STAT 446 Final Exam 40 points (48 points for Stat Grad students)

The exam is to submitted by 12:00 Monday, 12/7/15. You are to work alone on this exam. Evidence to the contrary will result in zero for the final exam grade. If you use R or SAS to answer a question, you must also submit the R or SAS code. E-mailed final exams and R/SAS code will not be accepted.

For any question demarked with ‡, your answer must be stated in complete sentences and using at most three sentences.

1. An inspector wants to estimate the average weight for filled cereal boxes at the packaging plant. Each box claims the weight is 16 ounces. The cereal is available to her in cartons containing 12 boxes. The inspector randomly selects 8 cartons from the population of 1000 cartons and weighs each box in the sampled cartons using a reliable (unbiased) scale. The results in ounces are shown below.

Carton	Box weight in ounces											
1	16.1	15.9	16.1	16.2	15.9	15.8	16.1	16.2	16.0	15.9	15.8	16.0
2	15.9	16.2	15.8	16.0	16.3	16.1	15.8	15.9	16.0	16.1	16.1	15.9
3	16.2	16.0	15.7	16.3	15.8	16.0	15.9	16.0	16.1	16.0	15.9	16.1
4	15.9	16.1	16.2	16.1	16.1	16.3	15.9	16.1	15.9	15.9	16.0	16.0
5	16.0	15.8	16.3	15.7	16.1	15.9	16.0	16.1	15.8	16.0	16.1	15.9
6	15.8	16.1	15.9	15.9	16.2	16.0	15.8	15.8	15.9	16.0	16.0	15.9
7	16.1	16.0	15.8	16.2	15.8	16.1	15.9	16.1	16.0	16.1	15.9	16.0
8	15.9	15.9	16.2	16.0	16.1	16.1	15.8	16.1	15.9	16.1	16.0	16.1

   (a) (2pt) Calculate the estimate of the $\overline{y}_U$.

   (b) (2pt) Calculate the standard error of the estimate.

   (c) (2pt) ‡ Find and interpret a 95% t-based confidence interval for $\overline{y}_U$ in the context of the study. No credit will be given if it is not in context.

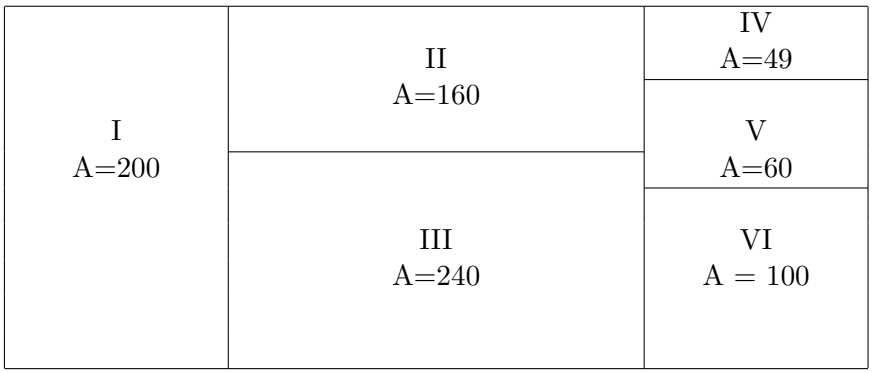
2. Let the 5 parenthetical values in the figure below represent the initial starting points for a systematic sample formed by sampling the secondary sampling units in the same locations in the remaining blocks.

(0)	3	3	2	0	0	1	0	0	1	0	1	1	0	1	0	0	3	2	1
1	2	2	(0)	2	0	4	3	0	0	0	1	5	1	1	1	1	0	2	1
(3)	3	(5)	5	2	0	6	2	3	2	4	3	2	1	2	3	3	2	4	2
2	7	6	5	4	3	9	4	1	4	6	8	4	6	4	6	7	4	4	1
6	(10)	8	2	5	7	6	6	2	4	4	3	8	7	7	3	6	4	3	3
4	3	10	7	8	8	7	6												
8	11	4	4	11	10	3	5												
8	7	10	8	5	9	6	6												
11	4	10	9	4	9	10	10												
10	10	10	6	3	9	9	7												
11	8	7	7	11	5	7	6												
5	9	5	8	10	9	6	11												
8	8	8	11	4	10	7	10												
3	11	3	10	4	7	3	6												
2	4	7	8	5	12	7	6												
0	3	3	2	0	0	1	0	0	1	0	1	1	0	1	2				
1	2	2	0	2	0	4	3	0	0	0	1	5	1	1	4				
3	3	5	5	2	0	6	2	3	2	4	3	2	1	2	5				
2	7	6	5	4	3	9	4	1	4	6	8	4	6	4	6				
6	10	8	2	5	7	6	6	2	4	4	3	8	7	7	7				

   (a) (2pt) Calculate an estimate of the population total t.

   (b) (3pt) Calculate the standard error of this estimate of t.

3. Consider the following population with $N = 6$ unequal-sized units, and $n = 2$ units are to be selected proportional to the area A with replacement.



   (a) (3pt) What are the selection (single draw) probabilities p_i for i = I, II, III, IV, V, and VI.

   (b) (1pt) The y values for the units are $y_I = 248$, $y_{II} = 204$, $y_{III} = 305$, $y_{IV} = 49$, $y_V = 78$, $y_{VI} = 126$. Calculate y_i/p_i for i = I, II, III, IV, V, and VI.

   (c) ‡ (2pt) In general, will Hansen-Hurwitz estimation provide an estimate of $\overline{y}_U$ with small variance? Justify your answer.

4. (2pt) **For stat grad students only:** In the previous example, suppose a sample of size $n = 2$ is taken without replacement and proportional to size. What are the first-order inclusion probabilities π_i for i = I, II, III, IV, V, and VI.

5. A student who did not take this course is interested in estimating the average genetic diversity of aspen trees within Rocky Mountain National Park. Aspen are clonal species meaning that they have very little genetic diversity within a stand. There are 100 aspen stands within the park. She takes a simple random sample of 10 stands and plans on genotyping every aspen tree within a stand.

   (a) ‡ (2pt) What sampling plan did she propose to use? Be as specific as possible.

   (b) ‡ (2pt) She only had enough resources to genotype 1000 trees, and the thousandth genotyped tree occurred in just the sixth sampled aspen stand. Her resources for gentyping are now exhausted and and has to stop collecting genotype data. She throws up her hands, swears something I do not want to repeat, and decides to terminate the project. Based on this information, what sampling design would you suggest if she could repeat the study (again with resources for genotyping at most 1000 trees)? Justify your answer.

6. ‡ (2pt) The Center for Disease Control (CDC) is tracking potential side effects of a new vaccine. There are 150 clinics distributing the vaccine. The CDC takes a simple random sample of 25 clinics. Then from each sampled clinic, they take a simple random sample of 50 patients that received the vaccine, and then record if the patient did or did not have any side effects. What type of sampling plan was used? Be specific.

7. ‡ (2pt) The CDC is tracking potential side effects of a new vaccine. There are 150 health clinics distributing the vaccine. The CDC takes a simple random sample of 10 patients that received the vaccine from each of the 150 clinics, and then record if the patient did or did not have any side effects. What type of sampling plan was used? Be specific.

8. A simple random sample of 85 US golf courses was selected was selected from a study population of 20000 US golf courses (6000 private and 4000 public). You will have data set with columns for the type of course (private or public), the total yardage for the 18 holes, and the green fee for one 18-hole round for each of these 85 courses.

   (a) (4pt) Compute a 95% confidence interval for the mean green fee for the study population using regression estimation assuming a simple linear regression model.

   (b) (2pt) ‡ Would you recommend generating a 95% confidence interval for the mean green fee using regression estimation or only using the simple random sample of green fee data (i.e., ignore total yardage in the analysis). Justify your answer.

   (c) ‡ (2pt) The researcher believes that stratified regression estimation using separate simple linear regression models (one model for public courses and one model for private courses) would be better than using a combined simple linear regression model for this data set. Is this a good suggestion?

   (d) (6pt) Whether or not you thought it was a good suggestion in (c), compute a 95% confidence interval for the mean green fee for the study population using stratified regression estimation using separate simple linear regression models.

   (e) ‡ (1pt) Suppose someone claims that this analysis is valid because we have a stratified SRS. Briefly comment on whether or not you agree with this statement.

   (f) (4pt) **For Stat Grad students only**: Fit separate quadratic regression models for each course type. You will notice a large change in the p-values for the estimates of the coefficients of the Yardage model terms. Why did this happen?

9. ‡ (2pt) **For Stat Grad students only**: Consider a population with $N = 7$ units. Consider the following sampling plan:

Sample number	Sample $\mathcal{S}$	$P(\mathcal{S})$
1	{1,2,3,4}	.2
2	{1,3,5,6}	.2
3	{2,3,4,7}	.2
4	{1,4,6,7}	.2
5	{2,4,6,7}	.1
6	{1,4,5,7}	.1

   Is the Horvitz-Thompson estimator $\widehat{V}(\widehat{t}_{ht})$ an unbiased estimator of the variance of $\widehat{t}_{ht}$? Justify your answer.