FINAL EXAM
STAT 5201
Fall 2019

Submit your answer on the class web site or in Room 313 Ford Hall
on or before Thursday, December 19 at 1:30 pm
In the second case please deliver it to the office staff
of the School of Statistics

READ BEFORE STARTING

You must work alone and may only discuss these questions with the TA or Professor Meeden. You may use the class notes, the text and any other sources of material you have access to.

Start each answer on a new page and make sure that you name is on each page.

If I discover a misprint or error in a question I will post a correction on my class web page. In case you think you have found an error you should check the class home page before contacting us.

1. Find a recent survey reported in a newspaper, magazine or on the web. Briefly describe the survey. What are the target population and sampled population? What conclusions are drawn from the survey in the article. Do you think these conclusions are justified? What are the possible sources of bias in the survey? Please be specific but brief.

2. Consider a population where you have in hand the values of an auxiliary variable which you know from past experience has a large positive correlation with the y characteristic of interest. For this problem use the auxiliary variable $xprob2$ which you can load into your R work space by doing the following

```
library("RCurl")
load(url("http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/f19prob2.RData"))
```

i) Find a sensible stratification based on $xprob2$. Briefly explain why you think your choice is a good one.

ii) For your stratification find the optimal allocation for a sample of size n.

3. To do this problem you need to load into your R working director the file $f19clus.RData$ using the commands.

```
library("RCurl")
load(url("http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/f19clus.RData"))
```

This file contains two R objects. The first is $clussmp$ which contains the results of a two stage cluster sample where 17 clusters where selected out of a population of 100 clusters. At both stages simple random sampling without replacement was used. The second object is $clussz$ which contains the sizes of the clusters in the sample. The average cluster size of the 83 clusters not in the sample is 24.205.

Give the value of your estimate for the population mean and an estimate of its variance. Briefly justify the choice of your estimate.

4. Let the distribution of $y_1, \ldots, y_n$ given θ be independent and identically distributed Bernoulli(θ) random variables and let the distribution for θ be Beta(α, β) where $\alpha > 0$ and $\beta > 0$.

i) Find an expression for the probability distribution $p(y_1, \ldots, y_n)$ and write a function in R which will calculate this value.

ii) Let f_1 and f_2 be two different beta distributions and let $0 < \lambda < 1$ be given. Then $f = \lambda f_1 + (1 - \lambda) f_2$ is a new prior distribution for θ. Given $\sum y_i$ find the posterior distribution of θ when f is the prior.

5. In Minnesota there are 327 school districts. School districts are classified as urban, suburban or other. In these districts there are 987 elementary schools. The total number of students in these schools was 395,449. A state agency wants an estimate of the total number of student absences there were in all the elementary schools last January. Note, an absence is a day when the school is open but a student is not there. Assume the agency has a list of the schools and the total number of students in each school.

Suggest a sampling design for the agency. What additional information, if any, could be provided by the state agency to help you choose the design. Do you see any problems in implementing your design? Specify your estimator and its estimated variance. Be brief but specific.

6. For this problem you need to get the data with the commands

```
library("RCurl")
load(url("http://users.stat.umn.edu/~gmeeden/classes/5201/moredata/f19popa.RData"))
```

The data set has three vectors. The vector y is the quantity of interest while $x1$ and $x2$ are auxiliary variables. Each are of length 2,000 and $x1$ is an increasing function of its labels. You be asked to do several simulation studies. In each one, the sample size $n = 100$ and you should take 300 samples. For both parts you should use three different designs: rep(1,2000), seq(3,1,length=2000) and seq(1,2,length=2000)

i) In this part you will be estimating the population total of y. You need to compare three estimators. The first is the usual Horvitz-Thompson estimator. The second is the calibrated HT estimator where the calibrate weights must satisfy three conditions. The sum of the calibrated weights should be 2000. In addition the sum of the calibrated weights times the sample values of $x1$ and $x2$ should equal the population totals of $x1$ and $x2$. Sometimes the design weights are not available when the data are analyzed. So for the third estimate we always assume that the design was simple random sampling without replacement even when is not. The corresponding design weights for this case are calibrated with the same three conditions that were used when creating the second estimator.

For each estimator you should compute its average value, its average relative bias,(est-tru)/tru, its average absolute error, the average length of its approximate 95% confidence interval and its frequency of coverage.

ii) In this part we are interested in estimating

$$\gamma(y) = \frac{\sum_{i=1}^{500} y_i}{\sum_{i=1501}^{2000} y_i}$$

the ratio of the total amount of y that belongs to the quarter of the poluation consisting of the 500 units with the smallest values of $x1$ to the quarter of the population consisting of the 500 units with the largest values of $x1$. In addition to the calibration constraints used in part i) you should add the constraints that the weights for the units in the smallest group in the sample and the weights for the units in the largest group in the sample should each sum to 500.