FINAL EXAM
STAT 5201
Fall 2017

Due on the class Moodle site or in Room 313 Ford Hall
on Monday, December 18 at 3:30 pm
In the second case please deliver 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 Glen 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 the class web page. In case you think you have found an error you should check the class home page before contacting us.

Sixty people took the exam. The top score was 100 and there were 3 in the 90's. Then there were 7 in the 80's, 5 in the 70's, 12 in the 60's, 13 in the 50's, 12 in the 40's, 3 in the 30's, 1 in the 20's, 2 in the 10's and one score below 10.

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. A doctor at a university with 3,100 students is interested in knowing how many of the students get a flu shot.

i) Last year in a random sample of 100 students she found that 65 had gotten a shot. Give the usual 95% confidence interval for the true proportion of students that got a shot.

ii) This year she would like to know how large her sample size should be for the resulting 95% confidence interval for p, the proportion of students who had a shot, is no longer than 0.08.

Ans i)

$$0.65 \pm 1.96\sqrt{\frac{1}{99}(1-100/3100)(0.65)(0.35)} = 0.65 \pm 0.092$$

ii) From class

$$n = \frac{n_0}{1+n_0/N} \quad \text{where} \quad n_0 = (1.96/.04)^2/4 = 600$$

$$n = \frac{600}{1+600/3100} = 503$$

3. On the class web page under Data sets you will see the link, house sales. This is a link to the R data file glen.rda which is a 597 by 3 matrix called glen. You can load this matrix directly into you working R directory using the command

```
load(url("http://users.stat.umn.edu/~gmeeden/classes/5201/datasets/glen.rda"))
```

For a recent year this contains information about 597 house sales in two zip codes in St Paul. A row gives y, the sale price of a home in thousands of dollars, x, the amount of taxes paid for the house in thousands of dollars and a zip code identifier.

To answer the following questions assume you know both the sales price and tax amount for every house. In each case assume that we are estimating the population mean.

i) Give the true variance of the sample mean under simple random sampling without replacement for a sample of size $n = 60$.

ii) Suppose we form three strata, the first 300, the next 200 and the final 97. Find the optimal allocation for a sample of size 60 and the true variance of this estimator.

iii) Find approximately the true variance of the Ratio Estimator for a simple random of size 60.

Ans i) Let y_i be the sale price and x_i be the tax for the ith house. Let μ_y and μ_x be their population means then the true variance of the sample mean is

$$(1-60/597)(\sum_{i=1}^{597}(y_i-\mu_y)^2/596)/60 = 276.4$$

ii) The optimal allocation is given by

$$n_h = n\frac{N_hS_h}{\sum_{i=1}^{3} N_hS_h}$$

where N_h is the size of the hth stratum and S_h is its standard deviation. For the three strata the values of N_hS_H are 14,477.0, 8,568.3 and 10,976.5 which gives the optimal strata sizes as 26, 15 and 19.

Using formula (3.5) of page 79 of the text the true variance of this estimator is $20.6 + 12.5 + 14.3 = 47.4$

iii) Let $R = \mu_y/\mu_x = 266.2/4.22 = 63.11$. Then the approximate variance is

$$(1 - 60/597)(\sum_{i=1}^{597}(y_i - 63.11x_i)^2/598)(1/60) = 81.43$$

4) For this problem you will be using the house sales population of problem 3. In your R working directory run the following three commands

```
> set.seed(878787)
> smp1<-sort(sample(1:597,30))
> smp2<-sort(sample(1:597,30,prob=tax))
```

i) For smp2 give the value of the Horvitz-Thompson estimator for the population total of the house sales and its estimate of variance.

ii) Let y_i denote the sale price of the ith house and x_i denote the amount of taxes for the house. Consider the model

$$Y_i = \beta x_i + Z_i \quad \text{for } i = 1, 2, \ldots, 597$$

where the Z_i's are independent random variables, $E(Z_i) = 0$ and $Var(Z_i) = \sigma^2 x_i$ for some unknown σ^2. For each of the two samples find the ratio estimate of the population total and their estimates of variance.

Ans i) Using the R handout on the class web page on the Horvitz-Thompson estimator (see section 6.2.2 of the text) you find the estimate for the total for the second sample is 158,386.6 with estimated variance 22,602,066

ii) Using the R handout on the ratio estimator we find the estimates for the total are 165,520.8 for the first sample and 155,409.3 for the second. Using the same handout we find the estimated sample variances for smp1 and smp2 are 32,362,075 and 25,589,547

5) For a population with a y of interest and an auxiliary x which is correlated with y and a design you need to write a program which allows you to compare the behavior of three estimators under repeated sampling from the design. The three estimators are the Horvitz-Thompson (HT) estimator, the HT estimator that **simultaneously** constrains the weights so that the add to the population size and are calibrated on x, (one way to do this is explained in the R handout on the class web page under the link calibration using quadprog) and the estimator which assumes that the design was srs with replacement and is again adjusted so that its new weights sum to the population size and are calibrated on x. For each estimator you need to compute its average value and average absolute error for 500 samples taken using the design.

Apply your function to the population of house sales in problem 3) for three different designs. The designs are pps using x, using x in reverse order, i.e. in R use $rev(x)$ and simple random sampling without replacement. Take the sample size to be $n = 30$.

Ans Each line gives the average value of an estimator followed by its average absolute error starting with the HT estimator, the adjusted HT estimator and finally the estimator that always assumes that the design is srs.

```
> comp3(sales,tax,tax,30,500)
[1] 159279.7   6019.8 159624.8   6150.0 159493.4   6183.0
> comp3(sales,tax,rev(tax),30,500)
[1] 161736.9  26429.5 159449.4   7782.5 153901.7   7955.0
```

```
> comp3(sales,tax,rep(1,597),30,500)
[1] 158921.8  11341.8 158879.7   6414.6 158879.7   6414.6
```

Note that sum(sales)=158922.3, so all three estimators are approximately unbiased under all three designs. Under the third design the last two estimators are the same estimator.

Here is the code for the function comp3.

```
comp3<-function(popy,popx,dsgn,n,R)
{
    #need library(quadprog)
    #this compares the ht est the const and normed ht and the
    #const and norm est which assmes the sample was srs.
    prob<-n*dsgn/sum(dsgn)
    N<-length(popy)
    totx<-sum(popx)
    tru<-sum(popy)
    wts<-1/prob
    ans<-rep(0,6)
    for(i in 1:R){
        smp<-sort(sample(1:N,n,prob=dsgn))
        htwt<-wts[smp]
        htest<-sum(htwt*popy[smp])
        hterr<-abs(htest-tru)
        cnhtwt<-calibrate(htwt,popx[smp],totx,N)
        cnhtest<-sum(cnhtwt*popy[smp])
        cnhterr<-abs(cnhtest-tru)
        cnsrswt<-calibrate(rep(N/n,n),popx[smp],totx,N)
        cnsrsest<-sum(cnsrswt*popy[smp])
        cnsrserr<-abs(cnsrsest-tru)
        ans<-ans+c(htest,hterr,cnhtest,cnhterr,cnsrsest,cnsrserr)
    }
    ans<-ans/R
    return(round(ans,digits=1))
}
```

6) Again the population of interest is the population of house sales used in problem three. In this case we are interested in estimating the median of the price of the houses sold. For 500 samples of size 30, where the sampling design is simple random sampling without replacement, find the point estimate and 95% confidence interval for the median based on the polya posterior. Also find the average absolute error of your point estimate and the frequency of coverage of your interval estimate.

Next do the problem but now use pps sampling proportional to tax as your design. In this case you will need to use the function *wtpolyap* that is described in the *R* handout *usingpolyapost* on the class web page.

Ans The true median is 235 and we see from the results given below that the srs design does a bit better than pps using tax when we are estimating a median.

```
> estmed(sale,rep(1,597),30,500,500)
```

```
[1] 236.88308  17.06963 195.88677 287.80968   0.93800
> estmed(sale,tax,30,500,500)
[1] 240.29447  19.05879 199.27818 290.40276   0.92200
>
```

Here is the code I used.

```
estmed<-function(popy,dsgn,n,Rsim,Rsmp)
  {
    N<-length(popy)
    tru<-median(popy)
    inprb<-n*dsgn/sum(dsgn)
    wt<-1/inprb
    ans<-rep(0,5)
    for(i in 1:Rsmp){
      smp<-sort(sample(1:N,n,prob=dsgn))
      smpy<-popy[smp]
      wts<-wt[smp]
      wts<-n*wts/sum(wts)
      dmed<-rep(0,Rsim)
      for(j in 1:Rsim){
         simpop<-wtpolyap(smpy,wts,N-n)
        dmed[j]<-median(simpop)
      }
      est<-mean(dmed)
      err<-abs(est-tru)
      bds<-as.numeric(quantile(dmed,c(0.025,0.975)))
      if(bds[1]<= tru & tru <= bds[2]) cov<-1
      else cov<-0
      ans<-ans + c(est,err,bds,cov)
    }
    return(ans/Rsmp)
  }
```