Practice Final Exam

1. In RNA sequencing experiments, the counts of expressed genes can be modeled as a Poisson distribution with pmf $f(x) = \Pr(X = x) = \lambda^x e^{-\lambda}/x!$. On the other hand, non-expressed genes have zero counts. This creates an issue in modeling the data, as genes with zero counts could be expressed and the zero is due to random sample or they could just not be expressed. To accommodate this, we can model all of the counts as a zero inflated Poisson distribution. To do this, we assume that for each gene there is some probability p of being expressed. Conditional upon being expressed, the counts follow a Poisson distribution with parameter λ. If the gene is not expressed, with probability $1 - p$, then the count is zero. Suppose we have 100 we investigate with counts $x_1, \ldots, x_{100}$.

   a. If $\lambda = 2$ and $p = 0.25$, calculate the probability that $X = 0$.

   b. Write out the log-likelihood function.

   c. Set up the maximum likelihood equations. Do not solve.

   d. Calculate the expectation and variance of a zero-inflated Poisson distribution with parameters p and λ.

   e. Set up the Method of Moments equations. Do not solve.

2. A survey was given to a random sample of college students to assess the support of gay marriage. 242 females and 207 males were surveyed. 176 females expressed support while 134 males expressed. a. Construct 95% confidence intervals for the percentage of support for gay marriage among college females and college males, seperately.

b. Construct a 95% confidence interval for the difference between the two proportions. Use the fact that the pooled standard error is $\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ with $\hat{p}$ equal to the pooled proportion.

c. If we were to do a hypothesis test to test for a difference between the two proportions, what would be the null and alternative hypotheses?

d. Is there a statistically significant difference between the two proportions at the 95% confidence level?

3. Suppose that Y, X_1, X_2 and ϵ are all Gaussian (aka Normal) random variables and that $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

a. If X_1, X_2, X_3, and ϵ are all independent standard normal random variables (mean 0 and variance 1) and $\beta_0 = 10$, $\beta_1 = 2$, $\beta_2 = 5$, and $\beta_3 = 1$, what is the variance of Y?

b. If X_1, X_2, and ϵ are all independent standard normal random variables (mean 0 and variance 1) and $\beta_0 = 10$, $\beta_1 = 2$, and $\beta_2 = 5$, what is the correlation of X_1 and Y?

c. We have 20 total observations of Y, X_1, X_2, and X_3. We fit linear models using X_1 alone, X_1 and X_2, and all three X's. State which of the three models will be chosen according to each of the below criteria.

criterion	X_1	X_1, X_2	X_1, X_2, X_3
BIC	125.5	128.2	128.1
AIC	122.5	124.2	123.1
adj R^2	-0.003	-0.045	0.048
5-fold CV	135.9	162.3	104.2

d. Below is a table for fitting the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. What is the interpretation of the coefficient for X_2?

Table 2: Fitting linear model: $Y \sim X_1 + X_2$

	Estimate	Std. Error	t value	Pr(>|t|)
X_1	1.969	0.4229	4.657	0.000226
X_2	4.747	0.2657	17.86	1.878e-12
(Intercept)	10.11	0.2968	34.07	4.337e-17

4. In a manufactured product you want to be certain that the delivered product is the weight advertised. The goal is 100 grams.

   a. To test the process you weigh 5 products and get weights of 106, 104, 99, 95, and 97. Calculate the sample mean and standard deviation.

   b. Obviously 5 samples is not sufficient so you take 20 more samples (for 25 total) and get a sample mean of 101.84 and sample standard deviation of 3.3. State the null and alternative hypotheses and then compute a p-value under the null distribution.

   c. If the average weight is in fact 101 with variance equal to 4, what is the power with 25 observations compared to null distribution of $N(100, 4)$ at the 95% confidence level?

5. Suppose that we want to estimate the size of the wolf population in Yellowstone. One method to do this is to go out, capture wolves, tag them, and release them. You then go out later and capture more wolves, noting how many are recaptured (this is known as mark-recapture). Suppose that there are in fact 400 wolves in Yellowstone. In the first round you capture and tag 50 wolves.

   a. Suppose you capture another 50 wolves in the second round. What is the expected number of wolves you will capture?

   b. Suppose you capture another 50 wolves in the second round. What is the probability you capture 10 tagged wolves in the second round of captures?

   c. The Lincoln-Peterson estimator for the total number of captured individuals is $\frac{Kn}{k}$ where K is the total number of capture in the second round, n is the number of tagged wolves in the first round, and k is the number of tagged wolves seen in the second round. Write out the formula to calculate the expectation of the Lincoln-Peterson estimator when $N = 400$, $n = 50$, and $K = 50$ (the only random part is k).

6. In basketball, baseball, and hockey playoffs, teams play each other for the best of 7 or 5 series.

   a. If team A and team B play each other and team A has a 60% chance of winning each game (whether home or away), what is the probability that team A will win the best of 7 series?

   b. Suppose that team A has a 60% chance of winning home games and a 50% chance of winning away games and that games alternate home and away (starting with a home game for A). What is the probability that team A wins a best of 7 series?

   c. Does the order that the teams play matter (as long as team A plays 4 home games and 3 away games)?

7. Your friend wants to sell his car, but they tell you they won't accept less than $10,000. You think the car is worth more like $8,000. Suppose offers will follow an exponential distribution with mean $8,000.

   a. What is the probability the first offer will be greater than $10,000?

   b. What is the probability your friend will have to take more than 5 offers before getting one that is $10,000 or larger?

   c. What is the expected amount of time they would have to wait for an offer larger than $10,000?

   d. Suppose they change their strategy to taking the largest of the first 5 offers. What would the expected sale price be?

8. Suppose you're on the game show *Let's Make a Deal*. The host reveals 3 doors. One contains a car and the other two contain goats, but you want the car. You pick one of the 3 doors. The host, Monty Hall, says, "Let's make a deal." He goes to the nearest (lowest in numerical order) door that contains a goat (for example, if you selected Door 1 and the car was indeed behind Door 1, then the host would always open Door 2, never Door 3). What are the probabilities that you will win the car if you stay with your original choice vs if you switch?

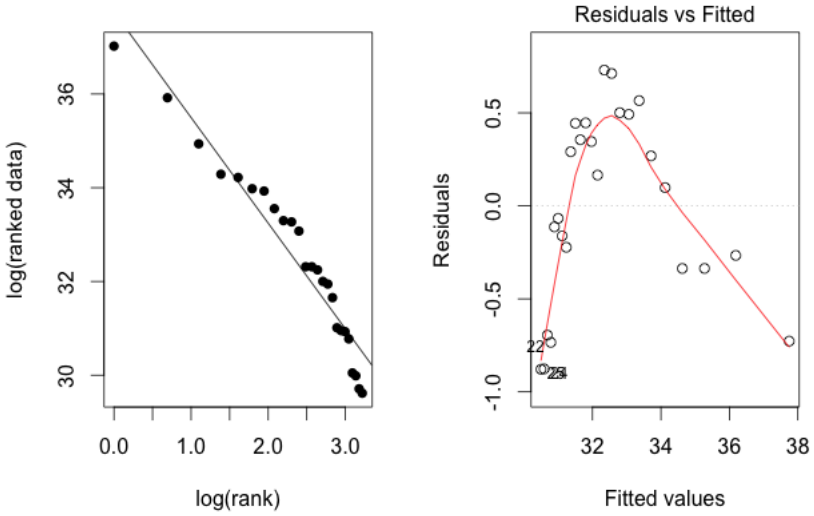
9. A popular model for long tailed distributions (meaning they are likely to take on large values) is power law type distributions. This has been used to model word frequency and city population. If X is power law distributions the X has cdf that is proportional to $x^{-\alpha}$ for $x > 1$ and some $\alpha > 0$.

   a. If $\alpha = 3$ then $f(x) = cx^{-3}$. Calculate the normalizing constant c.

   b. If $\alpha = 3$, calculate the expectation of X.

   c. Set up the maximum likelihood equation.

   d. A common method to fitting power law data is to plot the log of the ranks versus the log of the ranked data. Below is a figure of a power law linear fit. Does this fit look homoscedastic?



10. A random sample of 100 voters in a community produced 59 voters in favor of measure A. You would like to test to see if measure A passes.

    a. State the null and alternative hypotheses.

    b. Construct a 95% confidence interval for the support of measure A.

    c. Compute the p-value. Will measure A win with 90% confidence?

    d. Suppose the true support for measure A is 58%. What is the minimum sample size needed for a power of 90% at a 95% confidence level?

12. After the 1936 presidential polling fiasco, more rigorous polling methods were develop. The best known was the Gallup poll, who used a method known as quota sampling. Quota sampling is nothing more than a systematic effort to force the sample to fit a certain national profile by using quotas: The sample should have so many women, so many men, so many blacks, so many whites, so many under 40, so many over 40 etc. The numbers in each category are taken to represent the same proportions in the sample as are in the electorate at large. If we assume that every important characteristic of the population is taken into account when setting up the quotas, it is reasonable to expect that quota sampling will produce a good cross-section of the population and therefore lead to accurate predictions. For the 1948 election between Thomas Dewey and Harry Truman, Gallup conducted a poll with a sample size of about 3250. Each individual in the sample was inteviewed in person by a professional interviewer to minimize nonresponse bias, and each interviewer was given a very detailed set of quotas to meet. For example, an interviewer could have been given the following quotas: seven white males under 40 living in a rural area, five black males under 40 living in a rural area, six black females under 40 living in a rural area, etc. Other than meeting these quotas the ultimate choice of who was interviewed was left to each interviewer. Based on the results of this poll, Gallup predicted a victory for Dewey, the Republican candidate. The predicted breakdown of the vote was 50% for Dewey, 44% for Truman, and 6% for third-party candidates Strom Thurmond and Henry Wallace. The actual results of the election turned out to be almost exactly reversed: 50% for Truman, 45% for Dewey, and 5% for third-party candidates.

Can you identify a major problem or confounding factor with quota sampling as described above?

The following are true / false questions.

13. When there are more predictor variables (xs) overfitting is less likely in linear models.
14. Increasing the regularization parameter λ in lasso regression leads to sparser regression.
15. Increasing the regularization parameter λ in ridge regression leads to sparser regression
16. For a fixed size of the training and test set, increasing the complexity of the model always leads to reduction of the test error
17. A linear regression analysis allows you to make predictions for a variable using information about another variable and the relationship between them.
18. The standard error of the estimate is a measure of how far off are our predictions for X.
19. In any regression analysis, the total variance in Y can be broken down into two parts: the variance attributable to variation in X and that which is left over and unexplained.
20. If you wish to use many predictors to predict a single criterion, you must use each predictor in its own regression analysis. coefficients. coefficients.