1000004	$202,452	7	3
1000005	$200,000	5	2
1000006	$200,000	9	10
1000007	$158,971	11	2

   a. What would be your primary question of interest?
   b. Suppose now you'd like to craft a survey to get a bit more information. Take me through the Who, What, Where, When, How, and Why. Who is your audience, what are you asking, where will they fill it out, when will you send it out/close the survey, how will you sample, why are you asking these new questions/what does it add?
   c. What are the pros and cons of the different sampling methodologies?
3. *EDA*
   a. When do you use the median instead of the mean? How do you check this?
   b. The histogram of revenue shows the distribution is skewed right. Draw a rough, smoothed version of what this looks like. Calculate the appropriate measure of central tendency. What is the appropriate measure of spread?
   c. What is the mode for # Employees? Describe the variable in terms of its modality (uniform, unimodal, bimodal).
   d. Practically speaking, why do you think the mode # employees is as you claimed?
   e. What is a reasonable support for the distribution of the RV Revenue?
   f. Let's test whether revenue follows a U(150,000, 250,000) distribution. Here is the table for the U(150,000, 250,000) QQ Plot :

Index	x_i (observed)	p_i	q_i (theoretical)
1	198,000		
2	154,832		
3	167,130		
4	202,452		

5	200,000		
6	200,000		
7	158,971		

i. Calculate p_4 and q_4
ii. How do you make conclusions based on a QQ Plot?
iii. What qualities should any good plot have, generally speaking? Not just QQ-plots.

g. Is client ID categorical or quantitative?
h. I think we can all agree there is an outlier in the # of Employees. What are some ways to deal with this?
i. Suppose that a variable follows a $Gamma\left(1,\frac{1}{3}\right)$ distribution. Find the IQR of this random variable.

4. *Probability/Modeling*
   a. If a RV X, the number of trials the first snow day, has a geometric distribution with mean 5, find, given it's been 5 days without a snow day already, the probability that it will be more than 8 days until the first snow day. What property can help here? What is the continuous counterpart, the only continuous distribution that also has this property?
   b. Derive the cdf of an $Exp(\lambda)$ distribution.
   c. Let the RV Y be the number of clients with revenue higher than $190,000. From prior industry experience, we know that typically the probability of retaining a client with revenue higher than $190,000 is .42. What is the expected number of clients from our sample to have revenue higher than $190,000? If you use a distribution, please justify its use. Then you can just use the shortcut formula for expectation. Compare this to the actual number.
   d. Now, suppose that someone at the company tells you that the model they've been using for the number of clients with revenue higher than $190,000 is
   $$f(y) = \frac{3x}{n} \ \ for\ x = 1,2,3,\dots$$
   Why can't this formula be correct?
   e. Suppose that the company is considering advertising to specific new clients. They will spend $25,000 per client. Assuming that they will gain the expected revenue of the clients they already have, and with a probability of gaining the client's business of .24, what is the expected gain per client advertised to?
   f. Suppose now the company realizes the probability of actually gaining the client's business is .12. What is the amount of money they are willing to now spend on advertising per client in order to achieve the same expected gain from each client advertised to?
   g. Prove that the Poisson RV is a valid distribution.
5. *Inference*
   a. Recall the Pareto distribution that we encountered in Quiz 3:

Consider $X_1, X_2, \dots, X_n$ independently and identically distributed (i.i.d.) $Pareto(\alpha, x_m)$. This implies that

$$f(x_i; \vec{\theta}) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \text{ for } x_i \in [x_m, \infty) \quad i = 1,2, \dots, n$$

where $\vec{\theta} = (\alpha, x_m)$ both unknown.

i. Prove that the expectation of this distribution is $\frac{\alpha x_m}{\alpha - 1}$
ii. Estimate α using the method of moments.
iii. Estimate $\alpha^3 + 1$ and x_m using MLE.
iv. Is the MLE estimate for x_m unbiased?

b. What is the thought process for choosing the "best" estimator?
c. Derive the general formula for a 95% confidence interval surrounding the true population mean.
d. How can we make this interval from part (c) tighter?
e. Explain what a confidence interval really means.
f. Hypothesis Testing
    i. Lay out the basic structure of any hypothesis test (the 4 parts)
    ii. What is the fundamental assumption made during hypothesis testing?
    iii. Why can we never conclude that H_0 is true?
    iv. Define what a p-value is.
    v. What is the difference between a one-sided and two-sided alternative hypothesis?
    vi. What is the relationship between the power of a test and type one error?
    vii. What is one way in which p-values can be "abused."
    viii. What are the α and β of an ideal hypothesis test? What is the thought process in choosing the "best" hypothesis test in terms of α and β?
g. Suppose that we now want to perform a test to see if those clients with 3 or more employees (large) have higher revenue than those with 2 employees (small). We have that $s_{small} = 20491.68$, $s_{large} = 2229.82$, $\bar{x}_{small} = 170233$, and $\bar{x}_{large} = 200151$. We have reason to assume that the population variances are equal, and histograms show that both samples are symmetric.
    i. Which test is appropriate. Justify your claim by confirming the assumptions are met.
    ii. What are the null and alternative hypotheses?
    iii. Calculate the test stat and reject region. What is your conclusion at the alpha significance level of .05 (no context needed)?
    iv. Calculate the p-value. What is your conclusion at the alpha significance level of .05 ***in context***?
    v. What is a Type 1 error here in context? What is a Type 2 error here in context?
    vi. Provide a 95% C.I. for $\mu_1 - \mu_2$ and interpret it in context.

vii. Use the interval from part (vi) to conclude about the alternative hypothesis at the alpha significance level of .05. How do you know? (no context needed here).

viii. If we didn't know that the population variances are equal, how would the test change? (just name the test).