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.

Thirty five students took the exam. There were 2 scores in the 90's, 4 in the 70's, 6 in the 60's, 5 in the 50's, 6 in the 40's, 4 in the 30's, 4 in the 20's, 2 in the 10's and 2 in the 0's.

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 *xprob*2 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 *xprob*2. 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.

ANS

i) Looking at an histogram of *xprob*2 shows that it is strong skewed to the right. So just taking four equal strata of size 250 each is not a good idea. For this choice the strata variances of *xprob*2 are

$$0.03060133 \quad 0.11376839 \quad 0.47006139 \quad 7.84692110$$

A better choice would be strata that have more equal variances. Not clear what is the "best" choice but one that should do better has strata sizes of 500, 250, 125, 125 respectively with resulting strata variances of

$$0.2368053 \quad 0.4700614 \quad 0.7395922 \quad 1.0677196$$

ii) As always choose $n_h \propto N_h\sigma_h$ where N_h is the stratum size and σ_h is the stratum standard deviation of *xprob*2.

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.

ANS

You should use the ratio estimate since the cluster sizes vary.

Following the handout on working with cluster samples you find that the ratio estimate of the population mean is 11.62 and its estimated variance is 2.8. Note from the information given in the problem you can find the average population cluster size. This number is used when calculating the estimated variance.

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.

ANS

i) Here is the function:

$$\begin{aligned} p(y_1, \ldots, y_n) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{\sum y_i}(1-p)^{n-\sum y_i} p^{\alpha-1}(1-p)^{\beta-1}\, dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{\alpha+\sum y_i-1}(1-p)^{\beta-\sum y_i-1}\, dp \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)} \frac{\Gamma(\alpha+\sum y_i)\Gamma(\beta+n-\sum y_i)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned}$$

Note it is easy to compute this value in R, just use the gamma function.

ii) Let $y = (y_1, \ldots, y_n)$ then the joint distribution of y and θ is given by

$$\begin{aligned} p(y|\theta) f(\theta) &= \lambda p(y|\theta) f_1(\theta) + (1-\lambda) p(y|\theta) f_2(\theta) \\ &= \lambda p_1(y) \frac{f_1(\theta) p(y|\theta)}{p_1(y)} + +(1-\lambda) p_2(y) \frac{f_2(\theta) p(y|\theta)}{p_2(y)} \\ &= \lambda p_1(y) p_1(\theta|y) + (1-\lambda) p_2(y) p_2(\theta|y) \end{aligned}$$

Since

$$\begin{aligned} p(y) &= \lambda \int p(y|\theta) f_1(\theta)\, d\theta + (1-\lambda) \int p(y|\theta) f_2(\theta) \\ &= \lambda p_1(y) + (1-\lambda) p_2(y) \end{aligned}$$

we see that the weight given to $p_1(\theta|y)$ in the posterior is just

$$\lambda p_1(y) / (\lambda p_1(y) + (1-\lambda) p_2(y))$$

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.

ANS There are three strata: urban, suburban and other and the target population is the individual schools. The school districts are clusters. The size of the sample in each stratum should depend on the size of the stratum and the total number of students in the stratum. This should be information that the agency has. Nonresponse should not be a problem, but cost and how quickly he agency wants the answer could be a problem.

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(1,3,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.

ANS

i) In the following, ansdsgn is the HT estimator using the design weights. ansdsgncb is based on the calibrated design weights. ansrscb always assumes that the design was srs even when it was not and is calibrated in the same way. In all cases the sample size $n = 100$ and the results are based on 300 simulated samples.

```
 design=rep(1,2000)
             [,1]  [,2]     [,3]     [,4]  [,5]
ansdsgn   3588040 0.002 63491.49 302851.7 0.943
ansdsgncb 3587799 0.002 30831.24 351243.8 1.000
ansrscb   3587799 0.002 30831.24 351243.8 1.000

  design=seq(3,1,length=2000)
             [,1]   [,2]      [,3]     [,4] [,5]
ansdsgn   3602090  0.006 154703.55 773406.3 0.93
ansdsgncb 3594749  0.004  38764.59 827522.5 1.00
ansrscb   3576827 -0.001  44402.55 801835.2 1.00
```

```
  design=seq(1,3,length=2000)
             [,1]  [,2]     [,3]     [,4]  [,5]
ansdsgn   3590875 0.002 47868.81 229109.4 0.937
ansdsgncb 3583848 0.000 27435.49 289386.0 1.000
ansrscb   3618328 0.010 41825.89 280544.8 0.983
```

ii) First you need to calibrate the design weights and then use the function *wtpolyap* from the *R* package *polyapost*. Note my intervals are too long and I used the calibrated weights which summed to 2000, If you re-normed to sum to 100 as recommend in class the intervals would be even wider. The problem is that we have not taken into account the additional information we get by calibrating on $x1$ and $x2$.

```
dsgn<-rep(1,2000)
      [,1]  [,2]  [,3]  [,4] [,5]
cbans 0.620 0.003 0.010 0.209    1

 dsgn<-seq(1,3,length=2000)
Tue Dec  3 17:49:55 2019
       [,1]   [,2]  [,3]  [,4]  [,5]
cbans 0.598 -0.031 0.020 0.202 1.000

> dsgn<-seq(3,1,length=2000)
       [,1]  [,2]  [,3]  [,4] [,5]
cbans 0.643 0.041 0.026 0.218    1
```