FINAL EXAM
STAT 5201
Fall 2016

Due on the class Moodle site or in Room 313 Ford Hall
on Tuesday, December 20 at 11:00 AM
In the second case please deliver to the office staff
of the School of Statistics

READ BEFORE STARTING

You must work alone and may discuss these questions only with the TA or Glen Meeden. You may use the class notes, the text and any other sources of printed material.

Put each answer on a single sheet of paper. You may use both sides and additional sheets if needed. Number the question and put your name on each sheet.

If I discover a misprint or error in a question I will post a correction on the 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 brief.

2. In a small country a governmental department is interested in getting a sample of school children from grades three through six. Because of a shortage of buildings many of the schools had two shifts. That is one group of students came in the morning and a different group came in the afternoon. The department has a list of all the schools in the country and knows which schools have two shifts of students and which do not. Devise a sampling plan for selecting the students to appear in the sample.

3. For some population of size N and some fixed sampling design let π_1 be the inclusion probability for unit i. Assume a sample of size n was used to select a sample.

i) If unit i appears in the sample what is the weight we associate with it?

ii) Suppose the population can be partitioned into four disjoint groups or categories. Let N_j be the size of the j'th category. For this part of the problem we assume that the N_j's are not known. Assume that for units in category j there is a constant probability, say γ_i that they will respond if selected in the sample. These γ_j's are unknown. Suppose in our sample we see n_j units in category j and $0 < r_j \leq n_j$ respond. Note $n_1 + n_2 + n_3 + n_4 = n$. In this case how much weight should be assigned to a responder in category j.

iii) Answer the same question in part ii) but now assume that the N_j's are known.

iv) Instead of categories suppose that there is a real valued auxiliary variable, say age, attached to each unit and it is known that the probability of response depends on age. That is units of a similar age have a similar probability of responding when selected in the sample. Very briefly explain how you would assign adjusted weights of the responders in this case.

4. Use the following code to generate a random sample from of stratified population with four strata of size 2,000, 3,000, 8,000 and 5,000.

```
N<-c(2000,3000,8000,5000)  #these are the strata sizes
n<-0.01*N    #these are the sample sizes
mn<-c(900,700, 560, 1500)
std<-c(200,175,125,300)

set.seed(11447799)

strtsmp<-function(n,mn,std){
    ans<-list()
    K<-length(n)
    ans<-list()
    for(i in 1:K){
        ans[[i]]<-rnorm(n[i],mn[i],std[i])
    }
    return(ans)
}

smp<-strtsmp(n,mn,std)
```

Find the 95% confidence interval for the population mean given this sample. Note that we used

proportional allocation when selecting the sample. Given the results of the sample did this seem like a good idea. Include a copy of the code in your answer.

5. The Horvitz-Thompson (HT) estimator can be used when the model underlying the population is

$$y_i = \beta x_i + z_i$$

where the z_i's are independent random variables with zero means and variances that depend on the x_i's. In such cases the design is often taken to be sampling proportion to size using popx, i.e. pps popx. But in some cases the design will not be pps popx so it is of interest to see how the HT estimator behaves under other designs. In this problem the other design will be pps rev(popx). That is the unit with the smallest x value will have the largest inclusion probability and so until the unit with the largest x value has the smallest inclusion probability.

As was noted in class, given the sample, the weights used in the HT estimator will usually not sum to the population size. One way to modify the HT estimator is, given the sample, to rescale the HT weights used in the estimate to sum to the population size. It is easy to modify the code used in the homework for computing the HT estimator to calculate this second estimator as well.

In this problem we want to explore how important the model and the design are in the performance of the HT estimator. We will do this by comparing its performance to the alternative estimator described in the above.

The next bit of R code generates the population to be used in this problem.

```
set.seed(20122016)
popx<-sort(rgamma(500,7)) +20
popy<-rnorm(500,popx,sqrt(popx))
```

For this population generate 500 samples of size 40 using pps popx and find the average absolute error for the two estimators. Repeat this but now using pps rev(popx) as the design.

Finally repeat both of the two simulations but where now you replace each y_i with $y_i + 500$

Note, the population total for popy is 13,650.59 and the correlation between popx and popy is 0.493.

6. In the class you learned that for single stage cluster sampling it was sometimes a good idea to use the ratio estimator when estimating the population total instead of the standard estimator. In this problem you must construct such a population and show that the ratio estimator does better in a simulation study.

Let N be the number of clusters in the population and M_i denote the size of the ith cluster. When computing the ratio estimator you may assume that $M_0 = \sum_{i=1}^{N} M_i$ is known. The first step is to select your values for the cluster sizes, clssz=$(M_1, M_2, \ldots, M_{500})$, that is your population should contain $N = 500$ clusters. The units in the clusters should only take on the values 0 and 1. To generate these values for the clusters you must use the following function,

```
makecluspop<-function(a,b,clssz)
    {
        N<-length(clssz)
        ans<-matrix(0,2,N)
        for(i in 1:N){
            n<-clssz[i]
            p<-rbeta(1,a,b)
            ans[,i]<-c(n,rbinom(1,n,p))
        }
```

```
    return(ans)
}
```

where $a > 0$ and $b > 0$ are numbers you selected to generate you population. Once you have constructed your population you need to take 400 simple random samples without replacement of size 40 and find the average absolute errors for the two estimators.

7. Consider the problem of taking a sample of size n from a population of size N where n/N is small. Let d denote a vector of positive numbers of length N. Then the function *sample* in R lets you sample without replacement using d. Under this scheme the inclusion probabilities are (approximately) given by

$$\pi_i = n(d_i / \sum_{j=1}^{N} d_i)$$

Let $wt_i = 1/\pi$ be the weight associated with unit i. Now given a sample the sum of the weights of the units in the sample need not equal N. For this reason we will take as our weights

$$w_i = N(\frac{wt_i}{\sum_{i \in smp} wt_i})$$

and the resulting estimate of the population total is

$$t_w = \sum_{i \in smp} w_i y_i$$

For notational convenience we assume that the sample was the first n units of the population.

In class it was pointed our that given a sample and the resulting set of weights one way to simulate complete copies of the population to get an estimate of variance of this estimator is to do the following:

1. Observe a probability vector $p = (p_1, p_2, \ldots, p_n)$ from a Dirichlet distribution with the parameter vector, the vector with n 1's.

2. Calculate $n \sum_{i=1}^{n} w_i y_i p_i$ to get one simulated value for the population total.

3. Repeat R times to get R simulated population totals, say $t_1, t_2, \ldots t_R$ and then use $\sum_{i=1}^{R}(t_i - t_w)^2/(R-1)$ as our estimate of variance for the estimate t_w.

Here is a second way to get an estimate of variance. Let $v_i = (n/N)w_i$ for $i = 1, 2, \ldots, n$. Note the v_i's are just the w_i's rescaled to sum to the sample size n instead of the population size N.

1. Observe a probability vector $p = (p_1, p_2, \ldots, p_n)$ from a Dirichlet distribution with parameter vector, $v = (v_1, \ldots, v_n)$

2. Calculate $N \sum_{i=1}^{n} y_i p_i$ to get one simulated value for the population total.

3. Repeat R times to get R simulated population totals, say $t_1, t_2, \ldots t_R$ and then use $\sum_{i=1}^{R}(t_i - t_w)^2/(R-1)$ as our estimate of variance for the estimate t_w.

i) Show that for a given sample the expected value of the population total under the second scheme is t_w.

ii) Implement the following simulation study to compare the two methods of estimating the population variance. You might find it helpful to load into R the *rdirichlet* function using the command library(gtools). The population you will use is constructed as follows

```
set.seed(99887766)
popx<-sort(rgamma(500,5)) + 23
popy<-rnorm(500,(popx-25)^2 + 100,9)
 cor(popx,popy)
[1] 0.849
sum(popy)
[1] 57073.82.
```

For the three designs, rep(1,500), pps popx and pps rev(popx) select 500 samples of size 40 and find the average value of the estimator, its average absolute error, the length of its approximate 95% confidence interval and the frequency which the interval contains the true population total. Based on these results briefly compare these two methods.