Stat 446 **Name:** ____________________

Take home final

Due 12/13/16 at 8:00 AM

For this exam, you may use the textbook, any course materials provided on D2L, homeworks, and labs. You **may not** discuss questions or work together with classmates. You are welcome to contact the instructor with any questions related to better understanding of the questions themselves. Any relevant material from questions will be posted to D2L for the benefit of the entire class. For complete (and partial credit) please show all work, whether that be by hand or using R code.

1. Consider the dataset finalQ1.csv, which can be found on D2L. This contains data from a two-stage cluster sampling procedure. First a sample of 10 clusters is taken from a population of 50 clusters. Then a stratified sample is taken within each cluster, where there are 3 strata. Within each strata in each cluster 20 samples out of a population of 100 units are sampled. In other words, each cluster has a total of 300 secondary sampling units and a stratified random sample is taken to sample 20 in each stratum.

   (a) (10 points) Compute a point estimate of the total across the entire population.

   (b) (5 points) You will not find a variance formula in the notes to compute the variance of this two-stage cluster estimation procedure. Describe (explicitly) how this variance could be computed by integrating variance formulas from cluster sampling and stratified sampling or by a computational procedure (e.g. bootstrap).

2. (10 points) A sample has been taken using unequal probability sampling *with* replacement. There are 100 total units in the population and a sample of 25 has been taken, resulting in 20 unique sampling units. Using the dataset finalQ2.csv estimate the population total with a confidence interval. Note this dataset contains each y - value, the probability of being sampled p, and Q the number of times a unit was selected.

3. Assume there is a spatial grid of 400 sampling units divided into four different strata. We are interested in testing whether magpies prefer certain terrain, including residential areas. The response will be the number of magpies spotted in each grid. Each stratum consists of a habitat type with different costs to sample and a different number of sampling units:

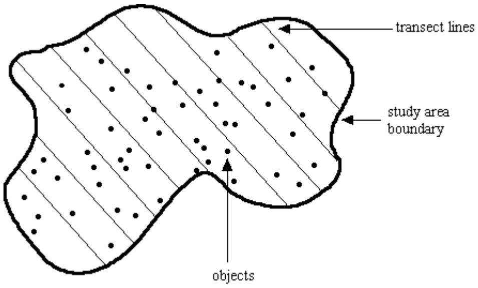
   - high alpine - 50 sampling units that cost \$40 each to sample.
   - prairie - 180 sampling units that cost \$20 each to sample.
   - residential - 100 sampling units that cost \$10 each to sample.
   - wetland - 70 sampling units that cost \$30 each to sample.

   (a) (10 points) You plan to employ a double sampling procedure to choose the optimum allocation of the sampling units across the strata. You begin by sampling 6 units from each stratum, from which the results can be found in finalQ3a.csv. Assume you have \$1000 more dollars to collect samples, choose the optimal allocation of those sampling units.

(b) (10 points) It turns your field technician forgot your sampling instructions and instead sampled each stratum proportionally such that twenty percent of the sampling units are sampled for each stratum. Use the data in finalQ3b.csv to estimate confidence intervals for the total number of magpies in each stratum as well as across the entire sampling grid.

(c) (10 points) While your field technician doesn't always follow instructions, the technician did manage to collect data on the number of almond poppy seed muffins for each residential sampling unit. It is known that almond poppy seed muffins attract magpies. Use the data in finalQ3c.csv to compute a confidence interval for the average number of magpies across the sampling units using a ratio estimator. It is known that the average number of muffins for each sampling unit in the residential area is 25.

4. (5 points) Transect sampling is commonly used in sampling ecological processes. Scientists will walk down a line (transect) to count the number of phenomenon observable from the path. The response could be the number of pine trees infected by mountain pine beetle. Below is an image where researchers will walk down each line to count the number of 'objects' that can be observed from each transect. Describe how this process is similar and different to sampling



procedures we have learned about in class.

5. (3 (extra credit) points) I will most likely teach this course next year. With an eye toward an improved version for the next iteration I'd appreciate a brief description of your thoughts. What did you like/dislike? What activities were helpful for your learning? What didn't work? How was the work load with HWs and labs?