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

Seventy three people took the exam and the high score was 84. There were 3 scores in the 80's, 5 in the 70's, 14 in the 60's, 7 in the 50's, 10 in the 40's, 11 in the 30's and 23 below 30.

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.

ANS

You should stratify on regions within the country. Within a region you should stratify on schools with two shifts and those with one shift. Note that schools with one shift are clusters and you might want to consider grades within schools as clusters as well. For schools with two shifts note each shift can also be thought of as a cluster as well. So you should probably sample from both shifts when sampling a shift school.

3. For some population of size N and some fixed sampling design let π_i 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_j 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.

ANS

i) $wt_i = 1/\pi_i$

ii) You should look at section 8.5 in the text. Now the the probabilty that the unit i responds in category j is $\pi_i\gamma_j$ and we can estimate γ_j by

$$\hat{\gamma}_j = (\sum_{rs_j} wt_i)/(\sum_{s_j} wt_i)$$

where s_j are the sample units in the jth category and rs_j are the responders in the jth category. Then the sampling weight for each respondent in class j is is $1/\pi_i\hat{\gamma}_j$.

iii) In this case since the N_j's are known, in category j we just readjust the wt_i's of the responders so that they sum to N_j.

iv) Form categories using successive intervals of the values of the auxiliary variable and use the answer from either part ii) or part iii) depending on what is known.

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.

ANS

Here is the code I used to get my answer.

```
eststrtmn<-function(smp,N)
  {
      N1<-sum(N)
      L<-length(N)
      usest <- 0
      usestvr <- 0
      smpsz<-rep(0,L)
      smpstd<-rep(0,L)
      strprop<-rep(0,L)
      for (i in 1:L) {
        ysmp<-smp[[i]]
        n<-length(ysmp)
        ff <- n/N[i]
        W <- N[i]/N1
        ysamp <- smp[[i]]
        usest <- usest + W * mean(ysamp)
        usestvr <- usestvr + W * W * (1 - ff) * var(ysamp)/n
        smpsz[i]<-n
        smpstd[i]<-sqrt(var(ysmp))
        strprop[i]<-W
    }
      dum<-sqrt(usestvr)
      lwbd<-usest - 1.96*dum
```

```
      upbd<-usest + 1.96*dum
      ans1<-c(usest,lwbd,upbd)
      n<-sum(smpsz)
      opt<-(strprop*smpstd/sum(strprop*smpstd))*n
      ans2<-rbind(smpsz,opt)

    return(list(ans1,ans2))
  }
```

Here are my results.

```
 eststrtmn(smp,N)
[[1]]
[1] 916.3534 893.7103 938.9966

[[2]]
          [,1]     [,2]     [,3]     [,4]
smpsz 20.00000 30.00000 80.00000 50.00000
opt   25.94901 32.89727 64.13293 57.02079
```

If we knew the strata standard deviations then the optimal sample size for stratum i is proportional to $W_i\sigma_i$ where $W_i = N_i/\sum N_j$ and σ_i is the stratum standard deviation. Here proportional allocation is fairly close to the optimal but the optimal would have a few less observations in stratum three and a few more in the others.

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.

ANS

Below is the output for my R functions which did the calculations.

```
> comparehtnormhtlp(popy,popx,40,popx,500)
          [,1]    [,2]     [,3]  [,4]
ansht 13651.92 306.551 1586.564 0.956
nrmht 13627.82 349.915 1592.077 0.934
> comparehtnormhtlp(popy,popx,40,rev(popx),500)
          [,1]    [,2]     [,3]  [,4]
ansht 13654.33 456.548 2348.826 0.960
nrmht 13374.13 421.716 1596.810 0.888
> comparehtnormhtlp(popy+500,popx,40,popx,500)
          [,1]     [,2]     [,3]  [,4]
ansht 263839.2 2987.920 15353.83 0.942
nrmht 263648.7  344.423 15555.94 1.000
> comparehtnormhtlp(popy+500,popx,40,rev(popx),500)
          [,1]     [,2]     [,3]  [,4]
ansht 263855.2 3339.788 17075.82 0.972
nrmht 263390.9  409.548 15686.78 1.000
```

As you can see both estimators are unbiased and HT is the best for popy when the design is pps popx, the first set of numbers. Everywhere else the HT estimator loses and sometimes by quite a bit when the underlying model is wrong. The design pps popx is clearly better than the other and is not as nearly sensitive to the model being correct.

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.

ANS

This problem is discussed in section 5.2.3 of the text.

We know the ratio estimator will be preferred when the cluster sizes vary but the proportion of ones in the clusters remain more or less constant across the clusters. To get my cluster sizes I used $round(rgamma(500, 20) + 1$ and I took $a = 16$ and $b = 24$. The true total for this population was 4169. The average values of the standard and ratio estimators were 4,176.2 and 4,168.4 so both were approximately unbiased. Their average absolute errors where 182.8 and 150.8 so the ratio estimator is better.

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 $wt_i = 1/\pi_i$ 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.

ANS i) Recall that if p is Dirichlet with parameter $(\alpha_1, \ldots, \alpha_n)$ then $E(p_i) = \alpha_i/\alpha_0$ where $\alpha_0 = \sum_{i=1}^{n} \alpha_i$ So for the second method we have

$$E(p_i) = v_i/sum_{i=1}^{n} v_i = v_i/n = (1/n)(n/N)w_i = w_i/N$$

and the result follows.

ii) Here are my results

```
         ave      av       ave      freq of
        value    err      len      coverage
Results for srs
1stans  57032.5 1408.582 6547.678 0.912
2ndans  57032.5 1408.582 6562.603 0.912

Results for design pps popx
1stans  57105.02 1304.497 4314.137 0.798
2ndans  57105.02 1304.497 7876.660 0.962

Results for design pps rev(popx)
1stans  57089.29 1518.431 9028.928 0.966
2ndans  57089.29 1518.431 6853.532 0.912
```

They are approximately unbiased but the second method is preferred since it gives better interval estimates.